COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017, Suppl. 5) 1062–1092

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12455

Generative Inferences Based on Learned Relations

Dawn Chen,[a] Hongjing Lu,[ab] Keith J. Holyoak[a]

[a]*Department of Psychology, University of California, Berkeley*
[b]*Department of Statistics, University of California, Los Angeles*

Received 7 November 2015; received in revised form 26 April 2016; accepted 2 September 2016

Abstract

A key property of relational representations is their *generativity*: From partial descriptions of relations between entities, additional inferences can be drawn about other entities. A major theoretical challenge is to demonstrate how the capacity to make generative inferences could arise as a result of learning relations from non-relational inputs. In the present paper, we show that a bottom-up model of relation learning, initially developed to discriminate between positive and negative examples of comparative relations (e.g., deciding whether a sheep is larger than a rabbit), can be extended to make generative inferences. The model is able to make quasi-deductive transitive inferences (e.g., "If *A* is larger than *B* and *B* is larger than *C*, then *A* is larger than *C*") and to qualitatively account for human responses to generative questions such as "What is an animal that is smaller than a dog?" These results provide evidence that relational models based on bottom-up learning mechanisms are capable of supporting generative inferences.

Keywords: Relation learning; Transitive inference; Deduction; Induction; Hypothetical reasoning; Bayesian models

1. Introduction

1.1. Generating inferences based on relations

A hallmark of human intelligence is the ability to learn and make inferences based on *relations* between entities, rather than solely on features of individual entities (for reviews

Correspondence should be sent to Dawn Chen, Institute of Cognitive and Brain Sciences, University of California, Berkeley, 3210 Tolman Hall MC 1650, Berkeley, CA. E-mail: sdawnchen@gmail.com

This paper is based in part on a Ph.D. dissertation completed at the UCLA Department of Psychology by DC, under the direction of HL and KH. A preliminary report of part of this research was presented at the 35th Annual Conference of the Cognitive Science Society (Berlin, August 2013). MATLAB code for the simulations reported here is available from the website of the UCLA Computational Vision and Learning Lab (http://cvl.psych.ucla.edu/generative.zip).

see Gentner & Forbus, 2011; Halford, Wilson, & Phillips, 2010; Holyoak, 2012). The ability to reason with relations increases over the course of cognitive development (e.g., Gentner & Rattermann, 1991; Halford, 1992), but it is already firmly established in preschool children (e.g., Gentner, 1977; Glass, Holyoak, & Kossan, 1977; Holyoak, Junn, & Billman, 1984). Core competences associated with higher level human cognition, including both inductive reasoning (e.g., analogy) and deductive reasoning (e.g., transitive inference), depend on the ability to acquire and manipulate relational representations.

Relational reasoning has several interrelated properties. Relations are *compositional* (Halford et al., 2010), in that constituent entities from which relations are constructed retain their identities and can be accessed within the relational structure. This property underlies the *generativity* of relations: A partial description of relations between objects can be extended to answer questions about relations between actual or hypothetical objects. For example, comparative relations such as *larger* exhibit the logical properties of transitivity and asymmetry, supporting deductions such as "If *A* is larger than *B* and *B* is larger than *C*, then *A* is larger than *C*," for arbitrary instantiations of the objects *A*, *B*, and *C*. Children as young as 5 or 6 years can make such transitive inferences reliably (Goswami, 1995; Halford, 1984; Kotovsky & Gentner, 1996).

The ability to make generative inferences based on relational knowledge thus appears to be a key aspect of human intelligence (Penn, Holyoak, & Povinelli, 2008), which needs to be accounted for by any model of relational reasoning that aspires to generality. The basic goal of the present paper is to describe a computational model of relational processing that addresses this requirement.

1.2. How can relations be learned?

The first step toward explaining how relations can be used to make generative inferences is to provide an account of how relations can be acquired in the first place. Doubtless, some relations are constructed in a top-down fashion, but there is strong evidence that some relations acquired early by children are learned through bottom-up processes (Mandler, 1992). For example, children seem to acquire comparative relations such as *larger than* in stages, first learning features of individual objects, then extracting specific attributes of individual objects (e.g., a size value), and eventually linking attributes of paired objects to form a binary relation (Smith, 1989). Thus, a basic problem for cognitive science is: How can relations be acquired from *non-relational* inputs?

A few models based on neural-network architectures (Doumas, Hummel, & Sandhofer, 2008; Smith, Gasser, & Sandhofer, 1997) have had some success in modeling bottom-up relation learning. However, it is difficult to fully evaluate the adequacy of proposed models of relation learning without first controlling the nature of the elementary inputs on which learning is based. For example, a well-known limitation of models of analogy (for which relational knowledge is central) is that modelers typically create their own "toy" input representations, which may be inadvertently tailored so as to reduce task difficulty (Chalmers, French, & Hofstadter, 1992). In modeling basic relation learning, it is critical to ensure that the non-relational inputs on which learning operates are

autonomously created (rather than hand-coded by the modeler) and are of realistic complexity. When a model of relation learning is forced to operate on realistic inputs, theoretical issues that might have gone unnoticed with simpler inputs are more likely to be brought to the fore.

