An integrated account of generalization across objects and features

Citation: Kemp, Charles, Patrick Shafto, and Joshua B. Tenenbaum. "An Integrated Account of Generalization across Objects and Features." Cognitive Psychology 64, no. 1–2 (February 2012): 35–73.

As Published: http://dx.doi.org/10.1016/j.cogpsych.2011.10.001

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/98843

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

An integrated account of generalization across objects and features[☆]

Charles Kemp[a,*], Patrick Shafto[b], Joshua B. Tenenbaum[c]

[a]*Department of Psychology*
Carnegie Mellon University
[b]*Department of Psychological and Brain Sciences*
University of Louisville
[c]*Department of Brain and Cognitive Sciences*
Massachusetts Institute of Technology

Abstract

Humans routinely make inductive generalizations about unobserved features of objects. Previous accounts of inductive reasoning often focus on inferences about a single object or feature: accounts of causal reasoning often focus on a single object with one or more unobserved features, and accounts of property induction often focus on a single feature that is unobserved for one or more objects. We explore problems where people must make inferences about multiple objects and features, and propose that people solve these problems by integrating knowledge about features with knowledge about objects. We evaluate three computational methods for integrating multiple systems of knowledge: the output combination approach combines the outputs produced these systems, the distribution combination approach combines the probability distributions captured by these systems, and the structure combination approach combines a graph structure over features with a graph structure over objects. Three experiments explore problems where participants make inferences that draw on causal relationships between features and taxonomic relationships between animals, and we find that the structure combination approach provides the best account of our data.

Keywords:
generalization, property induction, causal reasoning, knowledge integration

1. Introduction

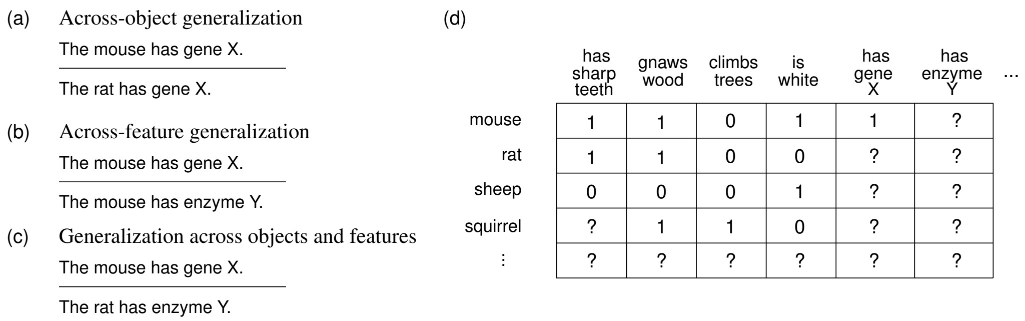
Will that berry taste good? Is that table strong enough to sit on? Questions like these require a reasoner to predict whether an object has a feature that has not yet been observed. Two versions of this basic inductive challenge can be distinguished. *Across-object generalization* is a problem where a reasoner observes one or more objects that have a given feature (e.g. Tim has feature F) then decides whether other objects have the same feature (does Tim's twin brother Tom have feature F?). *Across-feature generalization* is a problem where a reasoner observes one or more features of a given object (e.g. Tim is obese) then makes inferences about other features of the same object (does Tim have diabetes?). These two generalization problems form a natural pair, and both can be viewed as inferences about the missing entries in a partially-observed object-feature matrix. Figure 1 shows an example where the objects are animals of different species and the features are biological or behavioral attributes. Because the mouse and the rat are similar, observing that the mouse has gene X suggests that the rat is likely to to carry the same gene (across-object generalization). If gene X causes enzyme Y to be expressed, then observing that the mouse has gene X suggests that the mouse is likely to express enzyme Y (across-feature generalization).

Across-object and across-feature generalization are typically studied in isolation but these two forms of generalization often interact. For example, given that Tim is obese, we might predict that Tim's twin brother Tom is more likely to have diabetes than an unrelated individual called Zach. This prediction appears to rely on across-object generalization (since Tim is obese, Tom is likely to be obese) and on across-feature generalization (if Tom is obese, then Tom is likely to have diabetes). Similarly, if we learn that the mouse in Figure 1a carries gene X and that gene X causes enzyme Y

[☆] An early version of this work was prepared in collaboration with Allison Berke and presented at the Neural Information Processing Systems conference in 2006. We thank Keith Holyoak and three anonymous reviewers for helpful comments on the manuscript, and Bob Rehder and Brian Milch for valuable discussions. This work was supported in part by AFOSR MURI contract FA9550-05-1-0321, NSF award CDI-0835797, the William Asbjornsen Albert memorial fellowship (CK) and the Paul E. Newton Chair (JBT).

[*] Corresponding author

Email addresses: ckemp@cmu.edu (Charles Kemp), p.shafto@louisville.edu (Patrick Shafto), jbt@mit.edu (Joshua B. Tenenbaum)



(a) Across-object generalization

The mouse has gene X.

The rat has gene X.

(b) Across-feature generalization

The mouse has gene X.

The mouse has enzyme Y.

(c) Generalization across objects and features

The mouse has gene X.

The rat has enzyme Y.

(d)

	has sharp teeth	gnaws wood	climbs trees	is white	has gene X	has enzyme Y	...
mouse	1	1	0	1	1	?	
rat	1	1	0	0	?	?	
sheep	0	0	0	1	?	?	
squirrel	?	1	1	0	?	?	
⋮	?	?	?	?	?	?	

Figure 1: Generalization problems involving a set of animals and their features. (a) Across-object generalization is a problem where a reasoner makes inferences about the distribution of a single feature—here "has gene X." The example shown is a one premise argument: given that the statement above the line is true, the reasoner must decide whether the statement below the line is likely to be true. (b) Across-feature generalization is a problem where a reasoner makes inferences about the features of a single object. The argument shown here is strong if gene X is known to cause enzyme Y to be expressed. (c) Generalization problems may require a reasoner to generalize across both objects and features. Here the reasoner is told that a given animal (the mouse) has a given feature (gene X), then asked to decide whether a different animal (the rat) has a different feature (enzyme Y).(d) Generalization can be formalized as the problem of filling in the missing entries in an object-feature matrix. The three problems in (a)–(c) are all special cases of this matrix completion problem.

to be expressed, we might predict that the rat is likely to express enzyme Y (Figure 1c). Both of these predictions can be formulated as inferences about the missing entries in an object-feature matrix. We develop an account of generalization that handles inferences of this kind, and that includes both across-object and across-feature generalization as special cases.

Our approach is based on the idea of integrating multiple knowledge structures. An *object structure* can capture relationships among objects—for example, a structure defined over the three individuals previously introduced can indicate that Tim and Tom are more similar to each other than either is to Zach. A *feature structure* can capture relationships between features—for example, one feature structure might indicate that obesity tends to cause diabetes. We show how object and feature structures can be combined in order to reason about the missing entries in a partially-observed object-feature matrix.

Previous researchers have explored both object structures and feature structures, but most previous models work with just one kind of structure at a time (Figure 2a). Accounts of across-feature generalization (Waldmann et al., 1995; Ahn et al., 2000; Rehder, 2003) often use a structure that focuses exclusively on causal relationships between features. For example, a conventional causal model might specify that obesity causes diabetes without capturing any information about relationships between objects. To see the limitations of this approach, suppose that Tim, Tom and Zach are all obese, that Tim and Tom are identical twins, and that Tim has diabetes. Since Tom and Zach are both obese, the conventional model will predict that both men are equally likely to suffer from diabetes. It seems clear, however, that the causal relationship between obesity and diabetes is mediated by hidden causal factors, and that Tom and Tim are similar with respect to these factors. Since Tim's obesity led to diabetes, Tom's obesity is likely to have a similar effect, and we might therefore predict that Tom is more likely than Zach to suffer from diabetes.

Accounts of across-object generalization (also known as property induction, category-based induction, or stimulus generalization (Shepard, 1987; Osherson et al., 1990; Sloman, 1993; Heit, 1998; Hayes et al., 2010)) often work with a structure that focuses exclusively on relationships between categories or objects. For example, Kemp & Tenenbaum (2009) present a model that uses a tree-structured representation of relationships between animals in order to account for inferences about blank biological features (e.g. "has enzyme X"). Models of this kind, however, are unable to reason about features that are causally related to known features. Suppose, for example, you learn that whales "travel in a zig-zag trajectory" and need to decide whether bears or tuna are more likely to share this feature (Heit & Rubinstein, 1994). A model that relies on taxonomic relationships alone is likely to prefer bears, but a model that incorporates causal relationships between features might choose tuna on the basis that "traveling in a zig-zag trajectory" is related to other features like swimming and living in the water.

Accounts that rely on feature structures or object structures in isolation are fundamentally limited, but we show that combining these structures can lead to a more comprehensive account of generalization. Like many previous accounts of generalization, we take a probabilistic approach (Shepard, 1987; Anderson, 1991; Heit, 1998; Rehder, 2003; Kemp & Tenenbaum, 2009; Holyoak et al., 2010). Probability theory alone, however, does not specify how different knowledge structures should be combined, and we evaluate several alternatives. The *output combination* approach (OC approach for

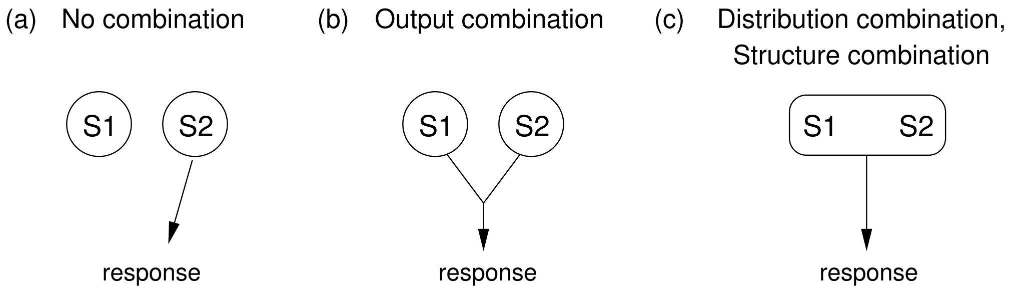


Figure 2: Approaches to generalization problems where two relevant systems of knowledge are available. (a) A response is generated that depends on only one of the systems. (b) A response is generated by using a simple mathematical function such as a weighted average to combine the outputs generated by each system in isolation. (c) The systems themselves are combined to generate a response. We consider two versions of this approach: the distribution combination model combines systems at the level of probability distributions, and the structure combination model combines systems at the level of graph structures.

short) combines two knowledge structures by combining the outputs that they produce (Figure 2b). This approach is related to previous accounts of knowledge integration that rely on simple mathematical functions such as sums and products to combine the predictions of multiple models (Medin & Schaffer, 1978; Anderson, 1981; Massaro & Friedman, 1990; Lombardi & Sartori, 2007), and is appropriate when the two knowledge structures correspond to independent modules (Fodor, 1983).

If the knowledge structures do not correspond to separate modules, the two may be combined more directly (Figure 2c). We consider two possibilities. The *distribution combination* approach (DC approach for short) combines two knowledge structures by multiplying the prior distributions that they capture. Multiplying prior distributions provides a probabilistic way to capture the intuition that an outcome is likely only if it is consistent with both component knowledge structures. The *structure combination* approach (SC approach for short) combines two knowledge structures by creating a third structure that corresponds to a graph product of the two component structures. One important difference between these approaches is that the DC approach predicts that object and feature structures are combined in a way that is not intrinsically causal. In contrast, the SC approach leads to a model that is defined over a causal graph and that therefore supports inferences about interventions and counterfactuals. Our experiments suggest that humans combine object and feature structures in a way that supports subsequent causal inferences, and we therefore conclude that the SC approach accounts for human inferences better than the DC approach.

Although previous models of inductive reasoning do not incorporate both feature structures and object structures, several researchers have explored whether these forms of knowledge are combined. Experiment 3 of Rehder (2006) suggests that causal relationships between features dominate similarity relationships between objects, and that similarity relationships are used only when causal information is unavailable. Other researchers have also considered cases where causal inferences and inferences based on surface similarity lead to opposite conclusions (Lassaline, 1996; Wu & Gentner, 1998; Hayes & Thompson, 2007; Lee & Holyoak, 2008; Holyoak et al., 2010), and the consistent finding is that causal inferences dominate similarity-based inferences. Causal relationships may indeed be primary, but the vast majority of real-world problems involve cases where causal relationships between features are known only partially. In cases of this kind similarity relationships between objects provide a guide to shared causal structure, and inferences should therefore exploit both causal relationships between features and similarity relationships between objects. Hadjichristidis et al. (2004) provide some evidence for this view, and show that inductive inferences are influenced both by the centrality of a feature in a causal structure and by similarity relationships between objects. Although we do not focus on feature centrality, our model can be viewed as a computational treatment of some key intuitions behind the work of Hadjichristidis et al. (2004). In particular, the model captures the idea that the taxonomic relationships between two objects can help to predict whether the two are similar with respect to unobserved causal variables.

We begin in the next section by introducing a general probabilistic framework for reasoning about partially observed object-feature matrices. The three combination models (OC, DC, and SC) all rely on prior distributions over matrices, and we show how priors of this kind can capture relationships between objects and relationships between features. The remainder of the paper describes three experiments that we conducted to evaluate our models. The results suggest that people are able to reason simultaneously about relationships between objects and relationships between features, and to make causal inferences that draw on both kinds of relationships. We demonstrate that the SC model accounts better for this ability than the OC and DC models, and that all three combination models perform better than alternatives that rely on a feature structure or object structure in isolation.

2. Bayesian generalization

Our account of inductive generalization is founded on Bayesian inference. Any Bayesian approach relies on a hypothesis space, a prior distribution that captures the background knowledge relevant to a given problem, and a general purpose inference engine. This section focuses on the inference engine, and the following sections describe how the prior distribution can capture knowledge about objects and knowledge about features.

Suppose that we are interested in a certain set of objects and a certain set of features. Let M be a complete object-feature matrix—a matrix that accurately specifies whether each object has each of the features. Figure 3 shows the sixteen possible object-feature matrices for a problem involving two objects (o_1 and o_2) and two binary features (f_1 and f_2). Suppose, for example, that o_1 and o_2 are two cousins, and that the features indicate whether the cousins are obese (f_1) and whether they have diabetes (f_2). It seems clear that the sixteen matrices M in Figure 3 are not equally probable *a priori*. For example, the case where both cousins have diabetes but only o_1 is obese seems less probable than the case where both have diabetes and both are obese. Assume for now that the prior probability $P(M)$ of each matrix is known. The prior in Figure 3 captures the idea that f_2 is probabilistically caused by f_1, and therefore tends to be present if f_1 is present. The specific probabilities shown are consistent with a causal model which indicates that obesity has a base rate of 0.3 and causes diabetes with probability 0.15.

Suppose that we observe M_{obs}, a version of the true matrix M with many missing entries. In Figure 3b, M_{obs} indicates that the first cousin is obese. Even though M_{obs} is incomplete we can use it to make inferences about all of entries in M. For example, learning that the first cousin is obese should make us more likely to believe that the first cousin has diabetes, and that the second cousin is obese. We propose that inferences of this kind are based on probabilistic reasoning. These inferences can be modeled using the posterior distribution $P(M|M_{obs})$, which captures expectations about M after observing M_{obs}. Using Bayes' rule, we rewrite this distribution as

$$P(M|M_{obs}) \propto P(M_{obs}|M)P(M). \tag{1}$$

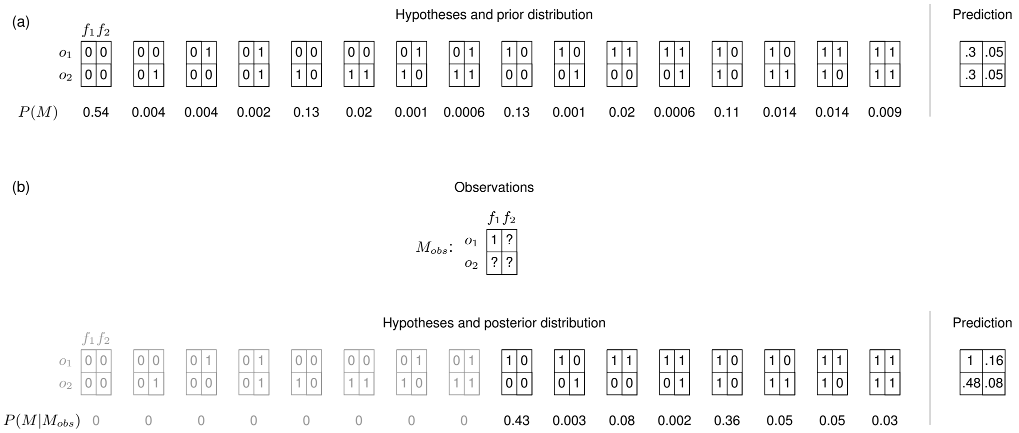


Figure 3: Bayesian generalization. (a) If there are two objects and two features, there are 16 possible binary matrices M. The figure shows one possible prior distribution $P(M)$ over these matrices. The prediction matrix on the right shows predictions about individual cells in the matrix computed by summing over the space of hypotheses. (b) Suppose that we observe the information shown in M_{obs}: we learn that o_1 has f_1. Eight of the matrices are no longer possible and are shown in gray. The posterior distribution $P(M|M_{obs})$ is computed by reweighting the prior distribution $P(M)$ on the eight matrices that remain. Relative to the prediction in (a), the prediction matrix now indicates that o_1 is more likely to have f_2 and that o_2 is more likely to have f_1.

