Theory Unification and Graphical Models in Human Categorization

David Danks

Department of Philosophy, Carnegie Mellon University; and

Institute for Human & Machine Cognition

Contact information:

Department of Philosophy

135 Baker Hall

Carnegie Mellon University

Pittsburgh, PA 15213

ddanks@cmu.edu

(412) 268-8047 (phone)

(412) 268-1440 (fax)

Introduction

Disparate, mutually incompatible theories of categorization are widespread in cognitive psychology. While there are various formal results connecting pairs of these theories, the primary research focus has been on particular empirical tests of people's favorite theories. This chapter steps back from the question of which single theory (if any) is "right," and focuses instead on understanding the intertheoretic relationships among these models. Specifically, I will use the framework of probabilistic graphical models—a set of closely related computational and mathematical model-types—to provide a common *lingua franca* for a significant subset of the

extant psychological theories of categorization. This unified theoretical framework thus enables us to better understand the systematic relationships between the theories. In particular, we can gain a clearer picture of the overlaps and differences in the models' empirical predictions and underlying assumptions. Furthermore, expressing these psychological models in a common framework helps to identify several natural generalizations of currently proposed models, as well as currently underexplored alternative theories.

This graphical framework for representing various alternative models of categorization has a further, less obvious, benefit. Recent categorization research suggests that at least some categories are defined or described by an underlying causal structure (Ahn, Marsh, Luhmann, & Lee, 2002; Hadjichristidis, Sloman, Stevenson, & Over, 2004; Rehder, 2003a, 2003b, this volume; Rehder & Burnett, in press; Rehder & Hastie, 2004). Lien & Cheng (2000) found that people preferentially attend to one category from a set of possible categories, possibly quite a large set, based on which category optimizes causal learning and inference. Categorization thus seems to rely (sometimes) on causal reasoning. At the same time, all causal learning theories currently available—whether associationist or computational, normative or descriptive—assume that people are trying to learn causal relationships among a fixed set of well-defined variables; in other words, all current theories of causal learning assume some fixed categorization of the world. We also know that causal learning and prediction can suffer significantly if we do not have the appropriate (in a still unclear sense) categories (e.g., Spirtes & Scheines, 2004).

These results and observations all point toward a deep interdependence between (at least parts of) the cognitive abilities of causal learning, inference, and prediction on the one hand, and categorization and category generation/learning on the other hand. As a result, we should aim to find a common representational language for categorization and causation, so that clear questions

can be simultaneously asked about both. Given the growing evidence (much of it described elsewhere in this book) that Bayesian networks—one particular type of probabilistic graphical model—underlie parts of causal cognition, this chapter's framing of categorization theories in terms of probabilistic graphical models provides an important early step towards understanding the relationships between causation and categorization.

In the next section, I will introduce three different categorization theories, all of which have figured prominently in recent research. I then introduce two different types of probabilistic graphical models—Bayesian networks and Markov random fields—and describe how these categorization theories can be straightforwardly understood in terms of inference in particular instances of these model-types. These formal equivalencies have various implications, both theoretical and experimental. Some of the implications are clear and immediate, including simple explanations for various model-fitting results in the experimental literature. Other implications are more suggestive. In particular, the mathematical equivalencies described in earlier sections suggest several substantive categorization theories that are, to my knowledge, novel to the psychological community (though not within machine learning). In the final substantive section, I will focus on one particular model and (programmatically) describe how it could account for a range of intriguing phenomena in various domains.

Three Similarity Functions

The general problem of categorization is to classify an object as belonging to a particular group. This classification can then be used for many different purposes, including inference of unobserved properties of this individual based on common properties within the group. For example, when hiking, I frequently (and quickly!) classify poorly-seen animals in terms of their

species. To make this judgment, I must attend to particular features and properties in the world, some of which I consciously attend to, others which I do not. In addition, my classification will depend (in part) on the other possibilities I consider. The same critter that I classify as a "pika" in Colorado might be classified as a "squirrel" in Pennsylvania (since I know that there are no pikas in Pennsylvania). Once I have classified the animal, I then decide whether to be concerned about the animal or not based on what I know about that species (e.g., a mountain lion or a squirrel, respectively). This latter task is typically referred to as feature inference or property induction: determining the likelihood that some novel instance of this category will have a particular property. In this section, I describe three different psychological theories that aim to model the cognitive representations and algorithms underlying this process.

Although there are some exceptions, most psychological models of categorization separate categorization into two stages. For a novel instance *X* and some set of possible categories, I first determine how representative *X* is of each potential category. These "similarity ratings" are then integrated in a second step to produce a behavioral response, such as my categorization of this critter as a "squirrel." In experimental settings, the relevant possible categories for a particular novel instance are invariably dictated by the cover story; in the real-world, the possible categories are selected on some poorly-understood bases, such as pragmatics or prior knowledge. Most psychological research has focused on the similarity rating function(s); relatively little empirical work has been done on the second stage integration of similarity ratings (though see Wills, Reimers, Stewart, Suret, & McLaren, 2000).

More formally, suppose that we represent individuals in terms of *n* (binary or continuous[i]) features, denoted by $F_1, \ldots, F_n$. These features are presumably selected by some process outside of the categorization theory itself. Throughout this chapter, I will make the

standard assumption for categorization theories that these features are well-defined and well-specified. Similarity ratings for a particular category are thus just functions on these *n* features. The standard second-stage integration rule for the similarity ratings is the Shepard-Luce rule (Shepard, 1957; Luce, 1963): if $S_C(X)$ denotes the similarity of X to category C and Q indexes over all of the potential categories, then $P(\text{Respond "}C\text{"} \mid X) = S_C(X) / \Sigma S_Q(X)$. That is, the probability of classifying X as a C is given by X's similarity to C, divided by the sum of X's similarity to every possible category (including C). Bias parameters are occasionally used (Logan, 2004), as well as other rules with significant formal connections to the Shepard-Luce rule (Ashby & Maddox, 2003).

In this section, I will focus on the similarity functions for standard versions of exemplar (e.g., Kruschke, 1992; Lamberts, 1998, 2000; Nosofsky, 1986; Nosofsky & Palmeri, 1997; Zaki, Nosofsky, Stanton, & Cohen, 2003; Logan, 2004 provides a good overview of recent work), prototype (e.g., Minda & Smith, 2001; Smith & Minda, 1998), and causal model (e.g., Rehder, 2003a, 2003b) theories of categorization. Substantial empirical support has been found for all three types of model, depending on the particular category, cover story, and task. And while these three similarity functions do not exhaust the space of proposed theories, they underlie the most widely discussed theories. In particular, this analysis includes Nosofsky's (1986) Generalized Context Model (GCM; described below), which is the almost universal standard against which new psychological theories are judged. Rule-based categorization theories (including Nosofsky, Palmeri, & McKinley's, 1994 RULEX model) are indirectly covered by this section, since single-feature rules are equivalent to exemplar/prototype models in which we attend to only one feature. More direct analysis of rule-based models is rarely possible, since simulations are almost always required to generate predictions for any realistic situations. Note

that dynamic measures of categorization, including category learning dynamics and response time predictions, will not be considered here.[ii]

The Generalized Context Model (GCM; Nosofsky, 1986) provides the basis exemplar similarity function for numerous other categorization theories (e.g., Erickson & Kruschke, 1998; Kruschke, 1992; Lamberts, 1998, 2000; Nosofsky & Palmeri, 1997). The core intuition behind the GCM is that the similarity or typicality of some novel instance *X* for category *A* is given by the average distance in the "category space" between *X* and some subset of previously observed category instances (the exemplars). In other words, I represent a category in terms of exemplars (previous instances known to be in the category). A new object is similar just to the extent that it is "close" to the previous observations. For example, my category of "bird" is defined by remembered previous instances of birds (e.g., a robin, an ostrich, and so on). My category of "squirrel" is defined by previously observed squirrels. And some new critter is classified as a bird rather than a squirrel just when its average distance to the bird exemplars is less than its average distance to the squirrel exemplars (and those are the only two possibilities considered).