We recently developed a discriminative Bayesian model termed *Bayesian Analogy with Relational Transformations* (*BART*) that can learn simple relations in a bottom-up fashion (Lu, Chen, & Holyoak, 2012). The inputs to BART are non-relational feature vectors derived independently of the model. A number of alternative feature representations have been used as inputs to BART, of which the richest and most complex feature representations were derived by applying the topic model (Griffiths, Steyvers, & Tenenbaum, 2007) to the English Wikipedia corpus. The output of the topic model is used to create real-valued feature vectors for individual objects. The BART model represents an *n*-ary relation as a function that takes a feature vector for an ordered set of *n* objects as its input and outputs the probability that these objects instantiate the relation. The model learns a representation of the relation from labeled examples (typically positive examples only), and then applies the learned representation to determine whether the relation holds for novel examples.

As the initial domain to investigate relation learning, Lu et al. (2012) examined learning of first-order comparative relations between animal concepts (e.g., a cow is larger than a sheep). Given feature vectors representing pairs of animals that exemplify a relation, BART acquires representations of comparative relations (e.g., *larger*, *smarter*) as weight distributions over the features. A key idea is that relation learning can be facilitated by incorporating *empirical priors*, which are derived using some simpler learning task that can serve as a precursor to the relation learning task (Silva, Heller, & Ghahramani, 2007). Just as children learn attributes of individual objects (e.g., *large*, *smart*) prior to binary relations (Smith, 1989), BART first learns representations of one-place categorical predicates (e.g., large animals, smart animals), which then serve as empirical priors to "jump-start" the acquisition of binary relations. For details on the operation of the model, see Lu et al. (2012).

BART's learned relations support generalization to new animal pairs. After receiving 100 training pairs represented using topic feature vectors, the model discriminated between novel pairs that instantiate a relation (e.g., *larger*, *smarter*) and those that do not with about 70%–80% accuracy. The model yields the classic symbolic distance effect (Moyer & Bayer, 1976), in which discrimination accuracy increases monotonically with the magnitude difference between items in a pair. Moreover, BART's learned weight distributions can be systematically transformed to evaluate analogies based on higher order relations between the learned first-order relations (e.g., *larger:smaller:: fiercer:meeker*). A simpler version of the model can predict magnitude values on specific dimensions based on human ratings for individual objects (Chen, Lu, & Holyoak, 2014).

1.3. The problem of generative inferences

Although the BART model shows some promise as a bottom-up model of relation learning, it nonetheless has many limitations. Perhaps most notably, it was developed as

a discriminative model, and thus can only perform tasks that involve evaluation of a relation (e.g., deciding whether the pair *sheep-fox* instantiates the relation *larger*) or evaluation of an analogy (e.g., deciding whether *larger:smaller:: fiercer:meeker* or *larger: smaller:: fiercer:slower* is the better choice to form a valid 4-term analogy). Unlike generative Bayesian models (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011), which naturally operate in a top-down manner to "fill in" incomplete representations, BART is unable to make inferences that require generation of new information that would complete a relation. For example, the model is unable to *generate* analogical completions (e.g., producing *meeker* to complete *larger:smaller:: fiercer:?*). The model is even unable to generate an answer to a simple factual question like "What is an animal larger than a fox?" Given the centrality of generative inferences in human cognition, as discussed above, the lack of such capacity is a severe limitation. Any model that aspires to account for human reasoning in a general way must specify mechanisms by which relations can be used to generate answers to inferential questions.

Several lines of work in artificial intelligence and cognitive science have examined the generation of exemplars and categories based on individual objects (for a recent review see Jern & Kemp, 2013). For example, Hinton and Salakhutdinov (2006) developed a multilayer neural-network model to reconstruct complex high-dimensional input, such as hand-written digits and grayscale patches in face images. Ward (1994) asked people to invent and draw imaginary animals from a different planet and observed that the novel constructions tended to be similar to familiar animals on Earth. Jern and Kemp (2013) had people study exemplars of an artificial category and then draw exemplars of the category. Their human data were broadly consistent with a model that learned the distribution of exemplars (represented by objects parts) for the category, and then generated exemplars by sampling from its learned distribution. In the area of category learning, there is evidence that instructions can guide learners to either focus on discrimination between exemplars of different categories, or on learning broader distributional properties of categories, with the latter learning style supporting a wider range of inferences (Hsu & Griffiths, 2010; Levering & Kurtz, 2015).