The likelihood term $P(M_{obs}|M)$ will depend on how the entries in M_{obs} were generated. Different formulations of this term can capture, for instance, whether the observations in M_{obs} are corrupted by noise, and whether a systematic method is used to select the occupied entries in M_{obs}. We will assume throughout that M_{obs} is created by randomly choosing some entries in M then revealing their true values. It follows that

$$P(M_{obs}|M) \propto \begin{cases} 1, & \text{if } M_{obs} \text{ is consistent with } M \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where M_{obs} is consistent with M if every entry that appears in M_{obs} matches the corresponding entry in M.

Combining Equation 2 with Equation 1 we see that

$$P(M|M_{obs}) \propto \begin{cases} P(M), & \text{if } M_{obs} \text{ is consistent with } M \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where the prior $P(M)$ captures our prior expectations about matrix M. Intuitively, Equation 3 states that any matrix M which is incompatible with the observations in M_{obs} has zero posterior probability, and that the posterior distribution over the candidates that remain is computed by reweighting the prior $P(M)$ (Figure 3).

The posterior distribution in Equation 3 can be used to make predictions about individual entries in matrix M. Suppose, for example, that we are primarily interested in entry M_{ij}, or the entry that indicates whether object i has feature j. The probability that this entry equals 1 is equal to the combined posterior probability of all matrices with a 1 in position (i, j):

$$P(M_{ij} = 1|M_{obs}) = \sum_{M:M_{ij}=1} P(M|M_{obs}) \qquad (4)$$

where the sum ranges over all candidate matrices M with a 1 in position (i, j). For example, the prediction matrix in Figure 3 indicates that $P(M_{12} = 1|M_{obs}) = 0.08 + 0.002 + 0.05 + 0.03 \approx 0.16$.

The Bayesian computations specified by Equations 3 and 4 are straightforward applications of statistical inference. Statistical inference is a general-purpose approach that can be applied across many different settings, and has previously been used to develop psychological accounts of property induction (Heit, 1998; Kemp & Tenenbaum, 2009), stimulus generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001), word learning (Xu & Tenenbaum, 2007), categorization (Anderson, 1991; Sanborn et al., 2010), identification (Kemp et al., 2010) and causal learning (Anderson, 1990; Griffiths & Tenenbaum, 2005). These different applications may address very different phenomena, but all of them use statistical inference to explain how prior knowledge and observed data combine to produce inductive inferences.

Even though the Bayesian approach emphasizes domain-general statistical inference, it recognizes that differences between inductive problems are critical. The prior distribution plays a fundamental role in any Bayesian model, and different prior distributions can capture the different kinds of knowledge that are relevant to different inductive problems. The next sections focus on the prior distribution $P(M)$ that plays a role in Equation 3. Formalizing this prior will require us to think carefully about the knowledge that guides generalization.

3. Knowledge structures

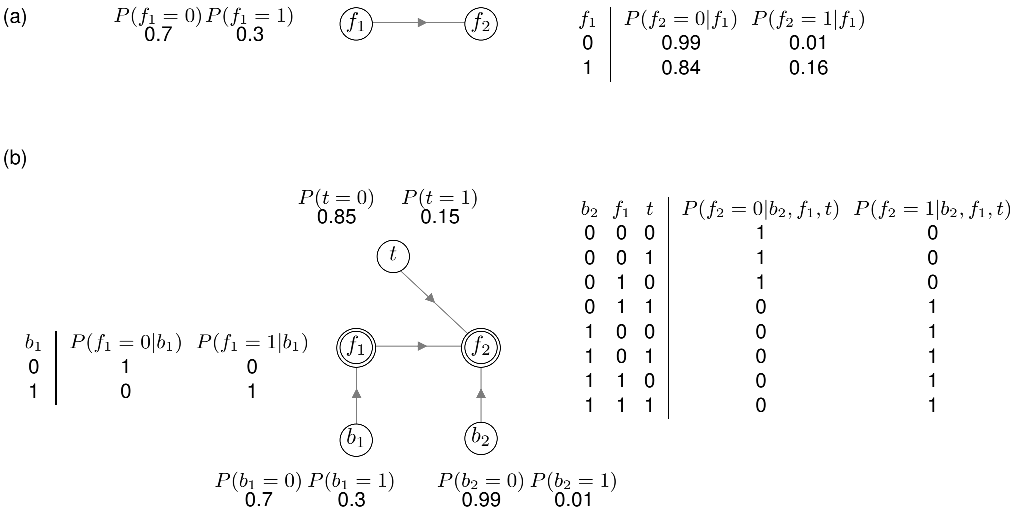
A reasoner may know about relationships between objects, relationships between features, and relationships between objects and features, and each kind of knowledge is useful for making inferences about the missing entries in an object-feature matrix. The prior $P(M)$ can capture all three kinds of knowledge. For example, suppose that M is a matrix that specifies the features of a set of pets. A reasoner might know that his pet mouse and his pet rat are similar (a relationship between objects), and might assign low prior probability to matrices where his mouse and rat have many different features. A reasoner might know that having sharp teeth and gnawing wood are linked (a relationship between features), and might assign low prior probability to matrices where many animals gnaw wood but do not have sharp teeth. Finally, the reasoner might know that his pet squirrel gnaws wood (a relationship between an object and a feature) and might assign low prior probability to any matrix that violates this condition.

We will work towards a prior distribution that simultaneously captures relationships between features and relationships between objects. Let F be a structure that captures relationships between features. For instance, F might be a causal model which specifies a causal relationship between having sharp teeth and gnawing wood. Let O be a structure that captures relationships between objects. For instance, O might be a model of similarity which indicates that mice and rats are similar. The next sections describe priors $P(M|F)$ and $P(M|O)$ that rely on a single structure, and we then consider priors $P(M|F, O)$ that take both structures into account.

3.1. Feature structures

Inferences about partially-observed object-feature matrices can draw on different kinds of relationships between features. Some of these relationships may capture non-causal correlations—for example, an appliance which is large and is found within the home is likely to be white. Here, however, we focus on causal relationships—for example, an appliance with an engine is likely to be noisy.

Building on previous work in psychology, artificial intelligence and statistics, we will formalize causal knowledge using graphical models, also known as Bayesian networks (Pearl, 2000). Bayesian networks can capture probabilistic relationships between variables—for example, the network in Figure 4a captures a case where feature f_1 (e.g. obesity) probabilistically causes feature f_2 (e.g. diabetes). Note that f_2 is present 16% of the time when f_1 is present, but only 1% of the time when f_1 is absent. Most psychological applications of Bayesian networks have focused on probabilistic causal relationships, but we will work with models that capture deterministic causal relationships. For example, the probabilistic relationship between obesity



$P(f_1 = 0)$	$P(f_1 = 1)$
0.7	0.3

f_1	$P(f_2 = 0 \vert f_1)$	$P(f_2 = 1 \vert f_1)$
0	0.99	0.01
1	0.84	0.16

$P(t = 0)$	$P(t = 1)$
0.85	0.15

b_1	$P(f_1 = 0 \vert b_1)$	$P(f_1 = 1 \vert b_1)$
0	1	0
1	0	1

b_2	f_1	t	$P(f_2 = 0 \vert b_2, f_1, t)$	$P(f_2 = 1 \vert b_2, f_1, t)$
0	0	0	1	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	0	1
1	0	1	0	1
1	1	0	0	1
1	1	1	0	1

$P(b_1 = 0)$	$P(b_1 = 1)$
0.7	0.3

$P(b_2 = 0)$	$P(b_2 = 1)$
0.99	0.01

Figure 4: Causal models. (a) A model that captures a probabilistic relationship between features f_1 and f_2. (b) A functional causal model that induces the same joint distribution over f_1 and f_2. Variables b_1 and b_2 indicate whether background causes for f_1 and f_2 are active, and variable t indicates whether the mechanism of causal transmission between f_1 and f_2 is active. All of the root variables (b_1, b_2 and t) are independent, and the double lines around f_1 and f_2 indicate that these variables are deterministically specified once the root variables are fixed.

and diabetes may be described more accurately as a deterministic relationship that depends on one or more genetic and environmental factors. Figure 4b suggests that the probabilistic relationship between f_1 and f_2 in Figure 4a can be reformulated as a deterministic relationship that depends on variables b_2 and t. Variable b_2 indicates whether some background cause of f_2 is active, and variable t indicates whether or not the mechanism of causal transmission between f_1 and f_2 is active. For example, suppose that fat cells produce a hormone that acts together with a special enzyme to cause diabetes. In this case the transmission variable t might indicate whether or not the special enzyme is present in a given individual. The distributions in Figure 4b show that variables f_1 and f_2 are both deterministic functions of their parents in the graph. For example, f_2 is true only if background cause b_2 is present, or if f_1 is present and the link between f_1 and f_2 is active (i.e. both f_1 and t are true). Note that the distributions in Figure 4b induce exactly the same distribution $P(f_2|f_1)$ that is captured by the model in Figure 4a.

Both models in Figure 4 are specified by defining probability distributions over graph structures. The edges in the graph indicate patterns of causal dependence, and each model specifies a conditional probability distribution on the value of each variable given the values of its parents in the graph. Together, these conditional probability distributions define a joint distribution over the values of all variables. For example, the joint distribution for structure S in Figure 4a is

$$P(f_1, f_2|S) = P(f_1)P(f_2|f_1) \tag{5}$$

This joint distribution can be written more generally as

$$P(v_1, \ldots, v_n|S) = \prod_j P(v_j|\pi(v_j)) \tag{6}$$

where $\pi(v_j)$ indicates the parents of variable v_j, or the set of all variables in S that send an edge to v_j. Variables with no parents will be referred to as *root variables*, and $P(v_j|\pi(v_j))$ is equivalent to the distribution $P(v_j)$ for any root variable v_j.

The models we consider rely on the factorization in Equation 6, but the conditional probability distributions for all variables other than the root variables must be deterministic. Models of this kind are often described as *functional* causal models (Pearl, 2000). Note, for example, that the functional model in Figure 4b specifies distributions $P(b_1)$, $P(b_2)$ and $P(t)$ on the three root variables, but that the distributions $P(f_1|b_1)$ and $P(f_2|f_1, t, b_2)$ are deterministic. At first it may seem that working with deterministic causal relationships is a severe restriction, but any network N that incorporates probabilistic relationships can be replaced by a functional model that is equivalent in the sense that it captures the same distribution over the variables in N. For example, Figures 4a and 4b both capture the same distribution over variables f_1 and f_2.

For our purposes, the primary reason to work with functional causal models is that they provide a natural way to combine causal relationships between features with relationships between objects. For example, the structure combination model developed in a later section is based on the intuition that similar objects (e.g. identical twins) ought to have similar settings for the hidden variables (e.g. genes) that influence observable features (e.g. health outcomes). There are, however, at least two additional reasons why functional models may be appealing. First, functional models are consistent with the proposal that people are causal determinists (Pearl, 2000; Goldvarg & Johnson-Laird, 2001; Luhmann & Ahn, 2005b; Frosch & Johnson-Laird, 2011), and with empirical results which suggest that people often invoke hidden variables to account for causal relationships that may appear to be probabilistic on the surface (Schulz & Sommerville, 2006). Second, Pearl (2000) has shown that functional causal models improve on networks that incorporate probabilistic relationships by providing a natural account of certain kinds of counterfactual inferences. Psychologists continue to debate whether humans are causal determinists (Cheng & Novick, 2005; Frosch & Johnson-Laird, 2011), but our work fits most naturally with the determinist position.

Although many applications of causal models focus on a single object at a time, a causal model F can be used to make predictions about an entire object-feature matrix M. Suppose that the feature values for object i are collected into a vector $\boldsymbol{o_i}$. The causal model F specifies a distribution $P(\boldsymbol{o_i}|F)$ on these vectors using Equation 6, and these distributions can be combined to produce a prior distribution on matrices M:

$$P(M|F) = \prod_i P(\boldsymbol{o_i}|F). \tag{7}$$

Equation 7 assumes that the object vectors $\{\boldsymbol{o_i}\}$ (i.e. the rows of the matrix) are conditionally independent given the feature model F (Figure 5a). Many models of causal reasoning make this assumption of conditional independence (Rehder & Burnett, 2005), and we refer to it as the assumption of *object independence.*

Even though previous causal models are rarely described as models for reasoning about entire object-feature matrices, most can be viewed as approaches that combine the Bayesian framework of Equation 3 with a prior (Equation 7) defined using a feature model. Approaches of this kind have been used to address problems including causal attribution, causal learning, and property induction (Waldmann et al., 1995; Rehder, 2003; Danks, 2003; Gopnik et al., 2004; Sloman, 2005). These approaches, however, suffer from a well-known and serious limitation (Luhmann & Ahn, 2005b,a). In most cases of interest, the feature model F will not capture all of the causally relevant variables, and hidden variables will ensure that the assumption of object independence is violated. Consider again the feature model F in Figure 4a which specifies that obesity causes diabetes with probability 0.15. Recall our earlier example where Tim, Tom and Zach are obese, where Tim and Tom are identical twins, and where Tim has diabetes. The assumption

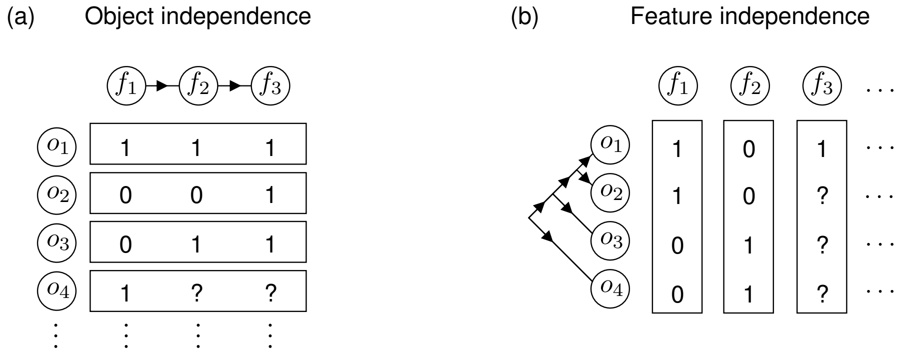


Figure 5: Independence assumptions made by models of generalization. (a) Models of causal reasoning generally assume that the rows of an object-feature matrix are conditionally independent given a structure over the features. These models are often used to account for across-feature generalization. (b) Models of similarity-based reasoning generally assume that the columns of the matrix are conditionally independent given a structure over the objects. These models are often used to account for across-object generalization.

of object independence implies that Tom and Zach are equally likely to suffer from diabetes, a conclusion that seems unsatisfactory. The assumption is false because of variables that are unknown but causally relevant—variables capturing unknown biological and environmental factors that mediate the relationship between obesity and diabetes.

One possible response to this problem is to work with a functional model that captures one or more unobserved variables. Consider, for example, the functional model in Figure 4b where transmission variable t indicates the presence of a certain gene that determines whether obesity causes diabetes. If we were confident that Tom carried the gene but were uncertain about whether Zach was a carrier, we might predict that Tom should be more likely than Zach to have diabetes. Note, however, that the gene variable is unobserved. In order to conclude that Tom carries the gene, we need to use the observation that Tim has diabetes, which suggests that Tim carries the gene, which in turn suggests that Tom also carries the gene. The final step depends critically on the knowledge that Tim and Tom are similar—knowledge that is violated by the assumption of object independence. In other words, even if we use a functional feature model F, we need to find some way to take relationships between objects into account.

We will relax the assumption of object independence by defining a prior distribution $P(M)$ that combines the feature structure F with a structure O that captures relationships between objects. Object structure O, for example, can capture the fact that Tim and Tom are identical twins, and are similar in many respects. First, however, we describe a prior distribution $P(M)$ that depends only on relationships between objects.

3.2. Object structures

We noted earlier that inferences may draw on different kinds of relationships between features, and knowledge about relationships between objects can be just as rich. The case of the three obese men shows that genetic relationships can matter, but many other relationships can guide inferences about unobserved features. Social relationships are relevant: for example, John is more likely to be obese if many of his friends are obese (Christakis & Fowler, 2007). Joint category membership may be relevant—if Rex and Spot are both dogs, then they are likely to have certain features in common, and if Rex and Rover are both Labradors, then even more inferences are licensed. Taxonomic relationships are often important when reasoning about animals, but other kinds of relationships including ecological relationships and predator-prey relationships may also play a role (Shafto & Coley, 2003). Finally, similarity can be treated as a relationship between objects, and there may be many kinds of similarity that guide inductive inferences (Medin et al., 1993).

Here we focus on a setting where the relationships of interest are captured by a single taxonomic tree. We previously described how Bayesian networks can capture relationships between features, and the same approach can capture relationships between objects.[1] Suppose that we are given a binary tree that captures taxonomic relationships among a set of m objects. The objects o_1 through o_m lie at the leaves of the tree, and we will use labels o_{m+1} through o_{2m-1} for the internal nodes of the tree. Figure 6a shows a simple case where the leaves of the tree represent four animals: a mouse, a rat, a squirrel and a sheep. The tree captures taxonomic similarity in the

[1] Although we focus in this paper on taxonomic relationships between *objects*, a taxonomic tree can also be used to capture inferences that rely on taxonomic relationships between *categories* (Tenenbaum et al., 2007). One possible approach is to use a tree where the leaves represent categories. The more general approach is to use a tree where the leaves represent objects and the internal nodes represent categories—the resulting representation can be used to make inferences about both objects and categories, and may be useful for modeling tasks like those described by Murphy & Ross (2010).

sense that objects nearby in the tree (i.e. nearby leaf nodes) are expected to have similar feature values. Let each feature be a vector $(o_1, \ldots, o_{2m+1})$ that assigns a value to each node in the tree, including the internal nodes. We will define a probability distribution over feature vectors which captures the idea that adjacent nodes tend to have the same feature value. Figure 6b shows two possible feature vectors. The first includes only one case where adjacent nodes have different values—object o_4 takes value 1 but the root node takes value 0. The second feature vector includes two cases where adjacent nodes take different feature values, and will therefore be assigned lower probability than the first feature vector.