Mathematically, we define a GCM (i.e., exemplar-based) category *A* by a set of exemplars $E_1, \ldots, E_m$, each of which is a full specification of values for the *n* relevant features. Let $Y(i)$ denote *Y*'s value for the *i*-th feature. The similarity between novel instance *X* and a particular exemplar E_j is then given by $Sim(X, E_j) = \prod_{i=1}^{n} \exp\left[-c \times \alpha_i \left| X(i) - E_j(i) \right|\right]$, where α_i is a salience parameter for the *i*-th feature, and *c* is a global weighting parameter.[iii] That is, the similarity is the product of (the exponential of) the distances between *X* and E_j on each of the feature dimensions. Note that, if the features are all binary-valued, then the similarity is just the product of $\exp[-c \times \alpha_i]$ for each feature F_i on which *X* and E_j differ. The overall similarity rating of novel instance *X* for category *A* in the GCM—that is, the output of the first stage of the

categorization model—is the weighted sum of similarities for all category exemplars: $GCM(X, A) = \sum_{j=1}^{m} W_j Sim(X, E_j)$. The similarity ratings for each GCM-category (e.g., *GCM*(*X*, *A*), *GCM*(*X*, *B*), and so on) are then combined using the Shepard-Luce rule to generate behavioral responses. In addition, the set of exemplars (i.e., the category definition) can straightforwardly be used for inference about unobserved features of objects placed into the same category: e.g., "this is a squirrel, and most of my squirrel exemplars were not aggressive, therefore this squirrel probably won't be aggressive."

Prototype-based theories offer a different picture of categorization from exemplar-based models. Instead of basing the category on a set of previously remembered instances, categories are defined by prototypes—single objects—that encode a summary or average of people's experiences with the category; these prototypes need not correspond to any actual category instance (and almost never will). A novel instance's similarity to the category then depends on its distance in "category space" to that single prototype. The prototypical bird, for example, is not any bird that has ever been observed, though various actual birds (e.g., robins) are closer to it than others (e.g., ostriches). Mathematically, the prototype model (e.g., Minda & Smith, 2001; Smith & Minda, 1998; see also versions in Johansen & Palmeri, 2002; Olsson, Wennerholm, & Lyxzèn, 2004) is almost always a GCM model with only one exemplar for the category, but where the exemplar might not have been observed.[iv] However, this standard, simple prototype model fails to do justice to the intuition behind prototype models. Information about inter-feature connections or correlations is an important part of any summary of a series of observations, and this information cannot be expressed in the description of a single instance. The standard prototype similarity function requires some augmentation to capture the underlying intuition.

A simple way to incorporate inter-feature correlations is with second-order features: features whose value is entirely determined by the values of two first-order features. (Second-order features are only one way to capture correlations; a more powerful option is discussed later.) For example, it might be important that both F_2 and F_{17} occur, perhaps because an observed correlation. In that case, we could define a second-order feature that "occurs" if and only if F_2 and F_{17} both occur. Second-order features consisting of logical AND functions are quite common (e.g., Rehder, 2003a, 2003b), but are certainly not the only kind of second-order feature that could be introduced; Danks (under review) gives a general, mathematical definition of (plausible) second-order features. If we allow second-order features into a category prototype, then we have to adjust the first-order feature prototype similarity function (which was just the GCM similarity function). For simplicity, I will use $d(i, j)$ to denote the distance between (instance) X and (prototype) E along the feature composed of F_i and F_j (if $i = j$, this is the appropriate first-order feature). Let α_{ii} be the salience of first-order feature F_i, and α_{ij} be the salience of the second-order feature composed of F_i and F_j. ($\alpha_{ij} = 0$ implies no second-order feature for F_i and F_j.) Given this notation, the second-order prototype (SOP) category similarity function is: $SOP(X, A) = \prod_{i=1}^{n} \prod_{j=i}^{n} \exp\left[-c \times \alpha_{ij} d(i, j)\right]$. That is, the similarity between some instance X and category A is the product of (the exponentials of) the distances between X and A for each feature, including second-order ones. Once X is categorized into a particular prototype-based category, feature inference is based entirely on the summary statistics encoded in the prototype itself. If the value of "Flies" is 0.95 for the prototypical bird (i.e., 95% of birds summarized in the prototype could fly), then the probability that this bird flies is 0.95.

The third psychological theory of categorization is causal model theory (Rehder, 2003a, 2003b, this volume). Causal model theory defines a category in terms of a particular causal

structure among the features, including possibly unobserved features (e.g., an animal's "essence"). The underlying intuition about similarity is that a particular instance is more likely to be a member of category *A* just when its observed features respect the causal relationships among the various features. Thus, the similarity function for a category in causal model theory is the probability that a particular novel instance would be generated by that category's causal structure (perhaps multiplied by some scaling factor). For example, a particular object is similar to "bird" when the combination of observed features would likely be produced by something with the causal structure underlying the category of "bird." Obviously, the mathematical details of causal model theory depend heavily on the particular representation of causal structures. Current versions of the theory model these structures using causal Bayesian networks (or causal Bayes nets). Details about causal Bayes nets are provided in the next section. For now, the relevant feature of a causal Bayes net is that it can be used to determine the probability of any particular combination of feature values given some causal structure; the causal model theory (CMT) similarity function is directly proportional to that probability. That is, $CMT(X, A)$ is proportional to $P(X \mid M)$, where M is the causal Bayes net for the category. Given a particular categorization, the causal structure can straightforwardly be used for feature inference (Ahn, *et al*., 2002; Hadjichristidis, *et al*., 2004; Rehder & Burnett, in press; Rehder & Hastie, 2004).

In this section, I have left out several different types of categorization theories; perhaps most notably, I excluded connectionist models (e.g., Gluck & Bower, 1988; McClelland & Rogers, 2003; Rogers & McClelland, 2004). There is reason for their exclusion. Connectionist models have the ability to model or approximate large classes of input-output functions. However, to determine the exact space of similarity ratings that can be modeled by a particular network, we must perform significant simulations, except in specific networks that can model *all*

possible input-output relationships. Without analytic results about the input-output relationships that can be modeled by a particular neural network structure, there is no definite target for expression in the framework of probabilistic graphical models. Moreover, it is notoriously difficult to determine what representations are contained in a connectionist model, since much depends on the particular connection weights that emerge from a learning history. As a result, process equivalencies that directly map the symbolic operations of the connectionist model onto a graphical model are also not forthcoming.

Probabilistic Graphical Models

The central theoretical claim of this chapter is that the similarity functions from the previous section can be usefully and interestingly described in the framework of probabilistic graphical models. In this section, I outline two types of graphical models—Bayesian networks and Markov random fields—and then describe how various similarity functions are proportional to calculating $P(X \mid Model)$, where *Model* is one of these probabilistic graphical models. That is, the various psychological theories make different predictions because they assume different graphical model types: a subclass of Bayesian networks for exemplar-based theories (the GCM), causal Bayesian networks for causal model theory, and a subclass of Markov random fields for prototype-based theories. Thus, these diverse theories can be viewed (from a mathematical point-of-view) as different parameterizations of a single unified theory. These mathematical observations raise a range of psychological implications and questions; I take up those issues in the subsequent two sections. Due to space constraints, I have omitted the full proofs and technical details about the various equivalencies; the relevant mathematical specifics can all be found in Danks (2004, under review).

In general, probabilistic graphical models provide a compact representation of a probability distribution by encoding various facts about independence and association in some type of graph. Strevens (this volume) explores the importance of (usefully) compact representations of probability and statistical facts. Bayes nets are one of the most popular probabilistic graphical models for such purposes; I here provide a very brief introduction to the framework. (Neapolitan, 2003; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; and other chapters in this volume all provide more comprehensive introductions to Bayes nets.) It is important to realize that, despite the name, there is nothing intrinsically Bayesian about a Bayes net; the name is due to the original uses of the framework. One can be, but need not be, a Bayesian about Bayes nets.