However, virtually all previous work relating learning to generative inferences has focused on categories and exemplars based on individual objects, rather than on tasks requiring generation of examples of instantiated *relations* between entities (the focus of the present paper). Although object concepts such as *dog* can be plausibly represented by distributions over features of objects (Fried & Holyoak, 1984), relation concepts such as *larger* are not directly definable in terms of features of individual objects (Doumas & Hummel, 2012). It therefore seems that generative inferences based on relations may involve different mechanisms than generative inferences based on object categories. There is ample evidence that people are in fact capable of making generative inferences in many relational tasks. For example, analogical inference involves generating inferences about a novel target situation by transferring knowledge from a more familiar source (e.g., Gick & Holyoak, 1980, 1983; Green, Fugelsang, Kraemer, Gray, & Dunbar, 2012). A number of computational models of inference generation by analogy have been proposed (e.g., Falkenhainer, Forbus, & Gentner, 1989; Halford, Wilson, & Phillips, 1998;

Holyoak, 2012; Hummel & Holyoak, 2003); however, with the important exception of Discovery of Relations by Analogy (DORA; Doumas et al., 2008), models of how analogical inferences can be generated have not directly addressed the issue of how relations are acquired in the first place. Notably, the DORA model (when applied in conjunction with *Learning and Inference with Schemas and Analogies* [LISA]; Hummel & Holyoak, 1997, 2003) is able to make generative relational inferences based on its acquired relations (Doumas, Morrison, & Richland, 2009).

This paper presents a model that begins to address the problem of making generative inferences based on relations that have been learned in a bottom-up manner from nonrelational inputs. We term the model *BART-g* (where "g" stands for "generative"), as it is an extension of the original BART model. Nonetheless, the new model is not strictly tied to BART, as its basic mechanism for making generative inferences could operate using outputs from any model that can take an ordered pair of objects and assign a probability that a specified relation holds for the pair. BART-g (like the DORA model; Doumas et al., 2008, 2009) thus provides an existence proof that a bottom-up model of relation learning has the potential to use its acquired representations to make basic generative inferences.

The rest of the paper is organized as follows: We first describe BART-g and the independently generated inputs that we use for relation learning. We then report simulations of the generative inferences required for two tasks: (a) transitive inference based on hypothetical objects; and (b) relation completion. The former task is one that is known to be within the capacity of preschool children. For relation completion, the performance of BART-g is compared to human data collected from adults performing the same task. Finally, we discuss both the promise and the limitations of the general approach to relation learning and generative inferences exemplified by BART-g.

2. Generative inferences in the BART-g model

2.1. Domain and inputs

We focus on the same domain used in the initial BART project (Lu et al., 2012): comparative relations between animal concepts. We also use the same basic inputs employed in previous work. To establish the "ground truth" of whether various pairs of animals instantiate different comparative relations, we used a set of human ratings of animals on four different continua (size, speed, fierceness, and intelligence; Holyoak & Mah, 1981). These ratings made it possible to test the model on learning eight different comparative relations: *larger*, *smaller*, *faster*, *slower*, *fiercer*, *meeker*, *smarter*, and *dumber*. Each animal concept is represented by a real-valued feature vector. In order to avoid the perils of hand-coded inputs (i.e., the possibility that the model's successes may be partly attributable to hidden representational assumptions made by the modelers), we use two sets of independently generated representations of objects as inputs, which we term "Leuven vectors" and "topic vectors," respectively.

2.1.1. Leuven vectors

We derived Leuven vectors from norms of the frequencies with which participants at the University of Leuven generated features characterizing 129 different animals (De Deyne et al., 2008; see Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). Each animal in the norms is associated with a set of frequencies across more than 750 features. We created vectors of length 50 based on the 50 features most highly associated with the subset of 44 animals that are also in the ratings dataset (Lu et al., 2012). Fig. 1 provides a visualization (for 30 example animals and the first 15 of the 50 features) of these high-dimensional and distributed representations, which might reflect the semantic representations underlying people's everyday knowledge of various animals.

2.1.2. Topic vectors

We created topic vectors by running the topic model (Griffiths et al., 2007) on a pre-processed version of the English Wikipedia corpus, which contained 174,792 entries and 116,128 unique words. We generated three Markov chains using the same corpus. The first sample in each chain was taken after 1,000 iterations, and sampling was repeated once every 100 iterations until eight samples were produced. Each sample yielded a matrix in which the (i, j)th entry is the number of times that word i has been assigned to topic j. From this matrix, we derived a vector for each word based on the conditional probability of each topic given that word. We averaged the word vectors created from different samples within a single Markov chain because they contained very similar topics (determined by examining the most probable words for each