We formalize these intuitions by turning the tree structure into a Bayesian network O. Suppose that λ_j is the base rate of feature f_j: in other words, the expected proportion of objects that have feature f_j. The Bayesian network O takes the base rate λ_j as a parameter, and specifies a distribution $P(f_j|O, \lambda_j)$ over possible extensions of feature f_j. Like all Bayesian networks, O includes a set of conditional probability distributions that specify how the value at each node depends on the values of its parents. The conditional probability distributions for O capture two basic intuitions: nodes tend to inherit the same values as their parents, but exceptions are possible, and become more likely when a child node is separated by a long branch from its parent. The following conditional distribution satisfies all of these requirements:

$$P(o_i = 1|\pi(o_i)) = \begin{cases} \lambda + (1-\lambda)e^{-l}, & \text{if } \pi(o_i) \text{ has value 1} \\ \lambda - \lambda e^{-l}, & \text{if } \pi(o_i) \text{ has value 0} \\ \lambda, & \text{if } o_i \text{ is the root node} \end{cases} \tag{8}$$

where l is the length of the branch joining object o_i to its parent. The last case in Equation 8 specifies that the probability distribution at the root node (o_{2m+1}) is determined directly by the base rate λ.

The conditional probability distributions in Equation 8 emerge from some simple assumptions about how features are generated. Suppose that feature f_j takes a value at every point along every branch in the tree, not just at the nodes. Imagine feature f_j spreading over the tree from root to leaves: the feature starts out at the root node with some value, and may switch its value (or mutate) at any point along any branch. Whenever a branch splits, both lower branches inherit the value of the feature at the point immediately before the split, and the feature now spreads independently along the two lower branches. Equation 8 follows from the assumption that

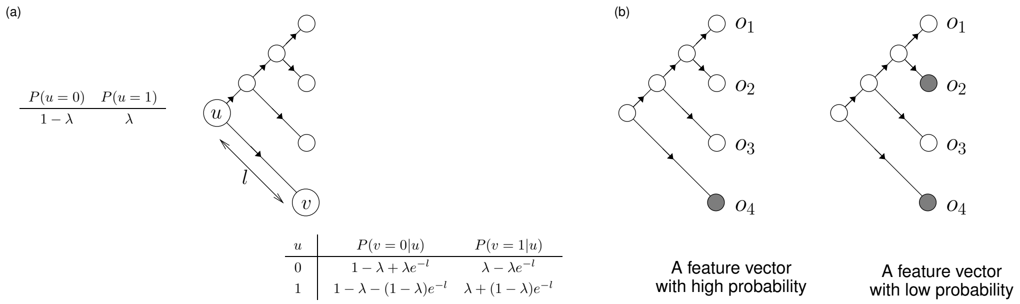

(a)
$P(u=0)$ $P(u=1)$
$1-\lambda$ λ
u
l
v
u $P(v=0|u)$ $P(v=1|u)$
0 $1-\lambda+\lambda e^{-l}$ $\lambda-\lambda e^{-l}$
1 $1-\lambda-(1-\lambda)e^{-l}$ $\lambda+(1-\lambda)e^{-l}$
(b)
o_1
o_2
o_3
o_4
A feature vector with high probability
A feature vector with low probability

Figure 6: Capturing taxonomic relationships between objects. (a) A tree structured graphical model. The conditional probability distribution for node v is shown, and all other conditional probability distributions are defined in the same way. (b) Prior probabilities assigned by the model in (a). Black nodes take value 1, and white nodes take value 0.

the feature value at any point in the tree depends only on the base rate λ and the feature value at the immediately preceding point.[2] Equation 8 has been previously used by computational biologists to study the evolution of genetic features (Huelsenbeck & Ronquist, 2001), and has also been proposed as a psychological model of property induction (Tenenbaum et al., 2007). Other methods for defining probability distributions over trees are possible (Kemp & Tenenbaum, 2009), and any model which captures the idea that nearby objects in the tree tend to have similar features is likely to work for our purposes.

The branch lengths in the tree help to capture the taxonomic relationships between objects. In Figure 6a, for example, the distance between o_1 and o_3 in the tree is twice the distance between o_1 and o_2, indicating that o_1 is more similar to o_2 than o_3. For all applications we assume that the topology of the tree and the relative magnitudes of the branch lengths are fixed, but that there is a single free parameter which corresponds to the total path length of the tree. If the total path length is very small then all of the objects are effectively very close to each other, and the prior distribution captured by network O will assign a prior probability of $1-\lambda$ to the

[2] Technically speaking, transitions between feature values are modeled using a continuous-time Markov chain with infinitesimal matrix:

$$Q = \left[\begin{array}{cc} -\lambda & \lambda \\ 1-\lambda & -(1-\lambda) \end{array} \right]$$

feature where all objects take value 0 and a prior of λ to the feature where all objects take value 1. If the total path length is very large then all of the objects are effectively very distant from each other, and the prior distribution captured by O will be similar to a distribution where the feature values for each object are independently generated by tossing a weighted coin with bias λ.

We have now defined a Bayes net O that specifies a distribution $P(f_j|O, \lambda_j)$ over single features. As before, this distribution can be used to define a prior distribution on object-feature matrices:

$$P(M|O, \boldsymbol{\lambda}) = \prod_j P(\boldsymbol{f_j}|O, \lambda_j) \tag{9}$$

where $\boldsymbol{\lambda}$ is a vector that specifies base rates for all features in the matrix. Equation 9 follows from the assumption of *feature independence*: the assumption that the features (ie. the columns of the matrix) are conditionally independent given the object structure O (Figure 5b).

The assumption of feature independence is relatively common in the psychological literature. Kemp & Tenenbaum (2009) describe four models of property induction that rely on this assumption, and Anderson's rational model of categorization is based on a very similar assumption. There are some cases of interest where this simplifying assumption appears to be justified. Consider, for example, inferences about blank features: given that whales have feature F, which other animals are likely to share this feature (Rips, 1975; Osherson et al., 1990)? Since little is known about feature F, it cannot be directly linked with any single known feature, and inferences about tasks of this kind tend to conform to taxonomic similarity. Participants might conclude, for example, that dolphins are more likely to have feature F than mice, since whales are more similar to dolphins than mice. A model that uses Equation 9 as its prior will account for results of this kind if the structure O captures taxonomic relationships between animals.

Although the assumption of feature independence is occasionally appropriate, models that make this assumption are limited in a fundamental way. A core finding from empirical work on property induction is that different features lead to different patterns of inductive inference (Gelman & Markman, 1986; Macario, 1991; Heit & Rubinstein, 1994; Shafto & Coley, 2003). Suppose, for example, that whales have a given feature, and that you need to decide whether bears or tuna are more likely to share the feature (Heit & Rubinstein, 1994). If the feature is anatomical (e.g. "has a liver with two chambers"), the anatomical match (bear) seems like the better response, but behavioral features (e.g. "travels in a zig-zag trajectory") support the behavioral match (tuna) instead. The assumption of feature independence cannot account for this result. Any model that makes this assumption predicts that two novel features will be treated in exactly the same way, since both are conditionally independent of all other features given a representation (O) of the relationships between animals.

The assumption of feature independence is unwarranted in part because people know about causal relationships between features. People know, for example, that "traveling in a zig-zag trajectory" is likely to be related to other features (like swimming and living in the water) that are shared by tuna and whales but not by bears and whales. Cases of this kind can be handled by combining causal relationships between features with taxonomic relationships between objects, and the next section considers how this combination can be achieved.

4. Combining knowledge structures

We have now described two models that can be used for reasoning about partially observed object-feature matrices. The feature model relies on a graph structure that captures causal relationships between features, and the object model relies on a graph structure that captures taxonomic relationships between objects. This section describes three approaches that can be used to combine these models. The approaches are summarized in Figure 7, and the critical difference between the three is the level at which the object and feature models are combined. The *output combination* approach combines the outputs generated by the two models, the *distribution combination* approach combines the probability distributions captured by the two models, and the *structure combination* model combines the graph structures over which the two models are defined. All three approaches seem plausible, but we will end up concluding that the structure combination approach provides the best account of our data.

4.1. The output combination model

Suppose first that the feature structure F and the object structure O are stored and used by two distinct reasoning modules. If these modules are informationally encapsulated (Fodor, 1983) there can be no direct interactions between these two structures. The predictions consistent with each structure, however, may be combined by some system that receives

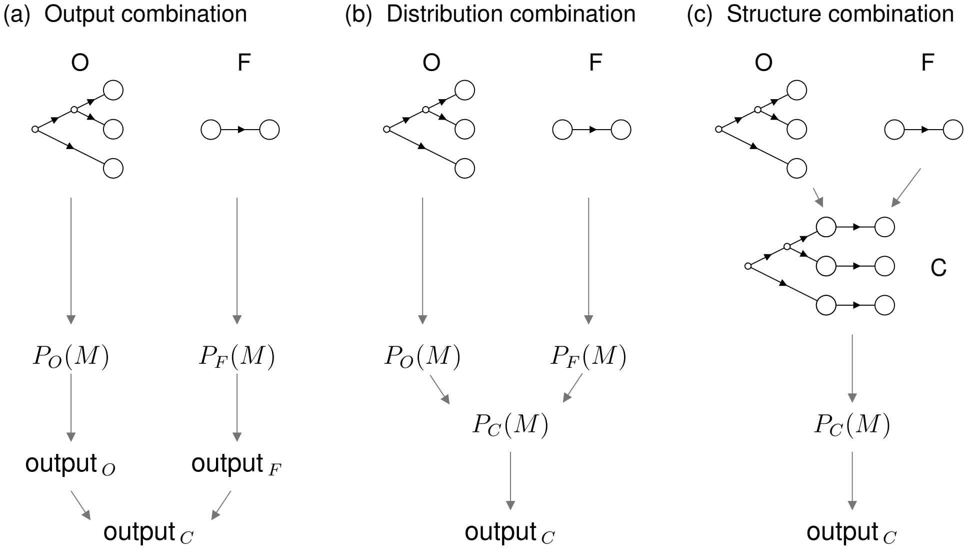


Figure 7: Three methods for combining an object structure O and a feature structure F. (a) The object and feature structures both induce prior distributions over object-feature matrices M. The output combination approach uses priors $P_O(M)$ and $P_F(M)$ to generate outputs consistent with each structure in isolation, then combines these outputs to generate output$_C$, the overall combined output. (b) The distribution combination approach combines the priors $P_O(M)$ and $P_F(M)$ to generate a combined prior distribution $P_C(M)$ over object-feature matrices. This prior is used directly to generate the overall output. (c) The structure combination model combines the object and feature structures to create a combined structure C over which a prior distribution $P_C(M)$ is defined.

input from both modules. This approach to knowledge integration is shown schematically in Figure 2b, and we refer to it as the output combination model (or OC model for short). Figure 7a shows how the OC approach can be applied given probabilistic models defined over structures F and O. The two models induce priors $P_O(M)$ (Equation 9) and $P_F(M)$ (Equation 7) over object-feature matrices M, and these two priors can be used to generate outputs in response to any given query. The overall or combined output is generated by combining these two outputs.

The OC model has been previously discussed in the literature on information integration (Anderson, 1981), which explores how multiple sources of information can be combined. The most typical approach in this literature is to combine multiple predictors using a simple mathematical function such as a sum (Lombardi & Sartori, 2007), a product (Medin & Schaffer, 1978; Massaro & Friedman, 1990; Ernst & Banks, 2002), or a weighted average (Anderson, 1981). We implemented all of these possibilities and the OC model evaluated in this paper uses a weighted average, which turned out to be the best performing combination function.

The OC approach is very general and can be used to combine the predictions of any set of models, including some that are probabilistic and some that are not. In addition to the psychological literature on information integration, the approach has also been explored in several other fields, including statistics, machine learning, and artificial intelligence. For example, the weighted average model evaluated in this paper will be familiar to machine learning researchers as a "mixture of experts" approach (Jacobs et al., 1991).

From a normative perspective, one limitation of the OC model is that it sometimes fails to draw out the full implications of the available information. For example, suppose that you know that obesity causes diabetes, that Tim and Tom are identical twins, and that Tim is obese. The OC model can infer that Tim is likely to have diabetes and that Tom is likely to be obese, since these conclusions follow from the feature and object models respectively. The OC model, however, cannot infer that Tom is likely to have diabetes, since this conclusion follows from neither component model in isolation. More generally, the OC model cannot make informed inferences about arguments where the objects and features mentioned in the conclusion do not appear among the premises. The argument in Figure 1c is one example: the OC model can infer that the rat is likely to have gene X, and that the mouse is likely to have enzyme Y, but cannot infer that the rat is likely to have enzyme Y. This potential weakness of the OC model can be addressed by combining the feature and object models directly instead of combining their outputs. This approach is shown schematically in Figure 2c, and the following sections describe two instances of this approach.

4.2. The distribution combination model

If the two component models are both probabilistic models, the two can be combined by combining the prior distributions that they capture. Figure 7b summarizes this approach and shows that the priors $P_O(M)$ and $P_F(M)$ are combined to create a prior $P_C(M)$ that is then used to compute the final output. We refer to this approach as the distribution combination model, or the DC model for short.

Just as there are several ways to combine the outputs of two modules, the prior distributions induced by two models could be combined using a weighted average or a product. The DC model evaluated in this paper uses a prior that is the product of the priors for the two component models:

$$P(M|F, O) \propto \prod_i P(\boldsymbol{o_i}|F) \prod_j P(\boldsymbol{f_j}|O, \lambda_j), \tag{10}$$

where the base rate λ_j for feature f_j is set by the marginal probability of f_j according to F. This model will be familiar to machine learning researchers as a "product of experts" model (Hinton, 2000). A DC model that relies on a weighted average of priors is also worth considering, but this model turns out to be equivalent to an OC model that relies on a weighted average. In general, however, a DC model will not be equivalent to an OC model—for example, a DC model that relies on a product of priors does not generate the same predictions as an OC model that relies on a product. As a result, the DC approach should be distinguished from the OC approach.

The DC prior in Equation 10 assigns high probability only to matrices that receive relatively high prior probability according to both the feature model (Equation 7) and the object model (Equation 9). For example, in the obesity and diabetes example, the matrices that receive high prior probability are those where Tim and Tom have the same feature values, and where individuals with diabetes are obese. This prior can then be used to generate predictions about unobserved entries in an object-feature matrix, as summarized by Figure 3. For example, the model can handle the argument in Figure 1c that proved challenging for the OC model. Given that the mouse has gene X, the DC model can predict that the rat is likely to have enzyme Y, an inference that relies on both the similarity between the mouse and the rat and the causal relationship between gene X and enzyme Y.

Although the DC model is capable of making sensible qualitative predictions about all of the generalization problems described so far, it may struggle to make accurate quantitative predictions. From a normative perspective, one limitation of the approach is that it assigns too much weight to feature base rates. Consider a very simple setting where there is one object and one feature with a base rate of 0.9. Matrix M is either 1 or 0, and the probability that $M = 0$ should intuitively be 0.1. The DC model, however, generates a prior distribution where $P(M = 0) \approx .01$. Since the prior is generated by multiplying a distribution over rows with a distribution over columns, the model effectively uses the base rate twice, which means that probabilities of rare events (e.g. $M = 0$) end up much smaller than they should. Our experiments do not address this limitation directly, but it may

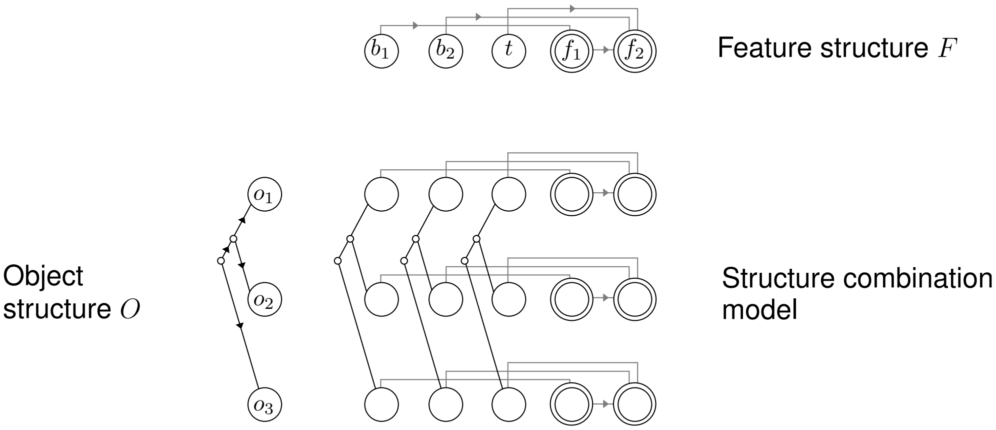


Figure 8: The structure combination model is created by combining an object structure O with a feature structure F. The feature structure F shown here is the same as the structure in Figure 4b, although the spatial layout of the variables has been altered. The SC model assumes that the root variables in F (here b_1, b_2 and t) are independently generated over O, and can be represented as a causal graphical model. The arrows on the edges of the graphical model are inherited from the component structures, but all except three have been suppressed for visual clarity. Note, for example, that all edges inherited from the feature structure F are oriented from left to right.

help to explain why the DC model achieves relatively low quantitative fits in some cases.