A Bayes net is defined relative to a set of variables; in our current setting, these are the observed features. One half of a Bayes net is a directed acyclic graph containing one node per variable/feature (see Figure 1). These nodes are (possibly) connected by directed edges (e.g., $F_i \rightarrow F_j$), indicating an asymmetric relationship. In "simple" Bayes nets, the asymmetric relationship is purely probabilistic. In contrast, a *causal* Bayes net (used by causal model theory, as well as multiple psychological theories of causal reasoning) is a Bayes net in which the edges in the graph are provided a causal interpretation. If the causal interpretation is justified by background knowledge, then $X \rightarrow Y$ indicates that X is a direct cause of Y, where no particularly substantive theory of causation is presupposed (see Woodward, 2003, for one possibility). We use family terminology (e.g., 'parent' or 'child') to describe the graphical relationships. The 'acyclicity' property of the graph means that there is no (non-trivial) arrow-following path from a variable back to itself (e.g., there cannot be a path like $F_1 \rightarrow F_3 \rightarrow F_{17} \rightarrow F_1$ in the graph).

The other half of a Bayes net is a joint probability distribution (or density, for continuous variables) that specifies the probability of any particular set of feature values. When the causal interpretation is justified, the joint probability distribution encodes information about the quantitative causal dependencies among the variables. The two Bayes net components—the directed acyclic graph and the joint probability distribution—are connected by a Markov assumption: every variable is probabilistically independent of its non-descendants conditional on its graphical parents. This assumption implies that the joint probability distribution (density) factors as $P(X) = \prod_{i=1}^{n} P(F_i \mid pa(F_i))$, where $pa(F_i)$ denotes the graphical parents of F_i. The components are also connected by the Faithfulness assumption: the only probabilistic independencies are those predicted by the Markov assumption. The primary effect of the Faithfulness assumption is to exclude the possibility of multiple pathways whose effects *exactly* cancel out (e.g., $X \rightarrow Y \rightarrow Z$ and $X \rightarrow Z$, but X and Z are unconditionally independent). Faithfulness is assumed (either explicitly or implicitly) by essentially every Bayes net learning algorithm. An example of a Bayes net is provided in Figure 1.

[Insert Figure 1 about here]

The causal model similarity function is already expressed using causal Bayes nets: the causal structure defining category A must be a causal Bayes net, and the similarity of X to A is given by the probability of X in the joint probability distribution represented by the causal Bayes net. That is, the similarity rating of X for category A is equal to $P(X)$, where the probability distribution is represented by a causal Bayes net.[v] Thus, this categorization theory can easily be represented in terms of inference for probabilistic graphical models.

Perhaps more surprisingly, Bayes nets can also be used to express the exemplar-based GCM similarity function. In general, the similarity functions used in these two-stage

categorization theories are defined for all possible instances. Therefore, the pattern of those ratings for a particular category is proportional to some probability distribution over those same possible instances. So for example, if we have some exemplar-based (i.e., GCM) category A with its corresponding similarity function, $GCM(X, A)$, then there is necessarily some probability distribution $P(X)$ such that $GCM(X, A) \propto P(X)$, for all instances X (i.e., there is some constant K such that $GCM(X, A) = K \times P(X)$ for all X). Hence, to establish an equivalence between the GCM and some probabilistic graphical model, it suffices to show that, for every probability distribution proportional to a possible set of ratings for a GCM category, there is a *perfect map* in some class of probabilistic graphical models, and *vice versa*. A graphical model provides a perfect map of a probability distribution if and only if the graph implies (by Markov and Faithfulness) all and only the probabilistic independencies that occur in that distribution. In general, the (high-level) strategy for expressing categorization theories in terms of probabilistic graphical models is: determine the patterns that could possibly be produced by (normalized) similarity functions, and then find a set of probabilistic graphical models that perfectly represent exactly those patterns.

In the case of the exemplar-based GCM, consider a Bayes net with the directed acyclic graph in Figure 2. E is an unobserved variable whose number of values depends on the category being modeled. By the Markov assumption, the joint probability distribution for this Bayes net factors into $P(E, F_1, \ldots, F_n) = P(E) \times P(F_1 \mid E) \times \ldots \times P(F_n \mid E)$. The structure of this model is similar to the oft-used naïve Bayes models of machine learning classification problems, though the role and interpretation of the (unobserved) common cause is different in this situation.

[Insert Figure 2 about here]

Regardless of whether the features are binary (e.g., either present or absent) or continuous (e.g., height), *every* GCM category is proportional to a probability distribution over the F_i's that

has a perfect map given by a Bayes net[vi] with this graph. That is, for every GCM category, there is a Bayes net with Figure 2 graph and associated probability distribution such that $GCM(X, A) \propto P(X)$ for every possible instance X over features $F_1, \ldots, F_n$. The converse of this claim holds with a slight addition: for every probability distribution over the observed F_i's for a Bayes net with Figure 2 graph *and* a "regularity" constraint on the form of the $P(F_i \mid E)$ terms, there is a GCM category whose ratings are proportional to that distribution. The exact regularity constraint depends on whether the features are binary or continuous, but neither constraint is particularly strong.[vii] Since similarity ratings are determined only up to a choice of scale, we can conclude that GCM categories and Bayes nets with a Figure 2 graph (plus regularity constraint) describe exactly the same set of similarity ratings; any responses that can be fit to one model can be fit to the other.

In contrast, there is no corresponding equivalence between Bayes nets and prototype-based categorization models with second-order features. These two types of models are fundamentally different, in that the inter-variable connections in a Bayes net are asymmetric (whether in a probabilistic or causal sense), while the second-order features are symmetric. Hence, we need a probabilistic graphical model with *un*directed edges between the features to indicate symmetric connections. This model-type is called a Markov random field (see, e.g., Darroch, Lauritzen, & Speed, 1980; Lauritzen, 1996 for more technical introductions). As with Bayes nets, Markov random fields are defined only relative to a set of variables (features), and are composed of a joint probability distribution (density) and a graph. In contrast with a Bayes net, though, a Markov random field graph contains undirected edges between the nodes (see Figure 3). Roughly speaking, two features being connected by an edge in the graph implies that

there is a probabilistic dependence between those features' values, but no explanation of the correlation is given or presumed (and so there is no asymmetry between the variables).

The graph and probability distribution in the Markov random field are connected by a Markov assumption: the probability of any feature value is dependent only on its graphical neighbors. So for example, in Figure 3, $P(A)$ depends only on B and C; A is probabilistically independent of D and E, conditional on B and C. The Markov assumption implies that the joint probability distribution can be factored into the product of functions (called *clique potentials*) over the *maximal cliques* in the undirected graph. A graphical *clique* is any set of nodes for which every pair is connected by an edge, and a clique is maximal if adding any other variable renders it no longer a clique. Thus, the Markov assumption for a Markov random field G implies that, if the maximal cliques in G are denoted by $C_1, \ldots, C_q$, we can express the probability of some novel instance X as $P(X \mid G) = \frac{1}{Z}\prod_{i=1}^{q} g_i(X)$, where $g_i(X)$ depends only on the values of variables in C_i (and Z is a normalization constant). Figure 3 provides an example of a Markov random field, including both the graph and the factorization of the joint probability distribution into clique potentials.

[Insert Figure 3 about here]

As with Bayes nets and exemplar-based models, we can successfully apply the same high-level strategy to connect Markov random fields and prototype-based models. The patterns of ratings produced by prototype-based similarity functions can be understood as probability distributions, and we can represent that space of probability distributions in terms of Markov random fields. More specifically, for a particular (second-order feature) prototype-based category, its Markov random field counterpart contains an edge between two nodes (features) just when there is a second-order feature for those two. Then, for every possible pattern of

similarity ratings, there is a corresponding (proportional) probability distribution that has a perfect map given by the category's Markov random field counterpart. And for every probability distribution with a Markov random field perfect map (whose clique potentials satisfy a further, relatively weak, regularity constraint), there is a corresponding prototype-based category whose similarity ratings are proportional to the distribution. Just as GCM-categories are equivalent to (probability distributions with perfect maps given by) Bayes nets with Figure 2 graph, (second-order) prototype-based categories are equivalent to (the probability distributions for) a subset of Markov random fields.