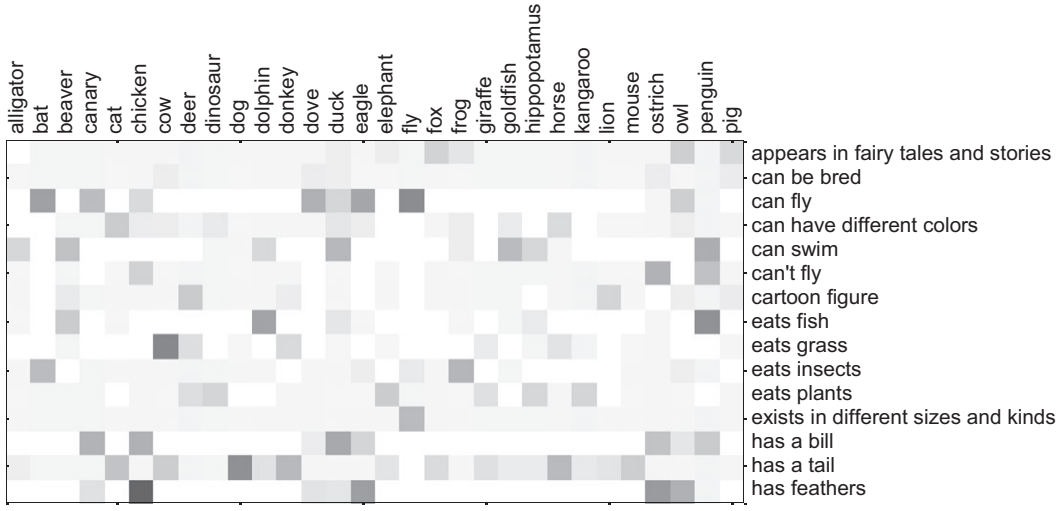

Fig. 1. Illustration of Leuven vectors for some example animals (De Deyne et al., 2008), reduced to 15 features to conserve space. The cell intensities represent feature values derived from response frequencies in a feature generation task (dark indicates high values and light indicates low values).

topic). However, samples from different Markov chains contained different topics, so they could be not be averaged. We merged the word representations created from the three Markov chains in the following way: First, for each chain, we selected the 30 features that had the highest values summed across the animals in the Holyoak and Mah (1981) norms. We then ran BART's relation-learning module using each chain's set of 30 features separately and ranked the chains in terms of the model's generalization performance (which did not differ very much across the different chains). We added all 30 features from the chain that resulted in the best performance to the final set of features. We then added features from the second-best chain that did not have very high correlations with any of the 30 features chosen so far (specifically, all correlations had to be < 0.8), which resulted in an additional 12 features. Finally, we added features from the last chain that did not have very high correlations with any of the 42 features chosen from the first two chains, resulting in a total of 52 features. Fig. 2 illustrates these topic vectors for the same 30 animals as in Fig. 1 using the first 15 of the 52 topic features, and Table 1 contains the top 10 words associated with each of these 15 topics.

2.2. Acquisition of relational representations in BART

BART-g is based on relations learned by the original BART model. BART represents a relation using a joint distribution of weights, $\mathbf{w}$, over object features. A relation is learned by estimating the posterior probability distribution $P(\mathbf{w}|\mathbf{X_S}, \mathbf{R_S})$, where $\mathbf{X_S}$ represents the feature vectors for object pairs in the training set, the subscript $\mathbf{S}$ indicates

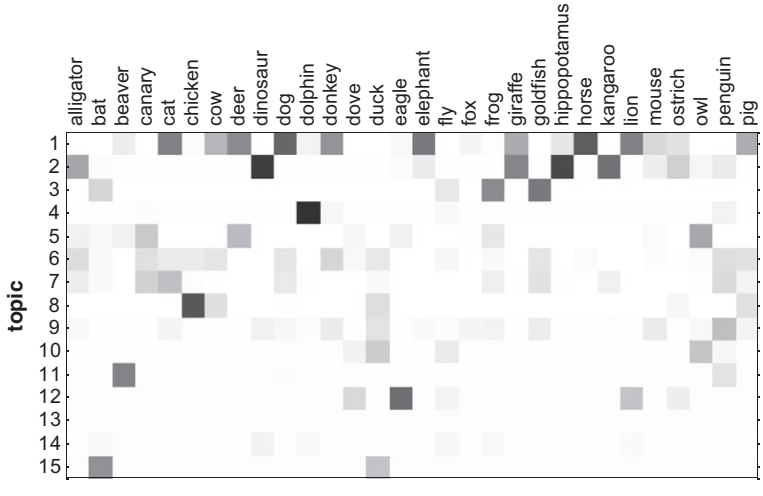

Fig. 2. Illustration of topic vectors (Griffiths et al., 2007), reduced to 15 features to conserve space, for the same example animals as in Fig. 1. The cell intensities represent feature values (dark indicates high values and light indicates low values). See Table 1 for the top 10 words associated with each topic.

Table 1
Top 10 words associated with example topics created using the Wikipedia corpus

Topic	Top 10 Words
1	Horse animals animal dog horses dogs wolf breed wild hunting
2	Found snake teeth years fossil large specimens genus modern largest
3	Species eggs found food body prey feeding egg feed insects
4	Island sea fish marine fishing islands coast water beach coastal
5	Forest species plant plants trees tree forests native areas habitat
6	Room back car door night find inside man front house
7	Character appears characters shown main named voiced revealed appearance series
8	Food made meat served called popular milk cooking cuisine dish
9	Disney animated Warner Walt cartoon voice animation featured television film
10	Species birds worldwide occur small bird family large long short
11	Oregon expedition North map Arctic Pacific Portland cook South ice
12	Cross mark flag arms symbol sign coat national official red
13	Human genetic evolution natural evolutionary humans biological selection life Darwin
14	Power battle form attack powerful ability fight sword fighting defeat
15	England runs match cricket Australia wickets made innings series scored

the set of training examples, and $\mathbf{R_S}$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular pair of objects instantiates the relation or not. The distribution of the vector $\mathbf{w}$ constitutes the relational representation, the mean of which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in $\mathbf{X}$ on judging whether the relation applies.