Moving from simple generalization problems to problems involving causal inferences about interventions and counterfactuals may raise some more fundamental challenges for the DC model. One common approach to causal reasoning makes use of a directed graph that captures causal relationships between variables, and manipulates this graph in order to reason about interventions and counterfactuals. The two components of the DC model are defined over directed graphs, but there is no overall graph structure that captures the way in which these two component graphs combine. It may turn out that manipulating each component graph separately then combining the two according to the DC approach is enough to account for human inferences about interventions and counterfactuals. Our third experiment explores this possibility, and to preview the results we find a qualitative mismatch between human inferences and the predictions of the DC approach.

4.3. The structure combination model

The DC model combines an object and a feature structure by combining the distributions induced by these structures, but the structure combination approach (SC approach for short) combines these structures directly. Figure 2 shows that the combined structure C induces a prior distribution $P_C(M)$, which can then be used to make inferences about generalization problems.

To explain how the two structures are combined, we use an example where feature structure F is the functional model in Figure 4b and O is a tree defined over three objects (Figure 8a). Recall that structure F indicates that obesity (f_1) causes diabetes (f_2), and suppose that structure O indicates that Tim and Tom (o_1 and o_2) are more similar to each other than either one is to Zach (o_3). Note that the relationship between obesity (f_1) and diabetes (f_2) in Figure 4b is mediated by a transmission variable t, which summarizes the influence of genetic factors that are unknown but relevant.

Even though variable t may capture one or more unknown factors, we do know something about this variable—we expect that the values it takes across the three objects will tend to respect the similarity relationships captured by O. For example, if the transmission variable t takes value 1 for just two of the three individuals, we might expect that these two individuals are more likely to be Tim and Tom (a similar pair) than Tim and Zach (a dissimilar pair). The other root variables should likewise respect the similarity relationships captured by O, and we therefore assume that all root variables in F are generated independently over O:

$$P(M|F,O) = P(\boldsymbol{b_1}|O,\lambda_{b_1})P(\boldsymbol{b_2}|O,\lambda_{b_2})P(\boldsymbol{t}|O,\lambda_t)P(\boldsymbol{f_1}|\boldsymbol{b_1})P(\boldsymbol{f_2}|\boldsymbol{b_2},\boldsymbol{f_1},\boldsymbol{t}) \tag{11}$$

where each matrix M now includes five columns for variables b_1, b_2, t, f_1 and f_2, and the base rates λ_{b_1}, λ_{b_2}, and λ_t are specified by the feature structure F. The last two terms on the right-hand side of Equation 11 indicate that variables f_1 and f_2 depend on the root variables but not the object structure O. There is no need to generate f_1 and f_2 over the object structure, since these variables are deterministically specified once the root variables have been fixed.

The prior distribution in Equation 11 can be represented as a causal Bayesian network defined over a graph product of feature structure F and object structure O. Figure 8 shows the graph for this network. Note that we have introduced a copy of O for each root variable in F, and that these root variables are connected to the deterministic variables f_1 and f_2 as specified by F. The graph product in Figure 8 inherits its conditional probability distributions from the feature structure F and the object structure O. As a result, the Bayesian network corresponds exactly to the prior distribution in Equation 11.

The SC model in Figure 8 is a network where the six variables of primary interest (i.e. the six variables representing values of f_1 and f_2 for the three objects) are deterministically specified given their parent variables. The model therefore qualifies as a functional causal model and offers all of the advantages of these models. For example, the SC model can be used in the standard way to reason about interventions and counterfactuals. Since the SC model can be represented as a Bayesian network, model predictions can be computed efficiently by standard algorithms for inference in Bayesian networks. All of the results in this paper were computed using the Bayes Net toolbox (Murphy, 2001). Figure 8 shows how our approach can be used to integrate one specific feature structure F and one specific object structure O, but the same approach can be used when F is any functional causal model and O is any tree structure. We will illustrate this flexibility by considering several different feature structures in our experiments.

Although the SC model is motivated by problems where object structures and feature structures should be combined, previous studies have documented cases where multiple structures are not combined. For example, Rehder (2006) describes some cases where causal relationships between features appear to dominate similarity relationships between objects, and the two are not combined. Any account of knowledge integration should therefore attempt to distinguish between cases where multiple structures are and are not combined. Since this paper focuses on taxonomic relationships between objects and causal relationships between features, we need some way to predict when taxonomic relationships should be taken into account. The SC model motivates the following *taxonomic influence* principle:

> Taxonomic relationships should be taken into account if and only if these relationships provide information about the distribution of variables that are unobserved but causally relevant to some feature of interest.

The taxonomic influence principle identifies two distinct cases where taxonomic relationships should play no role. The first case includes problems where taxonomic relationships are simply irrelevant to the features of interest. Suppose, for example, that an eccentric businessman is interested in buying animals with names that end in a consonant. In this case the assumptions of the SC model do not apply, since the causal variables that influence the businessman's decision do not respect the similarity relationships between animals captured by a taxonomic tree. The second case includes problems where taxonomic relationships *are* relevant to the features of interest, but where there are no unobserved root variables. In Figure 4b, for example, there are three root variables (t, b_1 and b_2), and the SC model in Figure 8 predicts that taxonomic relationships will not shape inferences once these three variables are observed for each object. Although this second case is a legitimate theoretical possibility, in real-world settings it is usually impossible to observe all of the causal root variables. We therefore expect that the first case will cover most of the real-world settings where taxonomic relationships are found to play no role in inductive reasoning.

4.4. Special cases of the structure combination model

Although this paper focuses on problems where object structures and feature structures must be combined, the combination models just described can also make inferences about cases where only one structure is relevant. We illustrate by explaining how our approach of choice—the SC model—subsumes previous probabilistic models that rely on either a feature structure or an object structure in isolation.

Many previous authors have used Bayes nets to account for inductive reasoning, and any model that corresponds to a Bayes net defined over features or a Bayes net defined over objects can be viewed as a special case of the SC model. Accounts of causal reasoning (Glymour, 2001; Rehder, 2003; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005) often rely on Bayes nets defined over features, and the SC model reduces to a Bayes net of this kind when the object structure O indicates that all objects are equally similar to each other. Suppose, for example, that object structure O is a tree where all objects are directly linked to the root node and lie at the same distance from this node. In this case the object structure plays no role and the SC model is identical to a model which assumes that objects correspond to independent samples from a Bayes net defined over features. Several previous accounts of across-object generalization rely on Bayes nets defined over objects (Shafto et al., 2008; Tenenbaum et al., 2007), and the SC model reduces to a Bayes net of this kind when the feature structure F indicates that all features are independent. In this case the feature structure plays no role, and the SC model is identical to a model which assumes that features correspond to independent samples from a Bayes net defined over objects.

Since the SC model subsumes most previous models that rely on Bayes nets, it accounts for the data that have been presented in support of these models in exactly the same way as the original models. The SC model should therefore be viewed not as a competitor to these previous models, but as an extension of these models. Previous models suggest that Bayes nets can be used to capture both relationships between features and relationships between objects, and the key contribution of the SC model is to demonstrate how these different kinds of knowledge can be combined.

5. Experiment 1: Generalization across objects and features

Our working hypothesis is that people find it natural to combine relationships between features and relationships between objects. Real-world examples like the case of the three obese men appear to support our hypothesis, and we designed three experiments to test this hypothesis under controlled laboratory conditions. All of our experiments used a set of four animals—a mouse, a rat, a squirrel and a sheep. These animals were chosen to include pairs that are similar (e.g. mouse and rat) and pairs that are not (mouse and sheep). A taxonomy that captures similarity relationships between these animals is shown in Figure 9a. A generalization task described in Appendix Appendix A confirmed that this tree matches human judgments about the relationships between these four animals. We explored several different feature structures and four examples are shown in Figure 9b.

Our first two experiments explore whether people make inferences that simultaneously draw on relationships between objects and relationships between features. The OC, DC and SC models all predict that object and feature structures are combined, and we compare these models with alternatives that rely on a feature structure alone or an object structure alone. Our third and final experiment focuses on counterfactual interventions, and we explore whether and how people combine object and feature structures in this setting.

Our first experiment considers a setting where participants are asked to make inferences about the missing entries in an object-feature matrix. The matrix is sparsely observed: for example, participants might be told only that the mouse has f_1, then asked to fill in the remaining entries. We were interested to see whether their responses would be guided by the causal relationships between the features, the taxonomic relationships among the four animals, or by both kinds of relationships.

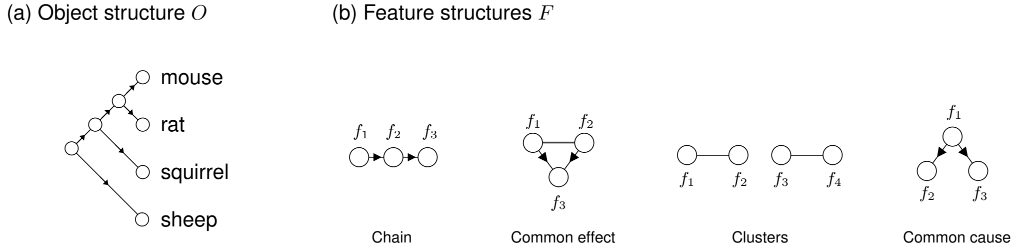


Figure 9: Object and feature structures used for experiments 1 and 2. (a) An object structure O that captures taxonomic relationships between the four animals in our experiments. (b) Feature structures F that summarize four kinds of relationships between the observed features in our experiments. The thick grey edge between features f_1 and f_2 in the common effect model indicates that these features are mutually exclusive. The undirected edge between f_1 and f_2 in the cluster structure indicates that these features are known to co-occur, but that neither feature causes the other. Functional causal models consistent with each structure are shown in Figure B.20 of the appendix.

5.1. Participants

16 MIT undergraduates participated in the experiment in exchange for pay. Participants were recruited through newsgroup and bulletin board postings and had no personal connection with the experimenters.

5.2. Materials

Participants read the following instructions:

> “You are a biochemist and you study enzyme production in mammals. Each mammal produces many enzymes, and different mammals can produce very different sets of enzymes. In your lab today you have a mouse, a rat, a sheep and a squirrel. You will be running tests on each animal to determine the presence of certain enzymes.”

The experiment had three within-participant conditions, each of which was associated with a different set of features. Each feature indicates the presence or absence of an enzyme. Pseudo-biological names like “dexotase” were used in the experiment, but here we use labels such as f_1 and f_2. The relationships between the observed features in each condition are summarized in Figure 9b. In the chain condition, participants were told that f_3 is known to be produced by several pathways, and that the most common pathway begins with f_1, which stimulates production of f_2, which in turn leads to the production of f_3. In the common-effect condition, participants were told that f_3 is known to be produced by several pathways. One of the most common pathways involves f_1 and the other involves f_2, although f_1 and f_2 are rarely found in the same animal. In the cluster condition, participants were told about four enzymes: f_1 and f_2 are complementary enzymes that work together in the same biochemical pathway, and f_3 and f_4 are complementary enzymes that work together in a different biochemical pathway.

To reinforce each causal structure, participants were shown 20 cards representing animals from twenty different mammal species (names of the species were not supplied). The card for each animal included a bar chart which showed whether or not that animal had tested positive for each enzyme in the current condition. The cards were chosen to be representative of the distribution induced by a functional model with known structure and known parameterization. The functional models and the cards used for each condition are described in Appendix Appendix A. Even though each condition is based on a functional causal model, note that the cards and all other experimental materials mention only some of the variables in these models, and all of the information participants received was consistent with the existence of probabilistic causal relationships between the observed variables. We chose not to train participants on functional models for two reasons. First, real-world causal problems often involve systems where many of the relevant variables are unknown, which means that causal relationships between observed variables typically appear to be probabilistic. Second, previous psychological studies typically focus on probabilistic relationships between observed variables, and we wanted to maintain continuity with the large body of existing work in this area.

5.3. Procedure

The experiment began with a preliminary taxonomic task that was designed to probe background knowledge about taxonomic relationships between the four animals in the study. Participants were told that scientists had recently identified four enzymes and were asked 12 questions of the following form:

> "You discover that the mouse produces enzyme Q84. How likely is it that the rat produces Q84?"

Responses were provided on a scale from 0 (very unlikely) to 100 (very likely).

Participants then moved on to the three within-participant conditions. In each condition, participants read a description of a given causal structure: chain, common-effect or clusters. As described above, participants were given a set of 20 cards showing samples consistent with the causal structure. After participants had studied the cards for as long as they liked, the cards were removed and participants responded to a preliminary causal task including questions about unidentified mammals. One group of questions asked about the base-rate of each feature:

> “You learn about a new mammal. How likely is it that the mammal produces enzyme f_1?

The remaining questions asked about relationships between features:

> “You learn that a mammal produces enzyme f_1. How likely is it that the same mammal also produces enzyme f_2?

The questions in this preliminary task were intended to encourage participants to reflect on the causal relationships between the enzymes.

In each condition, participants were told that they would be testing the four animals (mouse, rat, sheep and squirrel) for each enzyme of interest. In the chain and common-effect conditions there were 12 tests in total (three for each animal), and in the clusters condition there were 16 tests in total. Each condition included two tasks. In the chain condition, participants were told that the mouse had tested positive for f_1 and were asked to predict the outcome of the 11 remaining tests. Participants were then told in addition that the rat had tested negative for f_2, and asked to predict the outcome of the 10 remaining tests. Note that this second task requires participants to integrate causal reasoning with taxonomic reasoning: causal reasoning predicts that the mouse has f_2, and taxonomic reasoning predicts that it does not. In the common-effect condition, participants were told that the mouse had tested positive for f_3, then told in addition that the rat had tested negative for f_2. In the cluster condition, participants were told that the mouse had tested positive for f_1, then told in addition that the rat had tested negative for f_4.

Responses to all questions were provided on a scale from 0 (very likely to test negative) to 100 (very likely to test positive). Participants made their responses by filling in a matrix with a row for each object and a column for each feature. One or two entries in the matrix were already present: for example, if participants had been told that the mouse had tested positive for f_1, the corresponding entry in the matrix was set to 100.

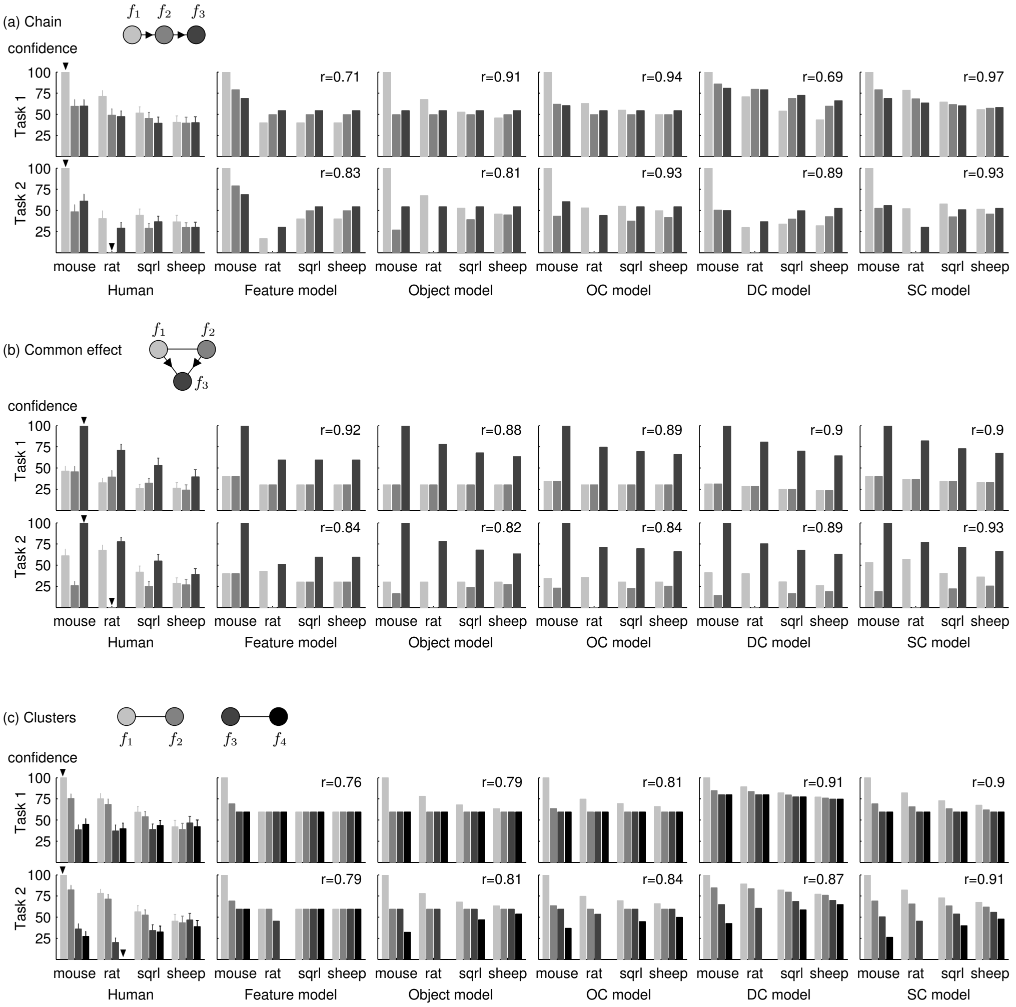


Figure 10: Experiment 1: Average human responses (column 1) and predictions for four models. All models generate probabilities as output, and these probabilities have been multiplied by 100 for comparison with the human data. (a) Results for the chain condition. In this figure and subsequent figures, known test results are marked with wedges. In task 1, participants were told only that the mouse had tested positive for f_1, and in task 2 they were told in addition that the rat had tested negative for f_2. Error bars represent the standard error of the mean. (b) Results for the common-effect condition. (c) Results for the clusters condition.