In summary, all three types of similarity functions (GCM, second-order prototype, and causal model) can be expressed (up to a scaling parameter) as computations of $P(X \mid Model)$, where the differential theory predictions arise from different assumptions about the underlying graphical model. The precise psychological model $\leftrightarrow$ graphical model relationships are:

GCM rating for $X \leftrightarrow P(X \mid$ Bayes net with Figure 2 graph and constraint)

Second-order prototype rating for $X \leftrightarrow P(X \mid$ Markov random fields with a constraint)

Causal model rating for $X \leftrightarrow P(X \mid$ Causal Bayes net)

The representation of category similarity functions as probability distributions has been previously explored by Myung (1994), Ashby & Alfonso-Reese (1995), and Rosseel (2002). In contrast to that work, the results detailed here use the framework of probabilistic graphical models, which allow us to extend the formal results to a broader class of prototype theories, as well as to include causal model theory. In related research, Nosofsky (1990) and Ashby & Maddox (1993) pursued a more direct strategy and found conditions in which exemplar models could be directly transformed into prototype models and *vice versa* without the framework of probability theory or graphical models (see also Barsalou, 1990). While important for

understanding those two theory-types, though, the direct results are not readily extensible to other psychological theories (e.g., causal model theory) because they do not situate the theories in a more general framework.

With these equivalencies in hand, I now turn to their implications. The next two sections demonstrate several pragmatic uses of the representation of these models as probabilistic graphical models, including better understanding of existing experimental results, suggestions of novel experiments, and more speculatively, the possibility of interesting generalizations of existing psychological theories.

Applying the Graphical Model Equivalencies

The most obvious application of these equivalencies is to facilitate rapid determination of the conditions in which the categorization theories make differential predictions, thus enabling us to both explain previous experimental results, and construct appropriate novel experiments. In particular, if the two probabilistic graphical model-types can perfectly represent different probability distributions *and* people's behavioral responses approximate the observed probability distribution for some category, then we can determine analytically whether some experiment is likely able to discriminate between the second-order prototype (=Markov random field) and causal model (=causal Bayes net) theories. In fact, the expressive potentials of Bayes nets and Markov random fields are appropriately overlapping. That is, there are probability distributions that can be represented perfectly by a Markov random field but not a Bayes net, and *vice versa*. There are also probability distributions (e.g., those equivalent to first-order prototype-based similarity functions) that can be represented perfectly by both Bayes nets and Markov random fields, as well as some that cannot be represented perfectly by models from either framework.

As a concrete example, there is no Bayes net that perfectly represents a probability distribution with the (Markov random field) factorization given in Figure 3. Thus, if Figure 3 describes the actual underlying category structure (i.e., the probability that any novel instance comes from that category), then a causal-model-based categorizer would not be able to perfectly learn the category structure. Similarly, there is no Markov random field that perfectly represents a probability distribution with the (Bayes net) factorization in Figure 1, and so a categorizer using second-order prototype-based categories would not be able to accurately learn that category. Finally, a simple category structure consisting of uncorrelated features can be equally well-represented by models from both frameworks, and so no experiment based on such categories will be able to distinguish between causal model and second-order prototype-based categorization (i.e., the psychological theories should have equally good model fits).

We can also apply this analysis to published—not just hypothetical—experiments. Rehder (2003a)'s common-cause condition uses a category probability distribution that can be equally well-represented by a Bayes net (= causal model) and a Markov random field (= second-order prototype). As predicted, he found no model fit difference between the corresponding psychological theories (see Table 5, p. 729). In contrast, Rehder (2003a)'s common-effect condition used a probability distribution that can be represented by Bayes nets but not Markov random fields. Thus, the two psychological theories should be distinguishable by that experiment: second-order prototype categorizers will do poorly, and causal-model categorizers should do well. Alternately, if we assume that people can learn a wide range of category structures, then we should expect the second-order prototype theory to have a significantly worse model fit than the causal model theory. The subsequent data analysis found exactly that significant model-fit difference in favor of causal model theory, which can represent the

underlying probability distribution (see Table 5, p. 729). (See also Experiment 3 in Rehder, this volume, for further evidence of an asymmetry between common cause and common effect networks.)

Finally, we can use this analysis to design experiments to push the outer boundaries of human category learning. As noted previously, there are probability distributions, and so categories, that none of the psychological theories can completely model. Correct theoretical predictions of cognitive failures (in this case, failure to correctly represent the category) are typically thought to constitute stronger evidence for a theory than predictions that people will behave close-to-optimally. Thus, a natural way to separate these three theory-types is to present individuals with categories whose structure cannot be mapped onto any of these representations without loss of information. In particular, we want to find categories for which each theory picks out different aspects of the structure, and so they predict different patterns of failure. Chain graphs are probabilistic graphical models that use both directed and undirected edges (further discussed below), and there are perfect map chain graphs for probability distributions with no Markov random field or Bayes net perfect map. One such graph is: $F_1 \rightarrow F_2 — F_3 \leftarrow F_4$.[viii] All three psychological theories predict that people will make significant, systematic, predictable errors when presented with a category with this structure, and those errors are predictable using the probabilistic graphical model equivalencies described here. The differential error predictions can then be used to better determine which theory best describes an individual's categorization process. To my knowledge, no such experiment appears in the literature.

In additional to methodological implications, these equivalencies suggest natural generalizations of existing psychological theories. The exemplar-based and second-order prototype-based similarity functions are equivalent with only subclasses of Bayes nets with

Figure 2 graphs and Markov random fields, respectively. In both cases, the equivalent graphical models have constraints on the probability distribution beyond those implied by the graphical model. From the probabilistic graphical model point-of-view, these additional constraints seem arbitrary, though they have a natural justification in terms of ensuring computational tractability. Setting aside computational issues, though, we might naturally consider generalizing the GCM to include patterns of similarity ratings that are proportional to *any* probability distribution with a perfect map Bayes net with Figure 2 graph. This generalization has a straightforward interpretation within the GCM framework: it corresponds to allowing exemplar-dependent feature saliences in the similarity function. Similarly, we can generalize the second-order prototype model to include any probability distribution with an arbitrary Markov random field perfect map. This generalization would significantly extend the scope of that theory, while retaining the basic intuition of prototype theories that the category representation is a summary of the observed category instances. Importantly, both of these generalizations remain bounded in explanatory power; there are experiments and patterns of similarity ratings, such as Rehder (2003a)'s common-effect condition, that can distinguish these generalizations from one another.

Finally, these equivalencies suggest alternate responses to two existing problems for categorization theories: empirical support for (seemingly) inconsistent theories, and (apparent) shifts in category structure during learning. The first problem is that there is significant empirical evidence supporting all three of these psychological similarity ratings, depending on the particular domain, presentation format, contrast class, and so on. One response to this fact has been to argue that there are distinct cognitive systems for different categorization strategies (e.g., exemplar vs. rule-based), and that contextual factors and background knowledge determine which system is activated. This idea is supported by evidence from reaction time (Allen &

Brooks, 1991) and neuroimaging (Grossman, Smith, Koenig, Glosser, DeVita, Moore, & McMillan, 2002; Patalano, Smith, Jonides, & Koeppe, 2001; Smith, Patalano, & Jonides, 1998) studies (see also Machery, in press). In a similar vein, Ashby and his colleagues have argued that different neural systems underlie implicit and explicit category learning, which are distinguished by whether participants can give a simple, verbal rule to differentiate the categories (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Waldron, 2000; Waldron & Ashby, 2001). These proposals all share the underlying idea that there are multiple processing systems in the brain responsible for the different types of categories.