The weight distribution can be updated based on examples of ordered pairs that instantiate the relation in the training set. Formally, the posterior distribution of weights can be computed by applying Bayes' rule:

$$P(\mathbf{w}|\mathbf{X_S}, \mathbf{R_S}) = \frac{P(\mathbf{R_S}|\mathbf{w}, \mathbf{X_S})P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R_S}|\mathbf{w}, \mathbf{X_S})P(\mathbf{w})}, \tag{1}$$

where the likelihood term is defined using a logistic function:

$$P(R = 1|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w^T x}}}. \tag{2}$$

The prior distribution of weights is a multivariate normal distribution. As in Lu et al. (2012), an empirical prior was used for the means of the prior weight distribution. Specifically, we trained BART on the eight categories of one-place predicates (e.g., *large*, *small*, *fierce*, *meek*) that can be formed with the extreme animals at each end of the four magnitude continua (size, speed, fierceness, and intelligence). For example, we used the 20 largest animals (e.g., whale, dinosaur, elephant) to learn the category of large animals and the 20 smallest animals (e.g., flea, fly, worm) to learn the category of small animals. When the model is then presented with ordered pairs,

BART automatically forms an empirical prior for the first role of the two-place relation by identifying the one-place predicate that best distinguishes the objects in the two relational roles (i.e., the category of which the first object is maximally more likely than the second to be a member). The potential priors on the second role are linked to those for the first role, by reversing the sign on the weights for the first role to form a contrast. The reliability of prior selection will naturally vary with the number of training examples, yielding an inherent source of variability in the acquisition of the relations. In an alternative "baseline" version of the model, the empirical prior on weights is replaced by an uninformative prior (standard normal distributions).

After learning the joint weight distribution that represents a relation, BART discriminates between pairs that instantiate the relation and those that do not by calculating the probability that a target pair $\mathbf{x}_T$ instantiates the relation R:

$$P(R_T = 1|\mathbf{x}_T, \mathbf{X_S}, \mathbf{R_S}) = \int_{\mathbf{w}} P(R_T = 1|\mathbf{x}_T, \mathbf{w}) P(\mathbf{w}|\mathbf{X_S}, \mathbf{R_S}). \tag{3}$$

2.3. Generative inferences in BART-g

The goal of the present paper is to endow BART with the ability to make generative inferences, so that the extended model BART-g can, for example, complete a partially instantiated relation, answering questions such as "What is an animal that is smaller than a dog?"[1] We use the weight representation for a relation learned by BART to complete generative tasks. When presented with a cue relation (e.g., *smaller*) and a cue object (e.g., dog), the model produces possible responses for the remaining object (e.g., cat) so that the ordered object pair satisfies the relation. More specifically, given the features of a cue object B, $\mathbf{x}_B$, and the knowledge that relation R holds for the object pair (A, B), BART-g generates a probabilistic description for the feature vector of object A, $\mathbf{x}_A$, by making the following inference:

$$P(\mathbf{x}_A|\mathbf{x}_B, R = 1) \propto P(R = 1|\mathbf{x}_A, \mathbf{x}_B) P(\mathbf{x}_A|\mathbf{x}_B). \tag{4}$$

This generative inference reflects a compromise between (1) maximizing the semantic similarity between A and B, which is reflected in the prior term, $P(\mathbf{x}_A|\mathbf{x}_B)$, and (2) maximizing the probability that the relation holds between the two objects, which is reflected in the likelihood term, $P(R = 1|\mathbf{x}_A, \mathbf{x}_B)$.

The term $P(R = 1|\mathbf{x}_A, \mathbf{x}_B)$ is the probability that relation R holds for a particular hypothesized object A, $\mathbf{x}_A$, and the known object B, $\mathbf{x}_B$. It is defined using a logistic function (consistent with the likelihood function used in BART, Eq. (2)):

$$P(R = 1|\mathbf{x}_A, \mathbf{x}_B) = \frac{1}{1 + e^{-\mathbf{w}_1^{\mathbf{T}}\mathbf{x}_A - \mathbf{w}_2^{\mathbf{T}}\mathbf{x}_B}}. \tag{5}$$

Relative to Eq. (2), we have only introduced small differences in the notation. The learned relational weights, $\mathbf{w}$, are written as two separate halves: weights associated with the first relational role ($\mathbf{w}_1$) and weights associated with the second relational role ($\mathbf{w}_2$). Correspondingly, the feature vector × for a pair of objects is separated into the feature vector for object A ($\mathbf{x}_A$) and the feature vector for object B ($\mathbf{x}_B$).

The prior for the features of object A, $P(\mathbf{x}_A|\mathbf{x}_B)$, is the conditional probability distribution given the features of object B, defined as the following:

$$P(\mathbf{x}_A|\mathbf{x}_B) = N\left(\mathbf{x}_B, \sigma^2\mathbf{I}\right). \tag{6}$$

We assume that object B (the cue object) serves a starting point for generating object A, so the means of $P(\mathbf{x}_A|\mathbf{x}_B)$ are taken to be the feature values of object B, reflecting a certain degree of semantic dependency between the two objects. The prior also encodes the assumptions that the features of A are uncorrelated and have the same variance σ^2, the value of which is a free parameter that determines the strength of the model's bias for generating A objects that are similar to B.