5.4. Model predictions

We will evaluate all models by considering both the qualitative effects that they predict and their quantitative correspondence with the human data. All models except the feature model rely on a tree that captures taxonomic relationships between the animals. We used the tree shown in Figure 9 which captures the idea that the mouse and the rat are very similar, that these two animals are somewhat similar to the squirrel, and that none of these three animals is very similar to the sheep. The branch terminating at the sheep is 3 units long, the branch terminating at the squirrel is 2 units, and all remaining branches are of length 1. Note that the four animals all lie at the same distance from the root. The tree component of each model has one free-parameter—the total path length of the tree. The smaller the path length, the more likely that all four animals have the same feature values, and the greater the path length, the more likely that distant animals in the tree (e.g. the mouse and the sheep) will have different feature values. For each model, the path length is set to the value that maximizes the average correlation with human data across Experiments 1 and 2. The values for the object model, OC model, DC model and SC model were 3.5, 1.9, 2.9 and 2.0 respectively.

All models except the object model rely on a functional causal model that captures relationships between the features. The functional models F used for each condition are shown in Figure B.20. Note that the functional models include no free numerical parameters, since the base rates for the root variables are fixed by the parameters of the network that generated the cards shown to participants during the training phase. The OC model has one additional parameter that specifies the weights assigned to the two component models. All results reported here use a weight of 0.42 for the feature model and a weight of 0.58 for the object model, and these values maximize the average correlation with human data across Experiments 1 and 2. The correlations achieved by the model are only marginally lower if the component models are weighted equally.

The second and third columns of Figure 10 show predictions for a *feature model* that uses the feature structure alone, and an *object model* that uses the object structure alone. In task 1 of the chain and common-effect conditions, neither approach predicts that inferences about all three features will decay smoothly over the tree. The feature model does not incorporate taxonomic relationships between the objects and makes identical predictions about the rat, the squirrel and the sheep. The object model does not incorporate causal relationships between features and therefore has no basis for making predictions about f_2 and f_3 given information about f_1.

The final three columns in Figure 10 show predictions for the three combination models. The predictions of all models are probabilities, and these probabilities have been multiplied by 100 for comparison with the human data reported in the next section. All combination models predict that responses will be guided by both the feature structure F and the object structure O. The predictions for task 1, however, reveal an important qualitative difference between the OC model and the DC and SC models. Given a single observed entry in an object-feature matrix, all of the combination models predict that humans will make informed inferences about the row and the column that contain the observation. For example, given that the mouse has f_1 and that f_1 causes f_2 (chain condition), all three models predict that participants will infer that the rat is relatively likely to have f_1 and that the mouse is relatively likely to have f_2. The DC and SC models predict that participants will use the single observation provided to make inferences about the rest of the matrix—for example, in the chain condition both predict that participants will infer that the rat is relatively likely to have f_2. In contrast, the OC model cannot make informed inferences about entries in the matrix that do not belong to the same row or column as the single observation, and predicts that participants will fall back on base rates when reasoning about these entries. Our model evaluation will focus on one important consequence of this qualitative difference between the models. In task 1 of each condition, the OC model predicts that inferences about the observed feature will decay over the tree, but that inferences about the remaining features will be identical for the rat, squirrel and sheep. The DC and SC models predict that inferences about all features (chain and common-effect) or about the first two features (clusters) will decay smoothly over the tree. For example, in task 1 of the chain condition the OC model predicts that inferences about f_2 and f_3 will be identical for the rat, squirrel, and sheep, but the DC and SC models both predict that the rat is more likely to have f_2 and f_3 than the sheep.

In task 2 of each condition, all combination models use the causal structure F and the object structure O to reconcile the two observations provided. In the chain condition, the second observation is unexpected: given that the mouse has f_1, the rat is similar to the mouse, and that f_1 causes f_2, it is surprising that the rat does not have f_2. All combination models infer that the second observation makes it less likely that the mouse has

f_2 and that the rat has f_1. In the common effect condition the second observation is less surprising: given that the mouse has f_3, this feature was probably caused by f_1 or f_2, and given that the rat does not have f_2, f_1 is the more likely of the two causes. All combination models therefore infer that all of the animals are more likely to have f_1 than f_2. Task 2, however, does produce some subtle qualitative differences between the models. As for task 1, we focus here on predictions about entries in the matrix that do not belong to the same row or the same column as an observed entry. In the chain condition, the OC model predicts that the squirrel and sheep are both equally likely to have f_3, but the DC and SC model predict that the sheep is marginally more likely than the squirrel to have this feature. In the common effect condition, the OC model predicts that the squirrel and sheep are both equally likely to have f_1, but the DC and SC model predict that the squirrel is more likely than the sheep to have this feature. In the cluster condition, the OC model predicts that the squirrel and sheep are equally likely to have f_2 and f_3. The DC and SC models, however, predict that all of the animals are more likely to have the features in the first cluster (f_1 and f_2) than the features in the second cluster (f_3 and f_4). All of these qualitative differences are consequences of the fact that the OC model can make informed entries only about matrix entries that belong to either the same row or the same column as an observed entry.

5.5. Results

Responses to the preliminary taxonomic task suggested that the tree in Figure 9 accurately captures background knowledge about taxonomic relationships between the four animals in our experiment. Responses to the preliminary causal tasks suggested that participants understood the causal structures in Figure 9. These results support the idea that the structures in Figure 9 are appropriate for modeling the data collected in the rest of the experiment. More details about the results of the preliminary tasks are provided in Appendix Appendix A.

Mean responses for the three conditions are shown in the first column of Figure 10. Before considering the quantitative fit of each model, we first assess the qualitative predictions identified in the previous section. In all three conditions, human inferences appear to be guided by both the causal relationships between features and the taxonomic relationships between objects. In task 1 of each condition, predictions about all features (chain and common-effect conditions) or about the first two features (clusters condition) decay smoothly over the tree. As predicted by the DC and SC models but not the OC model, participants are able to use a single observation to make informed predictions about the entire object-feature matrix. Comparisons between predictions for the rat and the sheep are especially revealing. In the chain condition, sign tests indicate that the rat is judged more likely than the sheep to have f_2 and f_3 ($p < 0.05$ in both cases). In the common-effect condition, the rat is judged more likely than the sheep to have f_1 and f_2 (sign tests yield $p < 0.05$ in both cases). In the clusters condition, the rat is judged more likely than the squirrel and sheep to have f_2 (sign tests yield $p < 0.05$ in both cases). All of these results are inconsistent with the feature, object, and OC models, but are captured by the DC and SC models.

Responses for the second task in each condition suggest that participants reconcile multiple observations as predicted by the combination models. After receiving a second observation in the chain condition, participants consider it less likely that the mouse has f_2 and that the rat has f_1. In task 2 of the common effect condition, participants infer that the mouse, the rat and the squirrel are all more likely to have f_1 than f_2. In task 2 of the clusters condition, participants infer that the mouse, the rat and the squirrel are all more likely to have f_1 and f_2 than f_3 and f_4. Unlike the results for task 1, the data for task 2 provide only partial support for the prediction that participants make informed inferences about entries in the object-feature matrix that do not belong to the same row or column as an observed entry. In the chain condition, the squirrel is judged more likely than the sheep to have f_3, but the DC and SC models generate a small difference in the opposite direction. In the common effect condition, the squirrel is judged more likely than the sheep to have f_1 but a sign test indicates that this difference is only marginally significant ($p < 0.1$). In the clusters condition, the squirrel is judged more likely than the sheep to have f_3 (again $p < 0.1$), but the sheep is judged equally likely to have f_2 and f_3. Overall, the qualitative effects identified for task 1 provide strong support for the DC and SC models ahead of the OC model, but the qualitative effects for task 2 do not distinguish as clearly between the combination models. Note, however, that the qualitative effects for task 2 all correspond to relatively small quantitative differences according to the predictions of the SC model.

To further assess the performance of the models we computed correlation coefficients between the human data and the predictions of each model. A correlation coefficient is a relatively crude measure of performance in this setting, but Figure 10 shows that the SC model achieves the highest correlations overall. The most important differences between the SC model and the OC model emerge in Task 1 of each condition. Although the correlations achieved by these models are similar, the OC model performs slightly worse than the SC model because it does not make informed inferences about rows and columns that do not include the single observation provided. In task 1 of the chain condition, for example, the model observes that the mouse has f_1 but does not infer that the probability that the rat has f_2 is now above baseline. Instead, the model predicts that the rat, squirrel and sheep are equally likely to have f_2. In contrast, the SC model successfully predicts that inferences about all features (including f_2 and f_3) will decay over the tree.

The SC and DC models perform similarly in the common-effect and clusters conditions, but the SC model provides the better account of the chain condition. In task 1, the DC model predicts that the rat, the squirrel and the sheep are all less likely to have f_1 than f_2 and f_3. This result is driven by the base-rate of f_1 specified by the causal model—note that the feature model also makes the same prediction. The poor performance of the DC model is therefore consistent with our earlier suggestion that this model tends to overweight base rate information. The correlation for task 2 of the chain condition is better, but note that the DC model still makes inaccurate predictions: unlike the SC model, the DC model predicts that the rat and the squirrel are more likely to have f_3 than f_1. Both tasks in the chain condition therefore suggest that the DC model provides an imperfect account of how humans combine causal relationships with taxonomic relationships.

The inferences made by the DC model depend on the free-parameter mentioned previously: the total path length of the tree O. When this parameter is very small, predictions about all four animals are very similar to predictions about the mouse, and when the parameter is very large, the DC model makes predictions very similar to the feature model. Adjusting the parameter allows the model to interpolate between these two extremes, but no setting of the parameter allows the model to strike the right balance between the causal relationships and the taxonomic relationships. The predictions in Figure 10 are for the parameter setting that maximizes model performance across all of the tasks, but if the parameter is fitted specifically for task 1 of the chain condition, the correlation achieved is still only 0.70.

Although the SC model provides a good account of the average response to each task, it may seem possible that the success of this model depends on averaging the predictions of participants with very different strategies. If some participants matched the feature model and others matched the object model, then the average response would be similar to the predictions of the OC model, which computes a weighted average of the predictions of the component models. The qualitative differences between the OC predictions and the human data provide some initial evidence that some participants are combining feature and object structures. More direct evidence is provided by partitioning participants into three groups depending on whether their responses relied on the feature relationships alone, the object relationships alone, or on a combination of these relationships. To create these groups we computed whether the responses of each participant correlated best with the feature model, the object model, or the SC model. The pathlength parameters used by the object and SC models were not fit to each individual participant but fixed throughout to the values that generated the predictions in Figure 10. Four participants matched the feature model, two participants matched the feature model, and 10 out of 16 participants matched the SC model. We can therefore conclude that the average responses in Figure 10 are representative of the responses of many participants, and that the majority of participants combined feature relationships with object relationships.

Overall, the results of Experiment 1 suggest that the combination models are superior to the models that rely on a feature structure alone or an object structure alone, and that the SC model is the best of the three combination models. These results suggest that participants can combine relationships between features and relationships between objects when making inductive inferences, and that the structure combination model helps to explain how these different kinds of information are combined.

6. Experiment 2: Generalization across objects and features

Experiment 1 provides strong evidence that humans can make inferences that draw on both causal relationships between features and taxonomic relationships between objects. This result may seem incompatible at first with the work of Rehder (2006), who found no evidence that people could combine causal and similarity-based reasoning. So far, however, our data are consistent with the hypothesis that causal relationships between features are primary and that taxonomic relationships are used only when no observations at all are available for some objects.[3] For example, taxonomic relationships may have played a role in the chain condition of Experiment 1 only because participants were never given any information about whether the squirrel and sheep had features f_1, f_2 or f_3. As a result, taxonomic relationships provided the only relevant information that participants could use to make inferences about these animals.

Our second experiment explores whether taxonomic relationships continue to play a role when observations for all four animals are available. Since these observations support causal inferences about each animal, it is possible that participants will now ignore taxonomic relationships and focus exclusively on causal relationships between features. We predict, however, that participants will continue to rely on taxonomic relationships in this situation. Our prediction is a consequence of the taxonomic influence principle introduced previously. Even if observations are provided for all animals in the experiment, taxonomic relationships should continue to play a role as long as there are variables that are unobserved but causally relevant to the features of interest.

6.1. Participants

18 MIT undergraduates participated in this experiment. The responses of one participant were removed because he left some pages in the experimental packet blank.

6.2. Procedure

Experiment 2 included two conditions: a chain condition and a common-cause condition. Each condition included two tasks. In the first task, participants were told only that the mouse had tested positive for f_1. In the second task, participants were told in addition that the rat, the squirrel and the sheep had tested positive for f_1, and that the mouse had tested negative for f_2. Note that the second task is a case where values for feature f_1 were provided for all animals. Apart from this task, the procedure for Experiment 2 was identical to that of Experiment 1.

6.3. Model predictions

Predictions for four models are shown in Figure 11. Predictions for the second task in each condition are most critical. Even though feature f_1 is observed for all four animals, the SC and DC models still predict a taxonomic

[3] We thank Bob Rehder for suggesting this hypothesis.

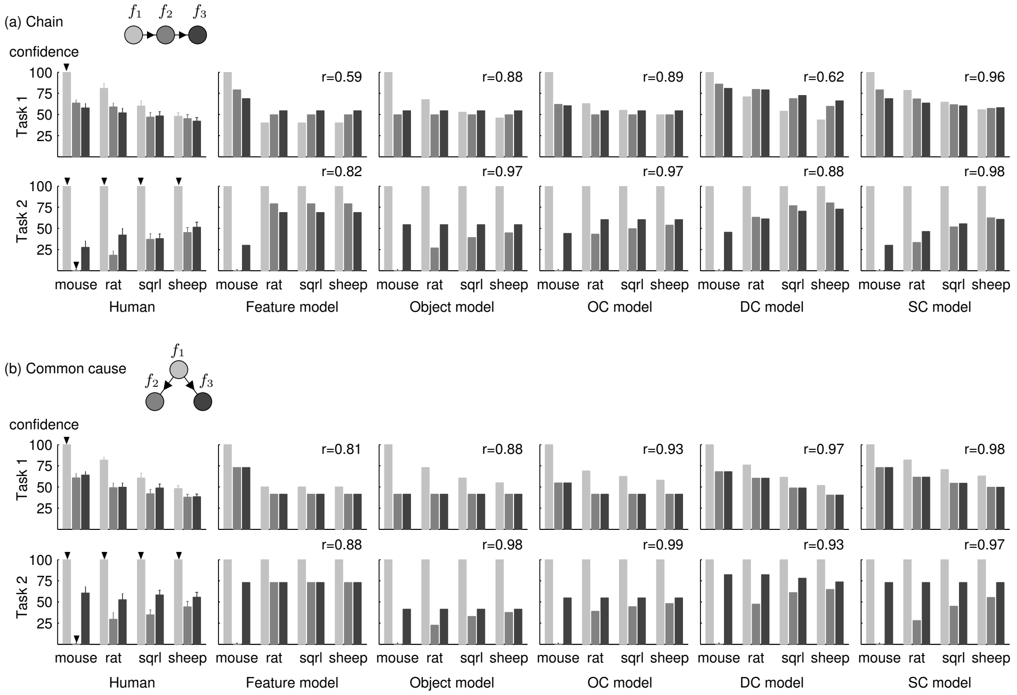


Figure 11: Experiment 2: Behavioral data and predictions of five models. In task 2 of each condition, feature f_1 is observed for all four animals.

effect: as taxonomic distance from the mouse increases, animals become more likely to have f_2. The feature model makes a different prediction—note that this model does not take taxonomic relationships into account, and therefore makes identical predictions about the rat, the squirrel and the sheep.

6.4. Results

Figure 11 shows mean responses for the participants and the four models described previously. The judgments for the first task in each condition replicate the finding from Experiment 1 that participants combine feature relationships and object relationships when just one of the 12 animal-feature pairs is observed. The results for the second task suggest that taxonomic information continues to be used even when observations for all four animals are provided. In both conditions, for example, participants infer that the rat is less likely than the sheep to have f_2 (sign tests yield $p < 0.05$ in both cases).

As for Experiment 1 we explored individual differences by dividing participants into groups depending on whether their responses correlated best with the feature model, the object model, or the SC model. Two participants matched the feature model, six participants matched the object model, and 9 out of 17 participants matched the SC model on the basis of their complete set of responses. Since the second task in each condition is critical for distinguishing between the feature model and the SC model, we ran a second analysis using data from the second task only and computing whether each participant better matched the feature model or the SC model. 14 out of 17 participants matched the SC model better than the feature model, and a sign test suggests that this result is statistically significant ($p < 0.05$). We can therefore conclude that the majority of participants relied on taxonomic information even in task 2.