The equivalencies described here suggest a different response to the range of empirical supports: the differential behaviors (perhaps) arise from differing parameterizations of a common categorization algorithm. That is, these distinct psychological theories might correspond to the same operation applied to three different representations (i.e., types of graphs), rather than distinct cognitive mechanisms. There might be only one process in which similarity ratings are based on $P(X \mid Model)$, but where the particular category model-type depends on factors such as experiential history, context, other background knowledge, and so on. Differential behavior arises from different inputs to the same process, rather than fundamentally different processes. If the cognitive representation of the category structure is a Bayes net with Figure 2 graph, the person will exhibit GCM category behavior. If the representation is a suitable Markov random field or causal Bayes net, categorizations will be best understood using second-order prototype or causal model theory, respectively. Of course, this suggestion is not intended to demonstrate that there cannot possibly be multiple processes; rather, it is intended to defeat the (too quick) inference from “support for multiple theories” to “multiple cognitive processes must exist.”

Understanding the different categorization theories in terms of probability calculations using different representations also provides a straightforward solution to the problem of integrating similarity ratings of categories with different structures. Regardless of category structure, all of the similarity ratings are on the same scale and have a clear interpretation, and so can easily be integrated into a single, coherent behavioral response. In contrast, the "multiple systems" hypothesis must provide some further account (perhaps in terms of probabilities) to explain how similarity judgments from entirely distinct cognitive processes are integrated to produce well-defined categorization judgments.

Finally, the underlying category structure *type* sometimes seems to change in response to repeated exposure to category examples (Johansen & Palmeri, 2002; Smith & Minda, 1998). For example, I might initially represent a category using a prototype, but shift to using an exemplar representation (or *vice versa*). Smith & Minda (1998) found that exemplar-based (specifically, GCM) structures were predominant throughout learning of small, poorly differentiated categories. During the learning of larger, more clearly delineated categories, however, there seemed to be a shift from prototype-based to exemplar-based category structures (see also Minda & Smith, 2001; Zaki, *et al*., 2003, and the overfitting worries of Olsson, *et al*., 2004). Johansen & Palmeri (2002) found a similar shift towards exemplar models during learning, though rule-based categories rather than prototype-based ones were more prevalent in early stages of their experiments. Rehder (this volume) suggests other trajectories for shifts in underlying category structure type.

The common framework of graphical models enables us to clearly articulate both theoretical and experimental questions about these phenomena. An immediate question that arises in these analyses centers on representational power. The experimental results were

analyzed by determining model fits for similarity responses at different times, and so the conclusions about likely shifts in category structure should be qualified by the precise model name. In particular, all of these analyses used only first-order prototype models; that is, they assumed that the category was represented by a single (perhaps unobserved) instance. The representational power of first-order prototype models is easily expressed in graphical model terms: they can only represent probability distributions whose perfect maps are Markov random fields with no edges between features, and so no inter-feature correlations. The GCM exemplar model, by contrast, can represent inter-feature correlations, though (for plausible instances) only of a certain type. As described above, the shifts from (apparent) prototype-based to (apparent) exemplar-based structures were more pronounced for categories with correlated features. Thus, given the significantly weaker representational power of first-order prototype models, it is entirely conceivable that these shifts in best-fitting model-type are due to this power imbalance, rather than actual cognitive changes. Reanalysis with a more sophisticated prototype model, perhaps one based on arbitrary Markov random fields, is warranted.

Some Speculations about Human Categorization

The previous sections focused on the equivalencies between three common similarity functions and computing $P(X \mid Category)$, where the theories differ about the exact form of *Category*. These similarity ratings are the first stage in a two-stage process. The second stage is typically the Shepard-Luce rule: the probability of responding with category C for novel instance X is the similarity rating between C and X, divided by the sum of similarity ratings for all other considered categories. Mathematically, if similarity ratings correspond to $P(X \mid Category)$, then use of the Shepard-Luce rule corresponds to computing $P(Category \mid X)$ if every category under

consideration is equiprobable.[ix] Use of bias parameters in the Shepard-Luce rule (as in Logan, 2004) then corresponds to allowing the possible categories to have different base rates. Thus, the complete (i.e., two-stage) theories solves arguably the central problem for categorization: given observations of a novel object's features, determine the probability that it falls in one or another category.

Moreover, the two stages of these theories are intended quite seriously: they are supposed to describe cognitively separable steps in categorization. So, for example, some experiments aim to obtain judgments of an item's similarity to a particular category without invoking actual categorization judgments (e.g., Rehder, 2003a, though Barsalou, 1985 argues that many other factors enter into similarity ratings). Alternately, we might suppose that people's conscious access to their categorization judgments is limited to the final output: in particular, $P(Category \mid X)$, where *Category* can be one of several probabilistic graphical models, including a Bayes net with Figure 2 graph, a Markov random field, or a causal Bayes net.[x] That is, rather than (explicitly) categorizing in two distinct stages, people directly determine the category probability given the instance. The most notable previous example of categorization based directly on $P(Category \mid X)$ is Anderson's (1991) rational analysis model. However, Anderson's model and subsequent extensions are not based on graphical models, and so are not considered here.

The central difference between the one- and two-stage views of categorization lies in the requirement of a contrast category to compute anything in the first view, while no contrast category is required to compute the similarity ratings of the second view. Any computation of $P(Category \mid X)$—whether by one or two stages—presupposes that there is at least one alternative category, else the probability is just one (since *Category* is the only possibility). In contrast, no information about any alternative categories is required to compute $P(X \mid Category)$,

which is all that is required for the first stage of the two-stage process. If categorization is really a one-stage process (i.e., the only conscious access is to the probability of the category given the instance, rather than intermediate similarity ratings), then one must explain the source of people's similarity ratings in experimental settings. The most natural explanation is that "similarity ratings" are actually categorization judgments with some implicitly assumed contrast class. That is, judgments of the typicality of a novel instance *X* for some category *A* are not actually similarity ratings, but rather are people's judgments of $P(A \mid X)$ relative to an implicit contrast category *not-A*.

If people's similarity judgments are really categorization judgments, then they should be influenced by variations in the contrast class structure and base rate. One experimental test to distinguish one- from two-stage views would be to present people with a novel category and instances of that category, all with an explicit contrast class (i.e., instances not in the novel category are definitely in the contrast class). We could then ask for ratings of the typicality of novel instances for the target category, where we vary between conditions either the structure or base rate of the *contrast* category, but not of the target category. The central prediction in this proposed experiment is: if categorization is (consciously) a one-stage process, then there should be a statistically significant difference between similarity ratings in the conditions; if categorization is (consciously) a two-stage process, then there should be no such difference. Even though the structure of category *A* does not change, $P(A \mid X)$ does change between conditions because of changes in the structure/base rate of the contrast class. The precise change will depend on the details of the contrast category structures (or base rates), but can be determined quantitatively. Importantly, note that this proposed experiment tests stability of *representativeness* (or similarity) judgments of some novel instance for a category, and not

categorization judgments. Both views agree that participants' categorization judgments should vary as the contrast class structure or base rate vary; the disagreement is about whether the contrast class matters for the typicality ratings.

Das-Smaal & De Swart (1986) performed an experiment similar in structure to this proposed one, and found limited evidence that representativeness (similarity) ratings for stimuli change depending on the contrast class. Unfortunately, they did not obtain typicality ratings for every possible combination of features, so we cannot use their experiment as even a first step towards development of a formal model. If these results can be suitably replicated and extended, then additional experiments can aim to determine (i) whether people have an implicit contrast class if not given an explicit one; and (ii) the structure of the implicit contrast class, if it exists.