To compute the inference in Eq. (3), we adapted the variational method (Jaakkola & Jordan, 2000) using the following updating rules for the mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ of the feature distribution, as well as the variational parameter ξ:

$$\begin{aligned} \mathbf{V}^{-1} &= \tfrac{\mathbf{I}}{\sigma^2} + 2\lambda(\xi)\mathbf{w}_1\mathbf{w}_1^{\mathbf{T}}, \\ \mu &= \mathbf{V}\left(\tfrac{\mathbf{I}}{\sigma^2}\mathbf{x}_B + \tfrac{\mathbf{w}_1}{2} - 2k\lambda(\xi)\mathbf{w}_1\right), \\ \xi^2 &= \mathbf{w}_1^{\mathbf{T}}(\mathbf{V} + \mu\mu^{\mathbf{T}})\mathbf{w}_1, \end{aligned} \tag{7}$$

where $\lambda(\xi) = \frac{\tanh\left(\frac{1}{2}(\xi+k)\right)}{4(\xi+k)}$ and $k = \mathbf{w}_2^{\mathbf{T}}\mathbf{x}_B$. The variational method is much more efficient than alternative sampling methods for making inferences based on high-dimensional representations.

Figs. 3 and 4 illustrate the operation of the model in generating an animal (object A) that is larger than a sheep (Fig. 3) or an elephant (Fig. 4), where the latter fill the role of object B. The feature distribution for A is updated from a prior, $P(\mathbf{x}_A|\mathbf{x}_B)$, favoring some degree of similarity between the two animals (left panel; top: high similarity, bottom: low similarity) to a posterior distribution, $P(\mathbf{x}_A|\mathbf{x}_B, R = 1)$, after taking into consideration the relation (i.e., *larger*) instantiated by the animal pairs (right panel). These distributions are shown in a simplified two-dimensional feature space (size and speed ratings for animals; Holyoak & Mah, 1981). In a later section, we apply the model to more complex and distributed feature representations.

3. Transitive inference based on hypothetical instances

The first test of BART-g evaluated whether the model enables transitive inferences on comparative relations. Comparative relations such as *larger* exhibit the logical properties

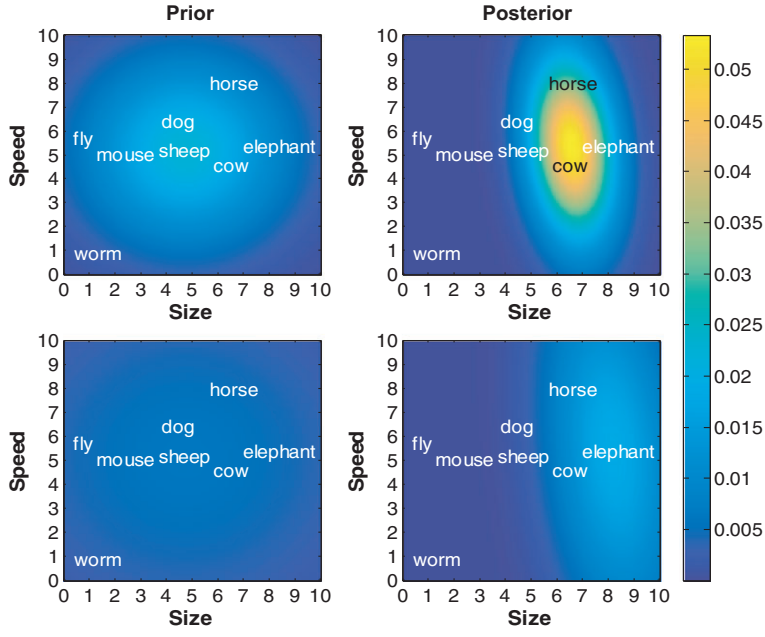

Fig. 3. Illustration of the model results for inferring an animal that is larger than a cued animal (*sheep*), in a simplified two-dimensional space (size and speed ratings for animals; Holyoak & Mah, 1981). Colors reflect probability densities (yellow indicates high values and blue indicates low values). The top panels show the prior and posterior distributions with $\sigma^2 = 7$ (favoring similarity-based completions such as *cow*), and the bottom panels show the prior and posterior with $\sigma^2 = 25$ (favoring landmark-based completions such as *elephant*). The locations of various animals in this two-dimensional space are labeled. The posterior distribution of inferred animals was generated using the relational weights that BART learned based on the four-dimensional feature vectors derived from the full set of human magnitude ratings. The feature space was reduced to two dimensions for the purpose of illustration.

of transitivity and asymmetry, supporting deductions such as "If *A* is larger than *B* and *B* is larger than *C*, then *A* is larger than *C*." Such hypothetical reasoning seems to depend on the ability to generate arbitrary instantiations of the relation without any guidance from object features (as the object representations are semantically empty). Note that the ability to make "one-shot" transitive inferences based on hypothetical instances is entirely distinct from the ability to learn an ordered series from repeated exposure to specific pairs. The latter phenomenon, termed "transitivity of choice," is within the capacity of many species (Merritt & Terrace, 2011). In contrast, one-shot transitive inference based on arbitrary instantiations has not been shown convincingly in any species other than humans (Halford, 1984). Human children reliably succeed on this task by about age 5 or 6 (Goswami, 1995; Halford, 1984; Kotovsky & Gentner, 1996).