Taken together, Experiments 1 and 2 provide strong evidence that humans combine causal relationships between features with similarity relationships between objects. This result may seem incompatible at first with previous studies which suggest that causal inferences often dominate similarity-based inferences (Lassaline, 1996; Wu & Gentner, 1998; Rehder, 2006; Hayes & Thompson, 2007). Experiment 3 of Rehder (2006) is a representative example. In this experiment, participants were presented with a source object with features C and E and told that C caused E. They then had to decide whether a target object also had feature E. Responses were primarily shaped by whether or not the target object had feature C, and there was only a small effect of the overall similarity between the source and target objects. Rehder (2006) uses this experiment to support his overall conclusion that causal reasoning and similarity-based reasoning often compete, but the results of this experiment are consistent with the predictions of the SC model. First, the model can account for the small but statistically significant effect of similarity. Second, the model can explain why the effect of similarity is relatively small in this case. If C is the only possible cause of E, and if the relationship between C and E is near-deterministic, then the SC model predicts that the similarity between source and target is relatively uninformative about whether the target has feature E. This prediction is a consequence of the taxonomic influence principle identified above, which suggests that similarity relationships are used only when they are informative about the distribution of unobserved but causally relevant variables. Other studies where causal inferences appear to dominate similarity-based inferences also use causal relationships that are plausibly interpreted as near-deterministic (Lassaline, 1996; Wu & Gentner, 1998), and this factor may explain the consistent finding that causal inferences tend to dominate similarity-based inferences (Stephens et al., 2009).

Although our theory accounts for the third experiment presented by Rehder (2006), it does not account for his first experiment. This experiment considers a case where some participants rely on similarity-based reasoning, others rely on causal reasoning, and no individual appears sensitive to both similarity relationships and causal relationships. The SC model can accommodate contexts where similarity-based reasoning appears to dominate causal reasoning, and others where causal reasoning appears to dominate similarity-based reasoning. The model, however, does not explain how a single context could produce both patterns of responses. We return to this issue in the General Discussion, and consider the implications for future models of generalization.

7. Experiment 3: Counterfactual interventions

Our previous two experiments explored how relationships between features and relationships between objects are combined in order to carry out generalization tasks. Our final experiment explores whether relationships between features and relationships between objects are combined in a way that is intrinsically causal. As mentioned earlier, a popular approach to across-feature generalization uses causal Bayes nets to capture relationships between features (Rehder, 2003). Many accounts of across-object generalization, however, are not intrinsically causal, including the similarity coverage model (Osherson et al., 1991) and Sloman's feature-based model (Sloman, 1993). Since accounts of across-feature generalization have emphasized causal knowledge but models of across-object generalization have not, it remains to be seen whether causal knowledge plays a critical role in settings that require generalization across both objects and features.

The combination models we have considered throughout suggest two qualitatively different ways in which people might make causal inferences that draw on multiple systems of knowledge. Suppose, for example, that a reasoner is asked to make predictions about a counterfactual intervention. The OC and DC approaches suggest a strategy where each component model is adjusted to allow for the counterfactual intervention and then the adjusted models are combined. Each component model may support causal reasoning, but once these models are combined the combination does not support causal reasoning in any obvious way. The SC approach suggests a different strategy where the component models are combined in a way that is intrinsically causal. As a result, the combination of the models can be directly used to address causal queries. Note that the SC model is defined over a causal graph structure, and that this structure can be manipulated in the standard way to make inferences about interventions and counterfactuals.

The critical difference between the three combination models is whether knowledge is combined at the level of predictions (the OC model), the level of probability distributions (the DC model) or the level of causal structures (the SC model). Experiments 1 and 2 suggest that all approaches can account for generalization to some extent, but the SC model may be uniquely able to predict certain inferences that rely on computations defined over a causal structure. Inferences about counterfactual interventions are one candidate (Pearl, 2000; Sloman & Lagnado, 2005; Rips, 2010) and our final experiment explores whether human counterfactual inferences rely on a causal structure that simultaneously incorporates relationships between features and relationships between objects.

7.1. Participants

32 CMU undergraduates participated in this experiment for course credit. All participants were drawn from the general CMU participant pool.

7.2. Materials

Experiment 3 used six feature structures F, each of which captures a relationship between a cause feature and an effect feature. The six structures were identical except that different pseudo-biological labels were used for each pair of features. Here we use f_1 to refer to each cause feature and f_2 to refer to each effect feature. For all six structures, participants were told that f_2 is known to be produced by several pathways, and that the most common pathway begins with f_1, which directly stimulates production of f_2.

7.3. Procedure

Participants were asked to reason about the same four animals used in Experiments 1 and 2. The experiment began with three preliminary tasks. The first task was the taxonomic task used in Experiments 1 and 2. The second and third tasks were an intervention task and an observation task, and each task used one of the six feature structures described in the previous section. In the intervention task, participants were told that "earlier in the day the mouse was injected with a syringe full of f_2". They were told that the mouse had subsequently tested positive for f_2, and were asked to predict the outcomes of the seven remaining tests involving the four animals and the two features. In the observation task, participants were told that the mouse had tested positive for f_2 and were asked to predict the outcomes of the seven remaining tests. The intervention and observation tasks were included in order to introduce the notion of causal interventions, and to give participants a chance to reflect on whether observations and interventions support different kinds of inferences. The order of these tasks was counterbalanced across participants.

After the preliminary tasks, participants were given four tasks where they were asked to make inferences about counterfactual interventions. In each case participants were presented with a complete matrix of objects by features. The four matrices used are shown in Figure 12. In each case, participants were asked to imagine that earlier in the day the mouse had been injected with a syringe full of enzyme f_1. They then rated the probability that the mouse would have tested positive for enzyme f_2 on a seven point scale where 1 was labeled "very unlikely" and 7 was labeled "very likely." The order of the four counterfactual tasks was counterbalanced across participants.

The preliminary intervention task and the counterfactual tasks both ask participants to think about cases where the mouse is injected with an enzyme "earlier in the day" and is later tested for the presence of one or more enzymes. The time interval between the injection and the tests is critical for the counterfactual task—the biological process by which f_1 causes f_2 presumably takes some time, which means that it makes sense to ask participants about a test for f_2 that is carried out some time after the counterfactual intervention. The intervention task, however, could be improved by asking participants about tests carried out immediately after the intervention. As a result, the intervention task is not ideal for assessing how participants reason about interventions, and is best viewed as a preliminary task that helps to set up the cover story for the counterfactual tasks.

Two of the preliminary tasks and all four of the counterfactual tasks used a feature structure where f_1 causes f_2. As for experiments 1 and 2, this information was reinforced by showing participants 20 cards for each feature structure. The distribution of cards appears in Table A.3, and is consistent with the functional causal model shown in Figure B.20e. As for Experiments 1 and 2, participants could study these cards for as long as they liked, but the cards were removed before they proceeded with the

Task 1

	f_1	f_2
mouse	1	0
rat	1	1
squirrel	1	1
sheep	1	1

Task 2

f_1	f_2
0	0
1	0
1	0
1	1

Task 3

f_1	f_2
0	0
1	1
1	1
1	1

Task 4

f_1	f_2
0	1
1	0
1	0
1	1

Figure 12: Observations for the four counterfactual tasks. In each case participants are asked to decide whether the mouse would have tested positive for f_2 if it had been injected with f_1.

experiment.

7.4. Model predictions

Suppose first that we are interested only in the mouse which has tested positive for f_1 and negative for f_2. Since the mouse already had enzyme f_1, injecting it with f_1 would probably have made little difference, and the mouse would probably still have tested negative for f_2. Suppose next that the mouse tests negative for f_1 but positive for f_2. If we had intervened and injected the mouse with f_1 it is reasonable to expect that the mouse would still have tested positive for f_2. Suppose finally that the mouse tests negative for both f_1 and f_2. Since f_1 causes f_2, we might expect that injecting the mouse with f_1 would have made it more likely that the test for f_2 would have been positive.

All of the inferences just described can be captured by working with a functional causal model that captures the relationship between f_1 and f_2 (Pearl, 2000). The first step is to create a *twin* graph that includes nodes for the counterfactual values of f_1 and f_2. In Figure 13a.i, these counterfactual nodes are labeled g_1 and g_2. Note that the parents of g_1 and g_2 correspond to the parents of f_1 and f_2, consistent with the idea that the causal mechanisms in the counterfactual world match the causal mechanisms in the actual world. We can now reason about a counterfactual intervention on f_1 by using the *manipulated graph* in Figure 13a.ii. All incoming edges to g_1 have been broken to capture the idea that variable g_1 is set by an intervention, and that observing the value of g_1 therefore provides no information about the values of its parent variables. Inferences about any other variables can now be made by carrying out Bayesian inference over the manipulated graph. In particular, we can compute the probability that $g_2 = 1$, which represents the probability that the mouse would have tested positive for f_2 after being injected with f_1.

The second plot in Figure 14 shows the predictions when the feature model is adjusted as just described to allow for the four counterfactual interventions. Since the feature model assumes that the rows of the object-feature matrix are conditionally independent given the feature structure, the predictions of this model can be computed by using the twin graph in Figure 13a.i to reason about the mouse in isolation. Note that the predictions for tasks 2 and 3 are identical—since the feature vectors for each animal are conditionally independent, information about the rat, the squirrel and the sheep does not influence counterfactual predictions about the mouse.

The third plot shows predictions according to the object model. Since the object model is not formulated as a functional causal model, the approach to counterfactual reasoning summarized by Figure 13a cannot be directly applied. The prediction of the object model, however, is straightforward. Since the features are assumed to be conditionally independent given the object structure, a counterfactual intervention on f_1 would have had no influence on f_2, which means that predictions about f_2 should track the actual values of f_2. In other words, adjusting the object model to allow for the counterfactual intervention leads to no change in the f_2 value observed for the mouse.

The fourth plot shows predictions of the OC model. These predictions are computed by adjusting the feature and object models to allow for the counterfactual intervention, then averaging the predictions of these adjusted models. The OC predictions therefore correspond to a weighted average of the predictions for the feature and object models, and are qualitatively similar to the predictions of the feature model.

The fifth plot shows predictions according to the DC model. As already described, the adjusted object model assigns a probability of 1 to the hypothesis that the counterfactual f_2 value for the mouse is identical to the f_2 value actually observed. As a result, the predictions of the DC model are identical to the predictions of the object model. Since the adjusted object model assigns probability mass to just one hypothesis, multiplying the distributions for the adjusted object and feature models produces a distribution that assigns nonzero probability mass to just one hypothesis—the same hypothesis favored by the adjusted object model.

The final plot shows predictions of the SC model. Figure 13b.i shows

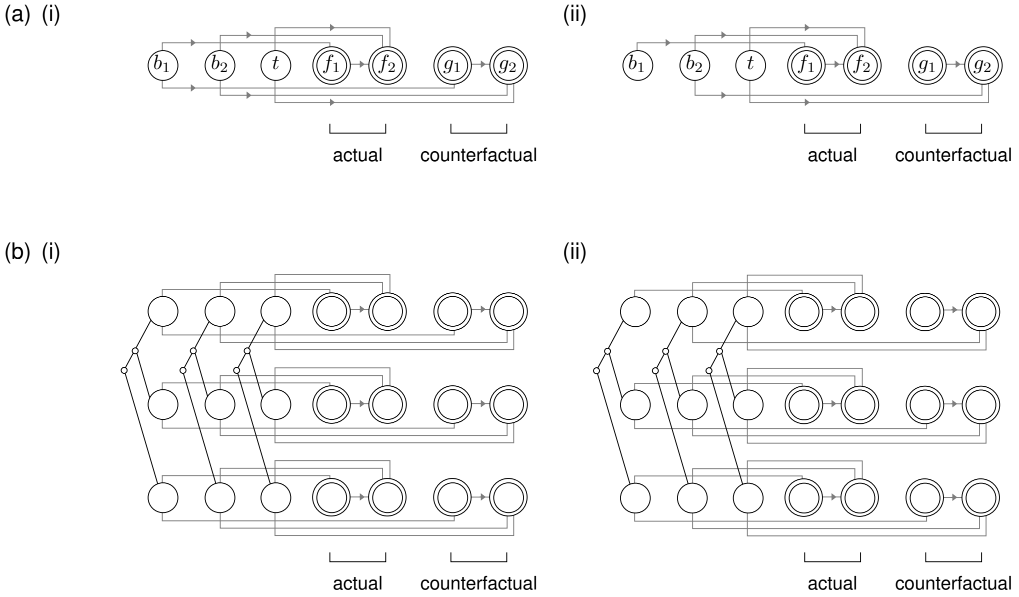


Figure 13: Counterfactual reasoning (a) (i) The causal model from Figure 4b has been supplemented with two additional nodes (g_1 and g_2) which represent counterfactual values of f_1 and f_2. A manipulated graph for reasoning about the value that f_2 would have taken if an intervention had fixed the value of f_1. (ii) The SC model from Figure 8 has been supplemented with six additional nodes that represent counterfactual values of the two features for the three objects. (ii) A manipulated graph for reasoning about what would have happened if the f_1-value for object o_1 had been fixed by an intervention.

how the SC model in Figure 8 can be converted into a twin graph. Manipulating this graph as shown in Figure 13b.ii allows us to make inferences about a counterfactual situation where o_1 is injected with enzyme f_1. Model predictions for all four of the counterfactual tasks in Experiment 3 are shown in Figure 14. The first task is a case where the mouse already has enzyme f_1 in the actual world. A counterfactual manipulation where the mouse is injected with f_1 should therefore make little difference, and the mouse should still test negative for f_2. The second task is a case where the mouse tests negative for f_1 and f_2. Note that the rat and the squirrel have f_1 but not f_2, suggesting that the mouse would also have tested negative for f_2 even if injected with f_1. The third task is similar, except that now the rat and the squirrel have both f_1 and f_2, which suggests that the mouse would also test positive for f_2 if injected with f_1. The final task is a case where the mouse already has f_2. Injecting the mouse with f_1 should therefore make

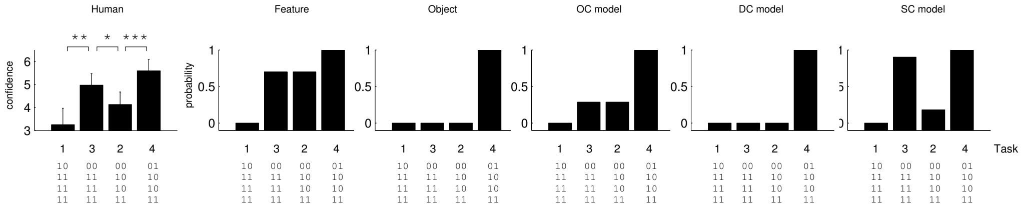


Figure 14: Results and model predictions for Experiment 3. The object-feature matrices for each task are reproduced from Figure 12.

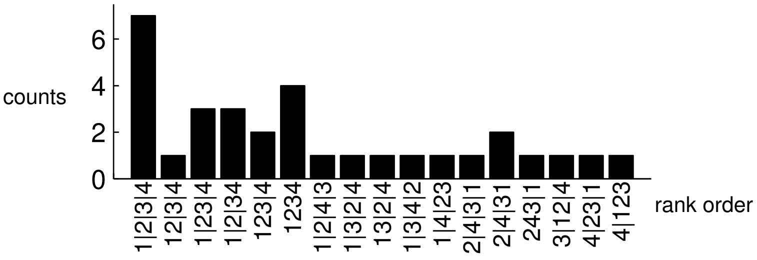


Figure 15: Individual differences analysis for Experiment 3. The classifications are based on the rank order of responses to the four tasks, and vertical lines indicate differences in rank. For example, 1|2|3|4 matches the rank order predicted by the SC model, and 1234 indicates that the same response was given to all four tasks.

little difference and the mouse should still test positive for f_2.

7.5. Results

Responses to the preliminary taxonomic task were similar to the responses to the taxonomic tasks in Experiments 1 and 2 and are not described further. Responses to the remaining two preliminary tasks suggested that observations and interventions were treated differently overall. Because these results do not differentiate among the models considered in this paper, full details are provided in Appendix Appendix A.

Average responses for the four counterfactual tasks are shown in Figure 14. Out of the five models in Figure 14, the rank order of the four responses is consistent only with the SC model. The most critical comparison occurs between tasks 2 and 3, and a sign test indicates that responses for task 3 are significantly greater than responses for task 2 ($p < 0.05$). The observed feature values for the mouse are the same for both tasks, and the first four models therefore make identical predictions about these tasks.

Only the SC model correctly predicts that observed feature values for the other animals will shape counterfactual inferences about the mouse. Sign tests also indicate that responses for task 3 are significantly greater than responses for task 1 ($p < 0.05$), and that responses for task 4 are significantly greater than responses for task 2 ($p < 0.001$). Note that both pairs differ only with respect to the observed feature values for the mouse.

Figure 15 summarizes the responses of individual participants, and shows that the modal response is consistent with the SC model. Seven out of 32 participants generated responses that match the rank order predicted by the model, and an additional 13 participants generated responses that collapse one or more of the distinctions present in the modal response.

Taken overall, our results suggest that people are capable of reasoning about counterfactual interventions in settings that draw on relationships between objects and relationships between features. Only the SC model accounts for the full pattern of results, which suggests that people combine multiple models at the level of causal structures rather than at the level of predictions or probability distributions. Our data therefore suggest that relationships between objects and relationships between features are combined in a way that is intrinsically causal.