The reason for exploring categorization as a one-stage process is because it opens a range of mathematical possibilities. The one-stage view draws attention to the importance of incorporating multiple potential categories into a single mathematical/representational structure (since categories are never considered in isolation). As in the existing psychological theories, categories in the one-stage theory can be represented as probabilistic graphical models. We can incorporate multiple probabilistic graphical models into the same structure by the use of a (qualitatively stated) theorem: there is no mathematical difference between (i) determining which of several probabilistic graphical models is most probable; and (ii) determining the most probable value of a new variable (*Category*) that ranges over the possible categories, and is a graphical parent of the relevant features. That is, given several probabilistic graphical models, there is a mathematically equivalent *single* structure with a new, unobserved *Category* variable that acts as a switch to produce the appropriate probabilistic graphical model depending on which category is actual. In the Bayes net literature, this unobserved variable is a context

variable for a model with *context-specific independence* (CSI; see Boutilier, Friedman, Goldszmidt, & Koller, 1996; Poole & Zhang, 2003; Zhang & Poole, 1999). Thus, instead of modeling categorization as calculations of $P(Category\ A \mid X)$, we can equivalently model it as calculations of $P(CATEG = A \mid X)$ for a context variable *CATEG* whose values are the various mutually exclusive possible categories. This context variable is similar to the "Being a…" variables of Strevens's (2000) minimal essentialism: the context (category) determines the structure (e.g., causal laws) for an individual, but no claims are made about *how* the context (category) does so.

As an example of these two ways of thinking about the same (theoretical, mathematical) process, consider the category of 'blicket' studied by Gopnik and her colleagues (e.g., Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004). Blickets are objects that cause a particular machine (a 'blicket detector') to light up and play music. Thus, the category structure for blickets is represented as a causal Bayes net (since it is a causal model), and is given in the left-hand side of Figure 4. The (possibly implicit) contrast class of '*not*-blickets' are all of the things that fail to activate the detector; that category's causal structure is given in the right-hand side of Figure 4. The equivalent CSI causal Bayes net is given in Figure 5, where the probability of *DetectorActivates* only depends on the value of *OnDetector* if the context node *Category* has the value 'blicket'. Purely as a visual aid, context variables will be indicated by a dashed circle. The theoretical equivalence here implies that there is no mathematical difference between calculating (i) $P(\text{left-hand structure} \mid X)$ in Figure 4; and (ii) $P(Category = \text{blicket} \mid X)$ in Figure 5.

[Insert Figure 4 about here]

[Insert Figure 5 about here]

The story gets a bit more complicated when we allow for the possibility that the different categories could have different underlying structures. In the above example, we could represent both category structures as causal Bayes nets. Suppose instead that one category structure is represented by a Markov random field (a second-order prototype-based category), and the other by a Bayes net (perhaps with Figure 2 graph). There is a third type of probabilistic graphical model—chain graphs—which contain both directed and undirected edges (Andersson, Madigan, & Perlman, 1996; Lauritzen & Richardson, 2002; Lauritzen & Wermuth, 1989; Richardson, 1998). The precise interpretation of chain graphs with mixtures of edge types is the subject of current research (Lauritzen & Richardson, 2002), but Markov random fields and Bayes nets emerge as special cases in which either all of the edges are undirected, or all are directed. Thus, all of the probabilistic graphical models used in this chapter can themselves be unified in a single framework, and that framework can also exploit the CSI (context-specific independence) equivalence between (i) multiple structures, and (ii) a single structure with a context variable.

This picture opens intriguing possibilities for developing integrated hierarchies of multiple category types (when the categories actually are hierarchical; see Malt & Johnson, 1992 and Sloman, 1998 for doubts about this condition).[xi] Suppose we have a set of mutually exclusive categories (e.g., ‘dog’, ‘cat’, ‘mouse’, ‘human’, etc.) that are complete relative to some super-category (e.g., ‘mammal’), so every instance of the super-category can be placed into exactly one of the target categories. Then the context (category) variable for that set corresponds to the super-category, and will be a graphical parent of any feature that is part of one of the category models. Since the GCM, prototype, and causal model categories can all be represented as probabilistic graphical models, a single CSI chain graph model can account for the possibility that these categories do not have the same structure. No special difficulties arise if, for example,

'dog' is a causal model category, 'cat' is an second-order prototype (i.e., Markov random field) category, and 'human' is an exemplar-based GCM (i.e., Bayes net with Figure 2 graph) category. The resulting single graphical structure might look something like Figure 6. (Recall that context variables are indicated by a dashed circle for ease of presentation.) Note the undirected edges between features, indicating the association between *# of Legs* and *Vocalization* in the categories.

[Insert Figure 6 about here]

Representing mutually exclusive categories in a single graph provides one picture of how multiple categories could be cognitively represented in a single category structure. Moreover, because this unified model does not require us to choose between the various psychological theories, it inherits their explanatory power (though no account has been given of why a category is represented using a particular structure). The proposal here is thus consistent with previous data supporting these psychological theories. In addition, this unified model provides a plausible mechanism for including one type of prior knowledge about contrast class. Background knowledge about a situation (e.g., I am on land, so all possible animals must be capable of living on land) is equivalent to conditioning on one or more features prior to categorization, which will change the prior distribution of category probabilities. In this example, P(*Category* = 'Whale') in Figure 6 will be very low, even before I observe any features of a particular instance.

Further hierarchies of categories can be straightforwardly modeled in this theory by introducing a node that is a parent (or neighbor) of the *Category* node. The values of this new variable will range over the super-class encoded in *Category*, as well as the categories that are complete and mutually exclusive for the super-class level of the hierarchy. The new variable (e.g., *AnimalType* in Figure 7) is the CSI context variable for its children (e.g., *BirdType*, *MammalType*, etc.). Moreover, since the super-class variable (e.g., *AnimalType*) is itself a node,

it can have various observed features as its (graphical) children, in addition to other context variables. One plausible resulting model structure is shown in Figure 7, where context variable names have been made more descriptive, and obviously not all relevant variables are included. (Inter-context connections are indicated by dashed edges. As with the dashed circles for context variables, this notation is intended purely as a visual aid.)

[Insert Figure 7 about here]

By attaching features to the super-class in addition to the sub-class, we can arguably explain two contrasting phenomena: (i) some properties (e.g., lactation) are more readily identified with the super-class (mammal) than with any of the sub-classes (e.g., dog); while (ii) the presence of those features nevertheless increases the likelihood of the sub-classes. These two phenomena can only be explained by a framework in which (i) features can be attached (in some sense) directly to the super-class; and (ii) there are meaningful connections between the super-class and the sub-class(es). In this model, *Lactates* is directly associated with *AnimalType* = mammal rather than any specific mammal, but the *Lactates* = Yes observation increases the probability that *AnimalType* = mammal, thereby increasing the (unconditional) probability that the individual has each possible *MammalType*. Thus, in this (potential) unified framework, we have a precise, mathematically well-specified representations of hierarchies of multiple specific types of category structure, where no particular level of the hierarchy is privileged *a priori*. Of course, the empirical adequacy of this admittedly quite complicated unified framework remains a substantial open question.

Conclusion

The central theoretical results of this chapter provide a common language for several major psychological theories of categorization: Generalized Context Models as Bayesian networks with a particular graphical structure; second-order prototype models as Markov random fields; and causal model categories as causal Bayesian networks. The expression of these theories as probabilistic graphical models opens up a number of methodological and theoretical possibilities. We can readily determine why some experimental designs are unable to distinguish among these theories on the basis of model fits, and so design better, more discriminating experiments. In particular, we can construct category structures which cannot be represented perfectly by any of the psychological theories to determine which theory (if any) correctly predicts people's systematic errors. These equivalencies also point towards natural generalizations of the psychological theories, corresponding to elimination of various non-graphical constraints on the probability distributions. These generalizations can easily be tested; for example, we could examine people's performance on categories that can be modeled by an arbitrary Markov random field, but not one that corresponds to an second-order prototype model. Moreover, this work suggests a different understanding of the categorization process, and not just similarity ratings.

Categorization judgments in these psychological theories all correspond to calculating $P(Category \mid \text{novel instance } X)$ when every considered category is equiprobable *via* a two-stage process: calculate the similarity ratings (i.e., $P(X \mid Category)$ for each category), and then use the Shepard-Luce choice rule to get $P(Category \mid X)$. In contrast, we can consider a categorization theory that directly computes $P(Category \mid X)$, without explicitly computing similarity ratings as an intermediate step. Experimental elicitations of similarity ratings are, on this model, just categorization judgments relative to an implicit, unspecified contrast class. There is some

preliminary evidence for the idea that all judgments, even typicality ones, are relative to a contrast class. However, significantly more experimental investigation is required. Finally, we can represent all of these probabilistic graphical models as chain graphs, and place them into a single graphical model hierarchy by exploiting various context-specific independencies. This process unifies multiple category-types into a single, coherent graphical structure.