3.1. Transitive inference in BART-g

The basic approach to transitive inference in BART-g is straightforward: The model "imagines" objects *A*, *B*, and *C* that instantiate the two given premises, as in the example

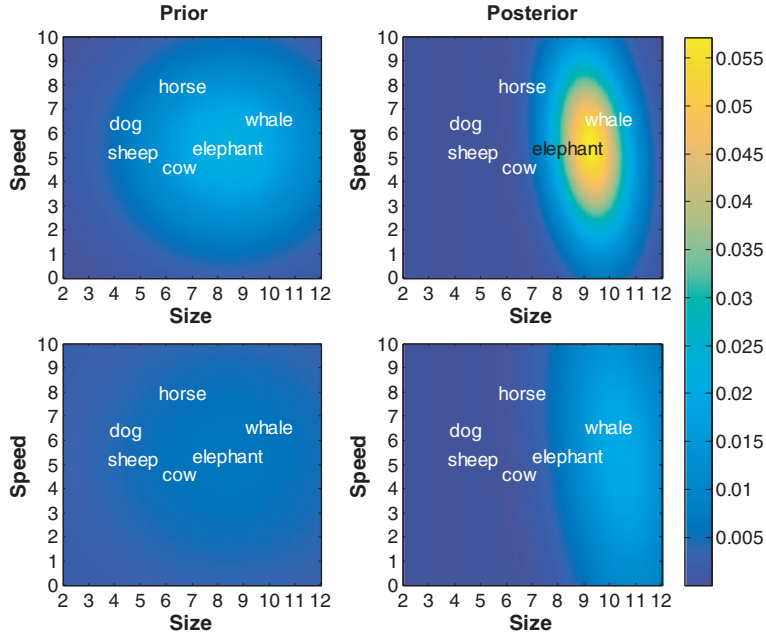

Fig. 4. Illustration of the model results for inferring an animal that is larger than an elephant in a simplified two-dimensional space. σ^2 is set to the same values as in Fig. 3 (seven for the top panels and 25 for the bottom panels).

above, and then tests the unstated relationship for the pair <*A*, *C*>. The model "imagines" an object by employing its generative inference computation, as described in the previous section. Specifically, for each of the eight comparative relations that BART learned, we first let BART-g imagine an animal *B* (because the statement "*A* is larger than *B*" implies that *B* is the referent against which *A* is being compared) by randomly sampling a feature vector from a distribution representing the animal category. This is an empirical Gaussian distribution with a mean vector and covariance matrix that were directly estimated from the feature vectors of the animals in the ratings dataset that had Leuven or topic vectors, respectively. There were 44 such animals for the Leuven inputs and 77 such animals for the topic inputs.

Given the sampled animal *B*, BART-g constructs a distribution for animal *A* (e.g., to satisfy the premise that "*A* is larger than *B*") by letting *B* fill the second role of the relevant relation. Similarly, the model constructs a distribution for animal *C* (e.g., to satisfy the premise that "*B* is larger than *C*") by letting *B* fill the first role of the same relation. Next, the model creates feature representations for specific animals *A* and *C* by setting their feature vectors, $\mathbf{x}_A$ and $\mathbf{x}_C$, to be the means of the inferred feature distributions for *A* and *C*, respectively. Note that these "imagined" animals are hypothetical: Although their features are sampled from the distribution of animal features, the results will seldom correspond to actual animals. To ensure that the premises have actually been satisfied, the model accepts the imagined animal *A* only if $P(R = 1|\mathbf{x}_A, \mathbf{x}_B) > 0.5$ and

$P(R = 1|\mathbf{x}_B, \mathbf{x}_A) < 0.5$, and the imagined animal C only if $P(R = 1|\mathbf{x}_B, \mathbf{x}_C) > 0.5$ and $P(R = 1|\mathbf{x}_C, \mathbf{x}_B) < 0.5$. If either animal A or animal C is rejected, a new animal B is sampled and A and C are generated again using the procedure just described.

Finally, if $\mathbf{x}_A$ and $\mathbf{x}_C$ have been accepted as satisfying the premises, the model calculates both $P(R = 1|\mathbf{x}_A, \mathbf{x}_C)$, denoting the probability that A is larger than C, and $P(R = 1|\mathbf{x}_C, \mathbf{x}_A)$, denoting the probability that C is larger than A. The model concludes that the relation holds for the pair $<A, C>$ if $P(R = 1|\mathbf{x}_A, \mathbf{x}_C) > 0.5$ and $P(R = 1|\mathbf{x}_C, \mathbf{x}_A) < 0.5$, implying that transitivity holds for the imagined *A-B-C* triad.