8. General Discussion

We formalized generalization as the problem of reasoning about object-feature matrices and evaluated three computational approaches that address this problem by incorporating both relationships between objects and relationships between features. All three models rely on a graph structure over objects and a graph structure over features, and our data suggest that combining these structures directly (the SC approach) provides a better account of human reasoning than combining the distributions induced by these structures (the DC approach) or combining the outputs produced by these structures (the OC approach).

Our first two experiments suggested that humans readily combine relationships between objects and relationships between features. The SC model accounts well for the results of these experiments and performs substantially better than alternatives that rely on object knowledge alone or feature knowledge alone. The model also accounts for an important qualitative effect that is inconsistent with the OC model. Given a partially-observed object-feature matrix, the OC model predicts that people are able to make informed inferences only about entries that belong to the same row or column as an observed entry. Experiment 1, however, showed that participants make informed inferences about an entire object-feature matrix after observing a single entry in the matrix.

Experiments 1 and 2 suggested that the SC model provides a better quantitative account of human reasoning than the DC model, but even so the DC model accounts relatively well for the results of these experiments. Experiment 3 explored a setting where the predictions of the SC model depart sharply from both the DC and the OC models. Our data suggest that the SC model alone is able to explain why inferences about a counterfactual intervention on a given object (e.g. a mouse) are shaped by the observed features of other objects (e.g. a rat and a squirrel).

Our comparison between the three combination models suggests that human knowledge about relationships between objects is tightly integrated with knowledge about relationships between features. The OC model explores the hypothesis that these two forms of knowledge are captured by distinct modules, but our data suggest that a modular approach will struggle to explain how humans make inferences about an entire object-feature matrix given only a handful of observations. Multiplying probability distributions provides one way to integrate systems of knowledge, but our data suggest that combining structured representations will provide the best way to explain how different systems of knowledge work together.

Although the SC model is constructed by combining existing models of inductive reasoning it goes beyond these models in several ways. First, it provides a unified view of two inductive problems—across-object and across-feature generalization—that are often considered separately. Second, unlike previous accounts of across-feature generalization, it acknowledges the importance of unknown but causally relevant variables, and uses taxonomic relationships to constrain inferences about the effects of these variables. Third, unlike most previous models of across-object generalization, the model can handle novel features that are causally linked to known features. Finally, the model helps to explain how counterfactual inferences are made in settings that simultaneously draw on relationships between objects and relationships between features.

8.1. Generalization and causal reasoning

Studies of inductive generalization in adults (Lee & Holyoak, 2008; Rehder, 2009) and children (Carey, 1985; Opfer & Bulloch, 2007; Hayes & Thompson, 2007) have suggested that inductive inferences often rely on causal theories. Our approach is consistent with this general claim. For expository convenience we have emphasized the distinction between causal relationships between features and taxonomic relationships between objects, but relationships between objects will often have a causal interpretation. A tree-structured taxonomy, for example, is a simple representation of the causal process that generated biological species—the process of evolution (Kemp & Tenenbaum, 2009). The graphical model in Figure 8 can therefore be viewed as a causal theory that incorporates causal relationships between features and causal relationships between species.

Our comparison between the DC and SC approaches supports the idea that causal reasoning plays an important role in human generalization. The most fundamental difference between these approaches is that the SC approach alone combines models in a way that is intrinsically causal. The DC approach is intuitive and accounts fairly well for our first two experiments. Our third experiment, however, demonstrates that the approach fails to account in full for human inferences about counterfactuals.

Although causal reasoning appears to contribute to many inferences, studies suggest that humans often rely on causal theories that are fragmentary or incomplete (Rozenblit & Keil, 2002). Results of this kind challenge causal accounts of generalization—how can humans make successful causal inferences if they do not understand the causal mechanisms that apply in any given setting? Our approach suggests a partial answer. Even though detailed causal theories are typically unavailable, general causal principles can still support accurate causal inferences. One such principle holds that objects with similar observable features are often influenced by similar causal factors, and this similarity can support causal reasoning even if the actual causal mechanisms remain unknown. Our work therefore begins to explain one of the most impressive aspects of human causal reasoning—the ability to make successful inferences in the presence of many hidden variables.

8.2. Combining knowledge structures

There are many previous accounts of inductive reasoning, including accounts that focus on relationships between objects (Osherson et al., 1991) and accounts that focus on relationships between features (Rehder, 2003). A distinctive aspect of our account is that it incorporates these two kinds of relationships. Hadjichristidis et al. (2004), Holyoak et al. (2010) and Stephens et al. (2009) have also developed accounts that incorporate relationships between objects and relationships between features, and here we compare these accounts to our own.

Hadjichristidis et al. (2004) focus on a problem similar to across-object generalization (Figure 1a), but consider arguments where the premise and conclusion refer to categories (e.g. mice and rats) rather than individual objects (e.g. a specific mouse and a specific rat). They propose that the strength of an argument depends on the similarity of the conclusion category to the premise category and the causal centrality of the feature involved. The causal centrality of a feature depends on its relationships to other features, and the account of Hadjichristidis et al. (2004) therefore incorporates both relationships between categories and relationships between features. Although the work of Hadjichristidis et al. (2004) is directly relevant to the problem of combining multiple knowledge structures, it differs from our approach in least two respects. First, our model addresses the general problem of completing a partially-observed matrix of objects by features, but Hadjichristidis et al. (2004) consider only the problem of across-object generalization. For example, their account would need to be supplemented in order to handle the generalization problem in Figure 1c. Second, Hadjichristidis et al. (2004) do not provide a computational model that indicates how causal centrality and similarity should be combined. Sloman et al. (1998) have developed a computational account of causal centrality, but additional work is needed to specify how this model might be combined with a formal model of similarity-based reasoning.

Holyoak et al. (2010) have developed a computational model that integrates causal inference with analogical reasoning and will be referred to here as the causal-analogical model (CA model for short). The CA model is motivated by the idea that a causal model learned for a source object can influence the model used to reason about the features of an analogous target object. Although we focused on taxonomic relationships between objects rather than analogical mappings, the SC model is motivated by a similar idea. We believe, however, that the two models have complementary strengths. Unlike the SC model, the CA model is designed to handle cases where the causal models for two objects may have different structures—for example, cases where there is no perfect correspondence between the causal edges in the graphs for source and target. Unlike the CA model, the SC model captures the idea that the causal parameters associated with two objects can be more or less similar depending on the overall taxonomic relationship between the two objects.[4] Ultimately it may be possible to

[4] Holyoak et al. (2010) assume that base rate parameters for the source and target are identical, and also assume that any causal relationship in the target that is analogous

develop a model that combines the strengths of both approaches, and that uses taxonomic relationships to shape inferences about both the structure and the parameters of the causal model for a given object.

In a separate line of work, Stephens and colleagues (Stephens et al., 2009) have used a paradigm similar to ours to study integrated reasoning and have developed an integrated causal model that is different from ours. The SC model combines a functional causal model and an object-based taxonomy by introducing a copy of the taxonomy for each root variable in the functional model. Stephens and colleagues use the same approach to combine a taxonomy with a causal model that incorporates probabilistic relationships—in other words, they introduce a copy of the taxonomy for each root variable in the probabilistic causal model. We will refer to the resulting model as the *probabilistic structure combination* model, or PSC model for short. The PSC model makes similar predictions to the SC model in some settings but suffers from two limitations. First, the PSC model predicts that taxonomic relationships have no further role to play once the root variables in the probabilistic model are observed for each object, and therefore cannot account for the result of our second experiment. Second, the PSC model does not rely on functional causal models, and is therefore unable to account for the counterfactual inferences explored in our third experiment. Given both of these limitations, we believe that the SC model should be preferred to the alternative that Stephens and colleagues consider.

The models discussed in this section combine knowledge structures in slightly different ways, and there are presumably many other ways in which knowledge structures could be combined. The combination strategy preferred by humans could well depend on the context, but our work suggests two basic principles that may be widely applicable. First, causal considerations will often dictate how knowledge representations should be combined. For example, the SC and DC models both incorporate the same components, but the SC model alone combines these components in a way that respects causality. As a result, the SC model provides a more accurate

to a causal relationship in the source has the same causal strength in these two cases. Equation 3 in their paper allows the possibility that corresponding causal parameters may differ, but the model as implemented appears to assume that corresponding causal strengths are identical as described in Equations 4 and A4 of their paper. In order to account for the results of our experiments, we believe that the CA model would need to use a version of their Equation 3 which captures the idea that the source and target are likely to have similar causal parameters to the extent that the two are taxonomically related.

account of inferences about counterfactuals. The second basic principle is that probabilistic inference provides a useful general framework for combining different kinds of knowledge. Different probabilistic models may capture qualitatively different forms of knowledge, but probability theory provides a *lingua franca* for binding them together.

8.3. When are knowledge structures combined?

As applied to our experiments, the structure combination model relies on two structures: the first captures taxonomic relationships between objects and the second captures causal relationships between features. Commonsense knowledge, however, includes many other kinds of structures, and models of generalization should ultimately aim to incorporate all of these structures. For example, inductive generalizations about animals may draw on ecological knowledge (animals that share the same habitat are likely to catch the same diseases) and social knowledge (animals chosen as pets are likely to share certain features).

Here we focused on problems where multiple knowledge structures are available and the best inferences tend to be compatible with all them. In other settings multiple knowledge structures may be relevant, but the best inference in any case may not depend on all of them. Suppose, for example, that inferences about biological features depend on a structure that captures taxonomic relationships between animals, and a second structure that captures ecological categories (Shafto et al., 2011). Inferences about some features will depend on both structures—for example, "has blubber" is a feature that tends to be shared only by marine mammals. Other features will depend on only one structure: for example, "is warm-blooded" is shared by all mammals, and "has a streamlined body shape" is shared by many marine creatures. A comprehensive account of generalization should be able to select the structures that are relevant to a given inference, and to flexibly combine these structures when needed.

The need for an account of structure selection becomes especially apparent as the number of different structures increases. Experiment 1 of Rehder (2006) suggests, however, that structure selection can play an important role even in relatively simple contexts. As part of Rehder's study, participants were shown a novel exemplar of a category and were told that this exemplar had a feature E that was caused by one of the characteristic features of the category. They were then asked to estimate the proportion of category members that had feature E. Two distinct strategies were observed—half of the participants relied on the fact that E was causally related to a characteristic category feature, and the other half relied on the similarity between the novel exemplar and the category prototype. No single participant gave responses that were sensitive to both the causal information and the similarity between the exemplar and the category prototype.

Rehder's data suggest that there are settings where multiple knowledge structures are relevant, but where participants make inferences based on the single structure that first comes to mind. Our data, however, suggest that there are similar settings where the majority of participants combine multiple knowledge structures. An important direction for future work is therefore to characterize the factors that determine whether reasoners select a single knowledge structure or attempt to combine multiple structures. One relevant factor is the salience of the structures involved. Unlike Rehder's first experiment, all of our experiments included a preliminary task that drew upon causal knowledge and a preliminary task that drew on taxonomic knowledge. As a result, both kinds of knowledge were presumably relatively salient when participants were completing the generalization tasks of primary interest. A second relevant factor is the principle of least effort (Zipf, 1949): even if multiple structures are salient, participants presumably incur the cognitive cost of combining them only if they believe that a single structure is unlikely to provide acceptable answers.[5] Shafto et al. (2007) use a speeded induction task to show that some knowledge structures are easier to access than others, and a similar paradigm could be used to test the proposal that combining multiple structures is relatively demanding.

Although the problem of structure selection is important, it is not addressed by structure combination model developed in this paper. The model applies in cases where the relevant structures have been selected already, and the remaining task is to decide how to combine them. Ultimately, however, it may be possible to develop a more comprehensive computational account where the input to the structure combination module is provided by a computational account of structure selection.

[5] We thank an anonymous reviewer for highlighting the issue of structure selection, and for pointing out that a principle of least effort is likely to be relevant.

9. Conclusion

Humans make many kinds of inductive inferences. Psychologists have made substantial progress by studying each kind of inference in isolation, but should ultimately aim for unifying accounts (Newell, 1989; Kemp & Jern, 2009) that can account for many kinds of inferences. We have taken a step in this direction by developing a model that accounts for across-object and across-feature generalization, including cases where both kinds of generalization must work in tandem. Our model simultaneously draws on taxonomic relationships between objects and causal relationships between features, and our experiments confirm that people are able to combine these two kinds of information.

Although our model incorporates multiple kinds of information, the knowledge captured by this model still falls well short of the complexity of commonsense knowledge. Accounts of generalization will eventually need to grapple with this complexity, and future studies of inductive reasoning will need to explore how many different pieces of knowledge are integrated and composed. Our work suggests that causal representations are useful for combining multiple systems of knowledge, and future studies can aim to use this approach to capture increasingly large systems of commonsense knowledge.

Appendix A. Experimental details

In each condition of our experiments, participants read a description of the causal relationships between a set of enzymes. Participants were then given a set of 20 cards that showed the distribution of these enzymes among a group of 20 unnamed mammals. The cards for each condition are shown in Tables A.1, A.2 and A.3. In the chain condition, for example, most animals that have f_1 also have f_2, and most animals that have f_2 also have f_3.

Experiment 1 included a preliminary task to confirm that participants were familiar with the taxonomic relationships between the four animals in the experiment (rat, mouse, sheep and squirrel). Each question on this test informed participants that one of the animals had tested positive for a novel enzyme, and asked them to predict whether the remaining animals would also test positive for this enzyme. The results (Figure A.16) are consistent with the taxonomic relationships captured by the tree structure O in Figure 9a, and are accurately predicted by a probabilistic model defined over this structure.

f_1	f_2	f_3	Chain	Common effect	Common cause
0	0	0	6	6	8
0	0	1	3	2	1
0	1	0	1	1	1
0	1	1	3	5	0
1	0	0	1	1	1
1	0	1	0	5	2
1	1	0	1	0	2
1	1	1	5	0	5

Table A.1: To reinforce the causal structure described in each condition, participants were shown cards that indicated the presence or absence of three enzymes (f_1, f_2 and f_3) in 20 unnamed animals. The final three columns of the table show card counts for the chain, common effect and common cause conditions.

f_1	f_2	f_3	f_4	Clusters
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	1	0	0	1
1	0	0	0	1
0	0	1	1	2
1	1	0	0	2
1	1	1	0	2
1	1	0	1	2
1	0	1	1	2
0	0	1	1	2
1	1	1	1	3

Table A.2: Card counts for the clusters condition.

f_1	f_2	
0	0	9
0	1	4
1	0	1
1	1	6

Table A.3: Card counts for the tasks in Experiment 3.

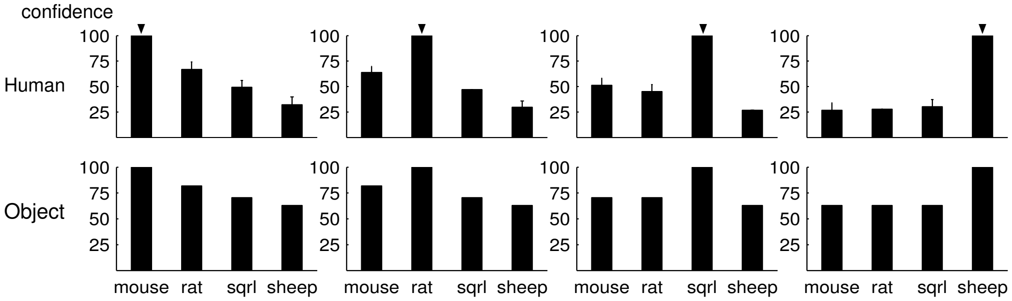


Figure A.16: A preliminary taxonomic task measured pre-existing knowledge about taxonomic relationships between the four animals in our experiment. Each plot in the top row shows human inferences about the distribution of a novel enzyme given that one animal (the column with value 100) is known to have the enzyme. The bottom row shows predictions of the tree-structured probabilistic model in Figure 9.