But the results described here matter for more than just categorization research. There is clearly a close interdependence between (at least some) causal cognition and (some) categorization. Causal beliefs and learning shape some of our categorization decisions, including categorizing some novel object, and selecting or learning categories for representing the world. In the other direction, our causal learning and decision-making depends upon both the ways in which we categorize our world, and the properties we infer about novel objects based on their category membership. Despite these connections, much of the research in the two fields, whether psychological, philosophical, or computational, has essentially ignored the importance of the other. Frequently, substantial allusions are made to the importance of the other field, but then the other field plays no theoretical or experimental role. Given the scope of these cognitive processes, this has been a reasonable research strategy: at least initially, we should divide-and-conquer. A range of recent psychological and theoretical research has started to shift this trend, but a common mathematical framework is needed for the two domains. The equivalencies detailed in this chapter thus provide an important early step for the integration of research on categorization and causation: the representation of a major component of categorization theories as probabilistic graphical models, which are the emerging consensus framework for modeling large portions of causal learning and inference.

Acknowledgments

This chapter has been greatly influenced by conversations with Bob Rehder and Chris Hitchcock. Clark Glymour, Alison Gopnik, Edouard Machery, and Laura Schulz all provided valuable comments on earlier drafts. Thanks also to the audience at the February 2004 CASBS workshop on Causation and Categorization for their insightful questions and comments.

References

Ahn, W.-K., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition, 30*, 107-118.

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.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1996). An alternative Markov property for chain graphs. In F. V. Jensen & E. Horvitz (Eds.), *Proceedings of the 12th conference on uncertainty in artificial intelligence* (pp. 40-48). San Francisco: Morgan Kaufmann.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216-233.

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

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400.

Ashby, F. G., & Waldron, E. M. (2000). The neuropsychological bases of category learning. *Current Directions in Psychological Science, 9*, 10-14.

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.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T.K. Srull & R.S. Wyer (Eds.), *Advances in social cognition, volume III: Content and process specificity in the effects of prior experiences* (pp. 61-88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Boutilier, C., Friedman, N., Goldszmidt, M., & Koller, D. (1996). Context-specific independence in Bayesian networks. In *Proceedings of the twelfth annual conference on uncertainty in artificial intelligence* (pp. 115-123). San Francisco: Morgan Kaufmann Publishers.

Danks, D. (2004). Psychological theories of categorization as probabilistic models. Technical report CMU-PHIL-149. Department of Philosophy, Carnegie Mellon University.

Danks, D. (under review). Psychological theories of categorization as probabilistic graphical models. Submitted to *Journal of Mathematical Psychology*.

Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics, 8,* 522-539.

Das-Smaal, E. A., & De Swart, J. H. (1986). Effects of contrasting category, conjoint frequency and typicality on categorization. *Acta Psychologica, 62,* 15-40.

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3-32.

Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., & McMillan, C. (2002). The neural basis for categorization in semantic memory. *NeuroImage, 17*, 1549-1561.

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science, 28*, 45-74.

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

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

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 695-711.

Lamberts, K. (2000). Information accumulation theory of categorization response times. *Psychological Review, 107*, 227-260.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretation. *Journal of the Royal Statistical Society, Series B, 64*, 321-361.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics, 17*, 31-57.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*, 87–137.

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology, 55*, 207-234.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 103–189). New York: Wiley.

Machery, E. (in press). Concepts are not a natural kind. *Philosophy of Science*.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language, 31*, 195–217.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*, 1-13.

Minda, J. P., & Smith, J. D. (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.

Myung, I. J. (1994). Maximum entropy interpretation of decision bound and context models of categorization. *Journal of Mathematical Psychology, 38*, 335-365.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. Prentice Hall.

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

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34*, 393-418.

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). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 936-941.

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1*, 360-370.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Poole, D., & Zhang, N. L. (2003). Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research, 18*, 263-313.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

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

Rehder, B., & Burnett, R. (in press). Feature inference and the causal structure of categories. *Cognitive Psychology*.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition, 91*, 113-153.

Richardson, T. S. (1998). Chain graphs and symmetric associations. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 231-259). Cambridge, Mass.: The MIT Press.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, Mass.: The MIT Press.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology, 46*, 178-210.

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

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*, 1-33.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1411-1436.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. 2nd edition, 2001. Cambridge, Mass.: The MIT Press.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science, 71*, 833-845.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition, 74*, 149-175.

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.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology, 53A*, 983-1011.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment.” *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1160-1173.

Zhang, N. L., & Poole, D. (1999). On the role of context-specific independence in probabilistic inference. In *Proceedings of the sixteenth international joint conference on artificial intelligence* (pp. 1288-1293). San Francisco: Morgan Kaufmann Publishers.

Captions

Figure 1: Example Bayesian network

Figure 2: Bayesian network analogue for the Generalized Context Model

Figure 3: Example Markov random field

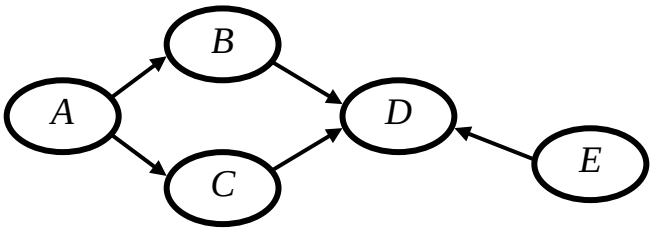
Figure 4: (a) Causal category structure for blickets; (b) Causal category structure for *not*-blickets

Figure 5: CSI Bayesian network for the blicket category

Figure 6: One-layer category structure

Figure 7: Multi-layer category structure

Figure 1



$$P(A, B, C, D, E) = P(A) \times P(B \mid A) \times P(C \mid A) \times P(D \mid B, C, E) \times P(E)$$

Figure 2

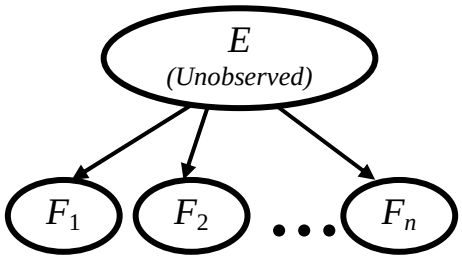
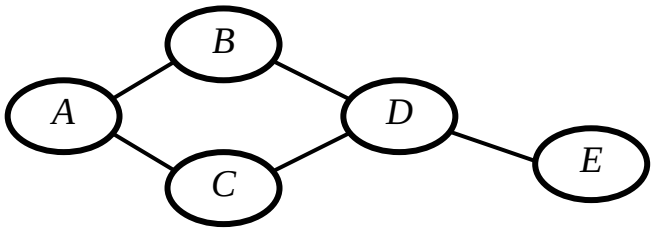
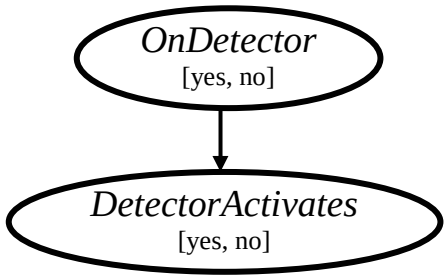


Figure 3

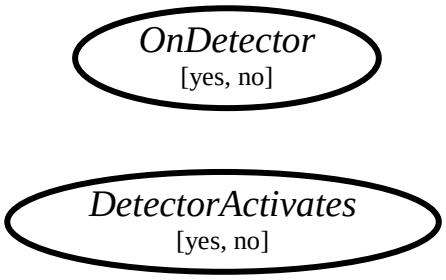


$$P(A, B, C, D, E) = G_1(A, B) \times G_2(A, C) \times G_3(B, D) \times G_4(C, D) \times G_5(D, E)$$