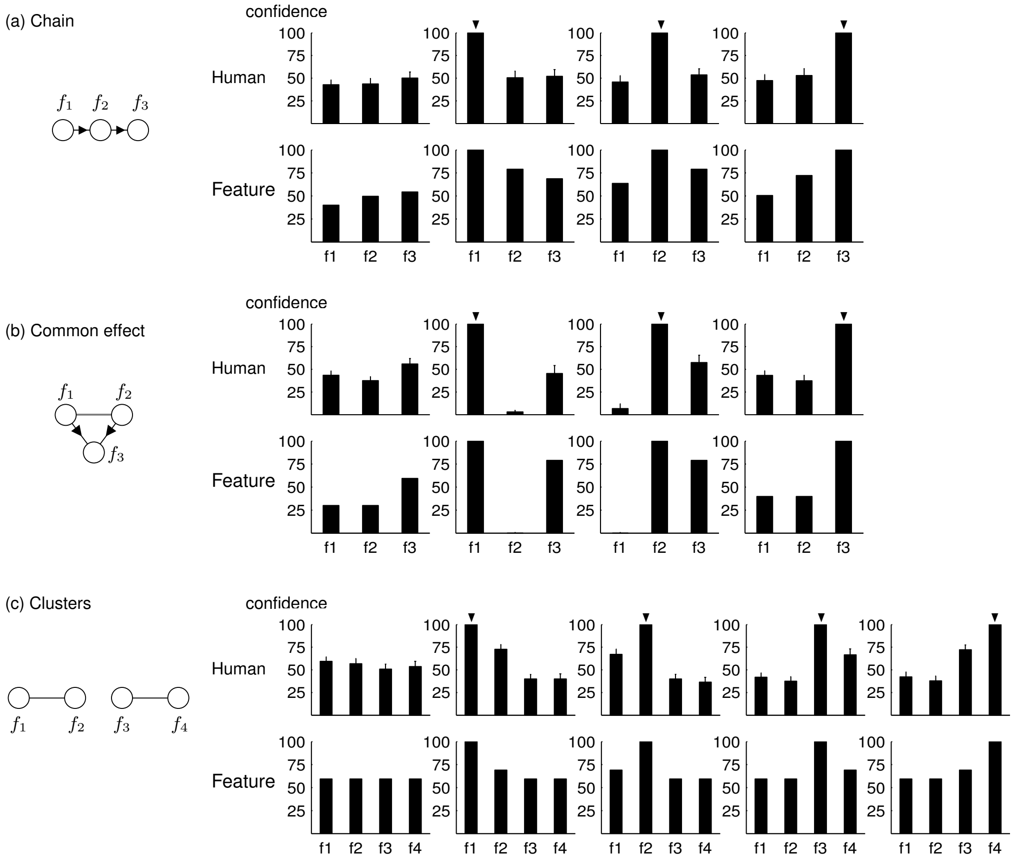


Figure A.17: A second set of preliminary tasks was included to determine whether participants had understood the causal information provided during the training phase. Each plot shows inferences about the enzymes expressed by a novel mammal. The first plot in each row shows inferences about a mammal that has not yet been tested for any of the enzymes. The remaining plots show inferences about a mammal that has tested positive for one enzyme (the column with value 100). The bottom row of each pair shows predictions of the functional model used to model each condition.

Experiment 1 also included a set of preliminary causal tasks to confirm that participants had learned the feature structures F for each condition. The results (Figure A.17) suggest that participants had learned these structures, and are well predicted by the functional models used to model each condition.

Experiment 3 included a preliminary observation task and a preliminary intervention task. Model predictions and mean responses to these tasks are shown in Figure A.18. All models predict that the two tasks will be treated differently. For example, *observing* that the mouse has f_2 should suggest that the mouse is likely to have f_1 and that the rat is likely to have f_2, but *intervening* so that the mouse has f_2 should not provide any information about the other entries in the object-feature matrix. We evaluated the prediction that the observation and the generalization tasks are psychologically different by using a two-way ANOVA with repeated measures to explore the relationship between the task (observation or intervention) and the seven missing entries in the object-feature matrix. There were main effects of task (F = 12.4, $p < 0.001$) and matrix entry (F = 62.1, $p < 0.001$), but no interaction between task and matrix entry (F = 1.63, $p = 0.13$). The main effect of task suggests that the observation and intervention tasks are treated differently, and we ran follow-up tests to explore the prediction that inferences were stronger for the observation task than the intervention task. Inferences that the mouse had f_1, that the rat had f_1 and that the rat had f_2 were all stronger for the observation task than the intervention task (sign tests yield $p < 0.01$ in all cases).

Although the observation and intervention tasks appear to be treated differently, the generalization gradients for the intervention task are not completely flat as predicted by the models. Analyzing the responses of individual participants suggests that these weak generalization gradients are produced by averaging the responses of two groups, where one group generates decaying generalization gradients and the other generates uniform gradients. Each participant generates a matrix of predictions for 4 objects by 2 features. The columns of these matrices can be classified as decaying (D), uniform (U) or other (O). A column is classified as decaying if the first entry is greater than the last entry and if no entry exceeds the entry in the previous row. A column is classified as uniform if all entries are identical, and all remaining columns are classified as "other." Since the intervention fixes the value of f_2 for the mouse, only the predictions for the remaining three values are used when classifying the second column. Depending on the

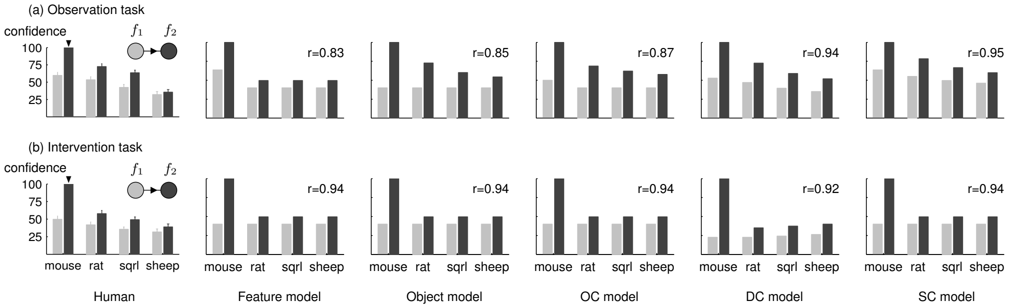


Figure A.18: Results for the preliminary observation and intervention tasks in Experiment 3.

classifications of its two columns, the matrix for each participant is assigned to one of nine possible categories. Counts for these categories are shown in Figure A.19. The most common pattern for the observation task is DD, which indicates that inferences about both f_1 and f_2 decay over the tree. 19 out of 32 participants produced this response, and the mean response for this group of 19 is shown in Figure A.19. The counts for the intervention task reveal two common responses. 12 participants generated DD responses, and eleven generated UU responses. Average responses across these two groups are shown in Figure A.19.

The responses of the UU group in Figure A.19b are consistent with the model predictions in Figure A.18b but the responses of the DD group are not predicted by any of the models. One possible explanation is that participants in the DD group recognized that the situation described is more complex than a simple intervention. The intervention task describes an intervention (earlier in the day the mouse is injected with f_2) and an observation (the mouse tests positive for f_2). If participants assume that any foreign enzyme is quickly broken down, the positive test result provides some evidence that f_2 is naturally present in the mouse's bloodstream, and therefore that the mouse is likely to have f_1 and that the rat is likely to have f_2. Interpreting the task in this way is possible because some time elapses between the injection and the test result, and removing this time interval would allow a cleaner test of how participants reason about interventions. As described in the main text, however, the time interval was deliberately introduced in order to set up the cover story for the counterfactual tasks.

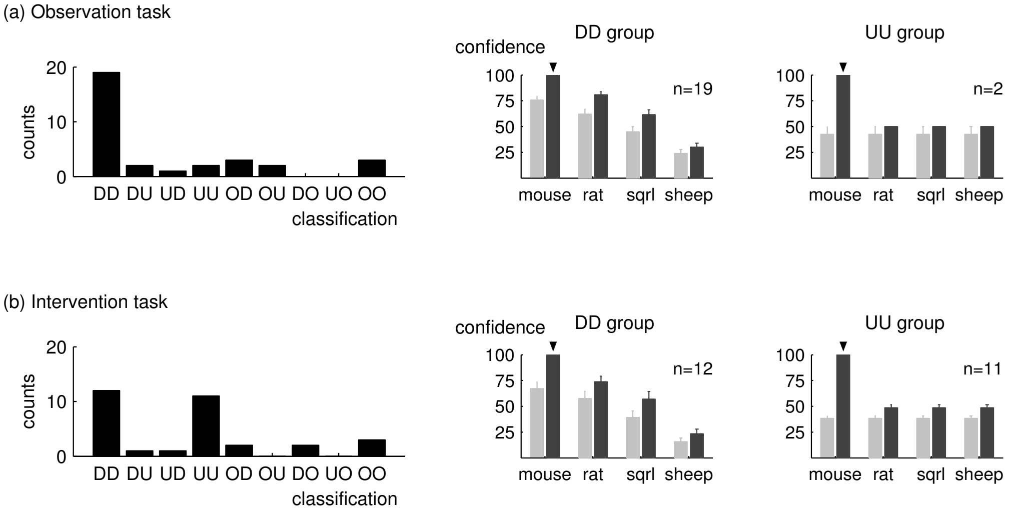


Figure A.19: Individual differences analysis for the preliminary observation and intervention tasks of Experiment 3. The DD group includes participants who generated ratings for both features that decay (D) over the tree. The UU group includes participants who generated ratings for both features that were uniform (U) over the tree.

Appendix B. Modeling details

The SC model makes use of a functional causal model F that captures causal relationships between features. Figure B.20 shows the structures F that were used to generate the training cards for each experiment and to model our experimental data.

References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 1–55.

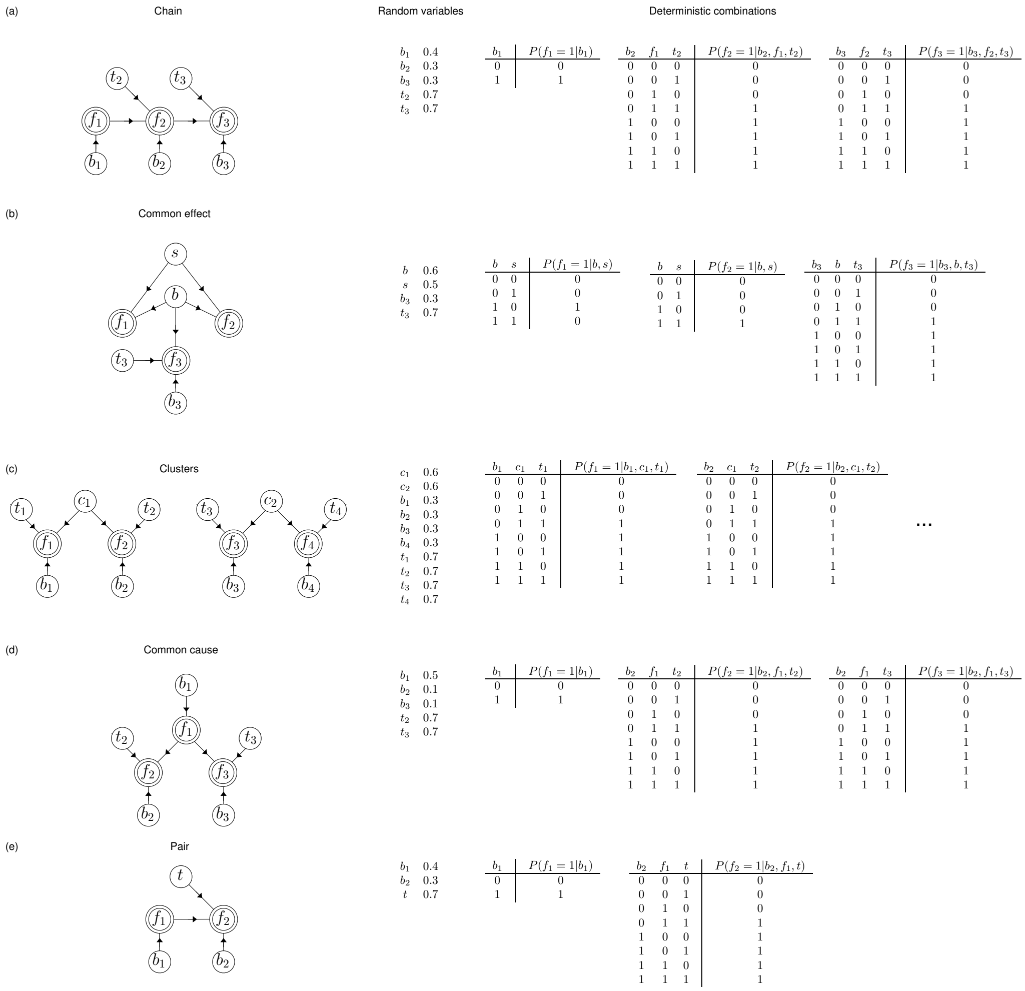
Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to white (2005) and to luhmann and ahn (2005). *Psychological Review*, *112*, 694–706.



(a) Chain

Random variables

b_1	0.4
b_2	0.3
b_3	0.3
t_2	0.7
t_3	0.7

Deterministic combinations

b_1	$P(f_1=1\|b_1)$
0	0
1	1

b_2	f_1	t_2	$P(f_2=1\|b_2,f_1,t_2)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

b_3	f_2	t_3	$P(f_3=1\|b_3,f_2,t_3)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

(b) Common effect

b	0.6
s	0.5
b_3	0.3
t_3	0.7

b	s	$P(f_1=1\|b,s)$
0	0	0
0	1	0
1	0	1
1	1	0

b	s	$P(f_2=1\|b,s)$
0	0	0
0	1	0
1	0	0
1	1	1

b_3	b	t_3	$P(f_3=1\|b_3,b,t_3)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

(c) Clusters

c_1	0.6
c_2	0.6
b_1	0.3
b_2	0.3
b_3	0.3
b_4	0.3
t_1	0.7
t_2	0.7
t_3	0.7
t_4	0.7

b_1	c_1	t_1	$P(f_1=1\|b_1,c_1,t_1)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

b_2	c_1	t_2	$P(f_2=1\|b_2,c_1,t_2)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

...

(d) Common cause

b_1	0.5
b_2	0.1
b_3	0.1
t_2	0.7
t_3	0.7

b_1	$P(f_1=1\|b_1)$
0	0
1	1

b_2	f_1	t_2	$P(f_2=1\|b_2,f_1,t_2)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

b_2	f_1	t_3	$P(f_3=1\|b_2,f_1,t_3)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

(e) Pair

b_1	0.4
b_2	0.3
t	0.7

b_1	$P(f_1=1\|b_1)$
0	0
1	1

b_2	f_1	t	$P(f_2=1\|b_2,f_1,t)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

Figure B.20: Functional causal models used to generate the training cards for each condition and to compute the predictions of the OC, DC, SC and feature models (Figures 10 and 11). In each network, the nodes without parents are independent random variables, and the base rate for each variable is shown. Each remaining node takes a value that is a deterministic function of the values of its parent nodes. (a) The chain structure is equivalent to a noisy-OR network. Variable b_i indicates whether the background cause for f_i is present, and variable t_i indicates whether the mechanism of causal transmission between f_{i-1} and f_i is active. (b) Variables f_1 and f_2 in the common effect structure are mutually exclusive. If the background cause b is present, then exactly one of these variables will be true, and this choice is a deterministic function of the switching variable s. (c)-(d) The cluster and common cause structures are equivalent to noisy-OR networks. Only two of the deterministic conditional probability distributions for the cluster structure are shown. (e) The structure used for Experiment 3.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, (pp. 370–379).

Danks, D. (2003). Equilibria of the Rescorla-Wagner Model. *Journal of Mathematical Psychology*, *47*, 109–121.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fodor, J. A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*, 280–291.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354–384.

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

Hayes, B., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, (pp. 278–292).

Hayes, B. K., & Thompson, S. P. (2007). Causal relations and feature similarity in children's inductive reasoning. *Journal of Experimental Psychology: General*, *136*, 470–484.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 411–422.

Hinton, G. E. (2000). Modelling high-dimensional data by combining simple experts. In *Proceedings of the 17th National Conference on Artificial Intelligence*. AAAI Press.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: a theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*, 702–727.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*, 754–755.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, *3*, 79–87.

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*, 216–233.

Kemp, C., & Jern, A. (2009). A taxonomy of inductive problems. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 255–260). Austin, TX: Cognitive Science Society.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 754–770.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 1111–1122.

Lombardi, L., & Sartori, G. (2007). Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, *57*, 101–125.

Luhmann, C. C., & Ahn, W. (2005a). BUCKLE: A model of unobserved cause learning. *Psychological Review*, *114*, 657–677.

Luhmann, C. C., & Ahn, W. (2005b). The meaning and computation of causal power: comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review*, *112*, 685–693.

Macario, J. F. (1991). Young children's use of color in classification: foods and canonically colored objects. *Cognitive Development*, *6*, 17–46.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*, 225–252.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

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

Murphy, G. L., & Ross, B. H. (2010). Category vs object knowledge in category-based induction. *Journal of memory and language*, *63*, 1–17.

Murphy, K. (2001). The Bayes Net Toolbox for MATLAB. *Computing science and statistics*, *33*, 1786–1789.

Newell, A. (1989). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Opfer, J. E., & Bulloch, M. J. (2007). Causal relations drive young children's induction, naming and categorization. *Cognition*, (pp. 206–217).

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science*, *15*, 251–269.

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

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

Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory and Cognition*, *34*, 3–16.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, (pp. 301–344).

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal*

Learning and Verbal Behavior, *14*, 665–681.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*, 175–221.

Rozenblit, L. R., & Keil, F. C. (2002). The missunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

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

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: causal determinism and children's inferences about unobserved causes. *Child Development*, *77*, 427–442.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 641–649.

Shafto, P., Coley, J. D., & Baldwin, D. (2007). Effects of time pressure on context-sensitive property induction. *Psychonomic Bulletin and Review*, *14*, 890–894.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*, 175–192.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*, 1–25.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.*. Oxford: Oxford University Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, (pp. 5–39).

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189–228.

Stephens, R. G., Navarro, D. J., Dunn, J. C., & Lee, M. D. (2009). The effect of causal strength on the use of causal and similarity-based information in feature inference. In W. Christensen, E. Schier, & J. Sutton (Eds.), *Proceedings of the 9th conference of the Australasian society for Cognitive Science*. Sydney: Macquarie Center for Cognitive Science.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based bayesian models of inductive reasoning. In A. Feeney, & E. Heit (Eds.), *Inductive reasoning: experimental, developmental and computational approaches* (pp. 167–204). Cambridge: Cambridge University Press.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, *124*, 181–206.

Wu, M. L., & Gentner, D. (1998). Structure in category-based induction. In M. A. Gernsbacher, & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1154–1158). Mahwah, NJ: Erlbaum.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.