Figure 4



P(*DetectorActivates* = Yes | *OnDetector* = Yes) = 1
P(*DetectorActivates* = Yes | *OnDetector* = No) = 0



P(*DetectorActivates* = Yes) = 0

Figure 5

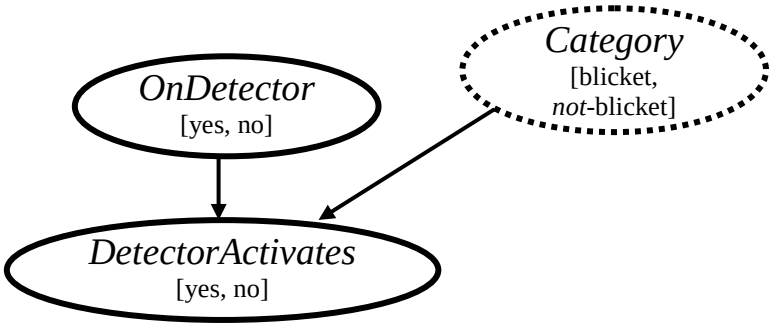

Category
[blicket,
not-blicket]
OnDetector
[yes, no]
DetectorActivates
[yes, no]

Figure 6

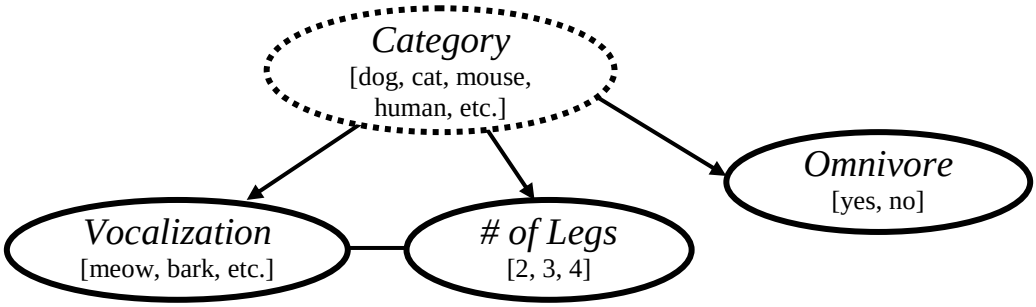
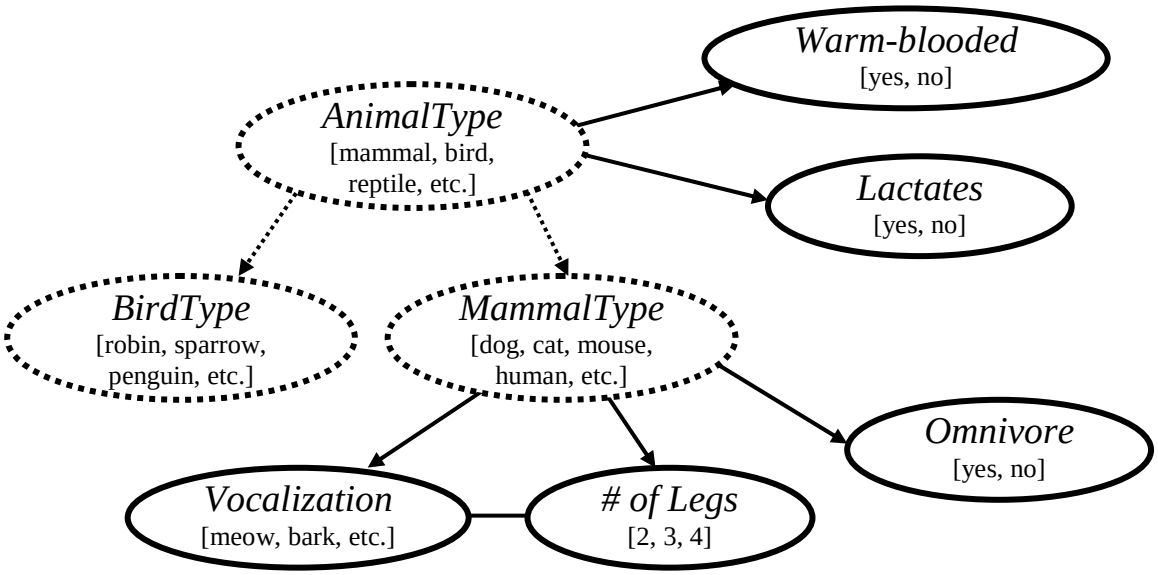

Category
[dog, cat, mouse, human, etc.]
Omnivore
[yes, no]
Vocalization
[meow, bark, etc.]
# of Legs
[2, 3, 4]

Figure 7


Warm-blooded
[yes, no]
AnimalType
[mammal, bird, reptile, etc.]
Lactates
[yes, no]
BirdType
[robin, sparrow, penguin, etc.]
MammalType
[dog, cat, mouse, human, etc.]
Omnivore
[yes, no]
Vocalization
[meow, bark, etc.]
# of Legs
[2, 3, 4]

[i] Many-valued features can be modeled as multiple binary features as in the SUSTAIN model (e.g., Love, Medin, & Gureckis, 2004), though at some computational cost. Note a feature is binary or continuous based on values that it might plausibly take, and not the actually observed values. For example, height is continuously-valued even though, in real life, we only see finitely many values for height in a population.

[ii] We thus collapse together, e.g., Erickson & Kruschke's (1998) ATRIUM model, Kruschke's (1992) ALCOVE model, Lamberts's (1998, 2000) EGCM model, and Nosofsky & Palmeri's (1997) EBRW, which are all equivalent to Nosofsky's (1986) Generalized Context Model for static problems. An interesting open question is whether the equivalencies described here can be used to understand the theoretic relationships among the response time models.

[iii] The GCM is equivalently (and more typically) expressed as the exponential of a sum of distances, rather than the product of exponentials used here. The GCM also allows for other distance measures (e.g., Euclidean); the equivalencies described in the next section continue to hold for other distance measures (though with different auxiliary conditions).

[iv] Early prototype models assumed that the similarity ratings were the *sum* of the distance on each dimension, rather than the product. Subsequent work has shown these additive models to be worse than multiplicative ones (see, e.g., Minda & Smith, 2001).

[v] More precisely, there must be a Bayes net in which the graph is a perfect map for (i.e., is Markov and Faithful to) the probability distribution.

[vi] It does not actually matter whether the graphical structure is treated as a Bayes net or a Markov random field (where every directed edge is converted to an undirected edge). The set of probability distributions is the same.

[vii] There is an apparent tension here between: (i) the condition of a regularity constraint on the $P(F_i \mid E)$ terms (suggesting that the GCM cannot model all probability distributions); and (ii) Ashby & Alfonso-Reese's (1995) proof that the GCM can, in the limit of infinitely many exemplars, model any probability distribution. Given that the notion of "infinitely many exemplars" is psychologically unrealistic, the equivalencies described here require the category to have only as many exemplars as E (the unobserved variable) has values. That restriction results in the need for the regularity constraint.

[viii] Using the Lauritzen-Wermuth-Frydenberg chain graph Markov property, this graph implies: $F_1 \perp\!\!\!\perp F_4$; $F_1 \perp\!\!\!\perp F_3 \mid \{F_2, F_4\}$; and $F_2 \perp\!\!\!\perp F_4 \mid \{F_1, F_3\}$.

[ix] Proof: Using the equivalencies in the Luce choice rule yields: $P(\text{respond "}A\text{"} \mid X) = P(X \mid A) / \Sigma P(X \mid M)$. Multiply the top and bottom by $P(A)$, and use $P(A) = P(M)$ for all M to reduce to $P(X \mid A) \times P(A) / P(X)$, which is just $P(A \mid X)$.

[x] This type of theory is frequently called 'Bayesian,' particularly by computer scientists and statisticians. I will avoid using that term here, since for many philosophers and psychologists, 'Bayesian' includes a much larger set of commitments than this theory requires.

[xi] The following discussion is quite qualitative, but the proposed framework has a precise mathematical description.