3.2. Evaluation of BART-g on transitive inference

We conducted tests of transitive inference with BART-g using the relational representations that BART learned based on 100 randomly chosen training pairs that instantiate a comparative relation, such as *larger* or *slower*. The training regime was essentially identical to that used by Lu et al. (2012), starting with an initial phase of learning simple attributes (e.g., *large* and *small*) that provided empirical priors for learning the corresponding two-place relations. If the *larger* relation that BART has learned is indeed transitive, then any instantiation of animal pairs for which "*A* is larger than *B*" and "*B* is larger than *C*" are both true will satisfy the conclusion, "*A* is larger than *C*." Thus, BART-g repeatedly imagines *A-B-C* triads that satisfy the premises and then draws the conclusion, in essence searching for a counterexample. If no counterexample is ever found, the transitive inference is accepted.

For comparison, we also tested a baseline model that substituted an uninformative prior for the empirical prior that guides BART's relation learning (for a full description, see Lu et al., 2012), but which is otherwise identical to BART-g. The learning process in the baseline model is essentially the same as standard logistic regression, and the model makes generative inferences using the same mechanism (with the same parameters) as does BART-g. For each of the eight comparative relations, the two relation-learning models were each run 10 times, each time with a different set of training pairs, resulting in a different learned weight distribution. For each of these learned weight distributions representing a comparative relation, we let the model generate 100 *A-B-C* triads satisfying the premises, testing the relevant relationship between A and C for each triad. To assess the influence of the free parameter in model predictions, the tests were conducted multiple times with different values of σ^2 ranging from 1 to 1,000 for the Leuven inputs and from 100 to 100,000 for the topic inputs.[2] The strongest tests are those in which σ^2 is set at low values, creating a strong prior preference that A, B, and C are similar to one another. When the similarity constraint is strong, the model is biased to generate animals that are similar to the cued animal, and hence more likely to yield a counterexample.

3.2.1. Results for Leuven inputs

Fig. 5 shows the mean proportion correct (i.e., the mean proportion of triads that satisfy the constraints based on transitive inference) for BART-g and the baseline model as a function of σ^2 (ranging from 1 to 1,000 for Leuven inputs). These results were

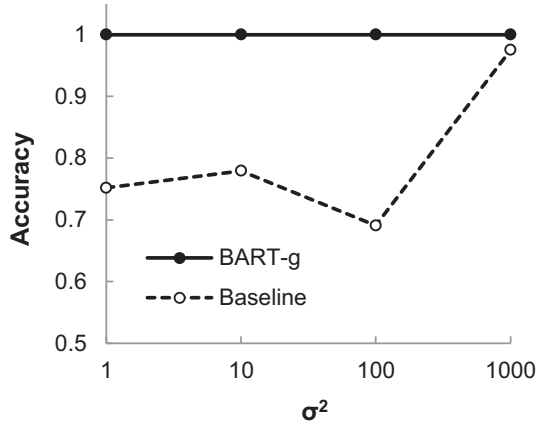

Fig. 5. Mean proportion correct on the transitive inference task for BART-g and the baseline model using Leuven vectors, as a function of the variance parameter. These results are averaged over 80 learned relational weight distributions.

averaged over all 80 learned relational distributions. The critical result is that BART-g's accuracy remained constant at 100% as σ^2 was reduced to the effective minimal value of 1. When the value of σ^2 was reduced below 1 for the Leuven inputs, the models produced many instantiations that failed to satisfy the required premises (i.e., $A > B$, $B > C$, and not vice versa), due to the very strong bias involved in generating highly similar objects. (Such failed instantiations of the premises were discarded and hence did not influence the results shown in Fig. 5.)

Thus, BART-g demonstrates what may be considered an inductive approximation to deduction for a wide range of parameter values for the prior: Despite exhaustive search for a counterexample to the transitive inference, no counterexample was ever found. In contrast, the baseline model often failed to satisfy the transitive premises to conclude that $A > C$ (and not vice versa) even when the value of σ^2 was as large as 100.

3.2.2. Results for topic inputs

Fig. 6 shows the models' performance on transitive inference for topic inputs as a function of σ^2. BART-g remained at 100% accuracy for a range of values of σ^2. In stark contrast, the baseline model often found only a small fraction of the desired 100 *A-B-C* triads that satisfied the premises. For values of σ^2 from 100 to 100,000, the baseline model, respectively, found on average 1.54, 22.7, 67.71, and 73.43 triads that satisfied the premises after generating 10,000 triads in total. The curve for the baseline model in Fig. 6 shows the mean proportion correct for whatever number of triads satisfied the premises, whereas the curve for BART-g plots the first 100 triads that BART-g found satisfying the premises. Hence, for topic inputs, although the baseline model finds many counterexamples to the transitive inference, BART-g demonstrates that the comparative relations it has learned are indeed transitive and asymmetric. This result demonstrates the necessity of learning adequate relational representations in order to achieve successful generative inference.

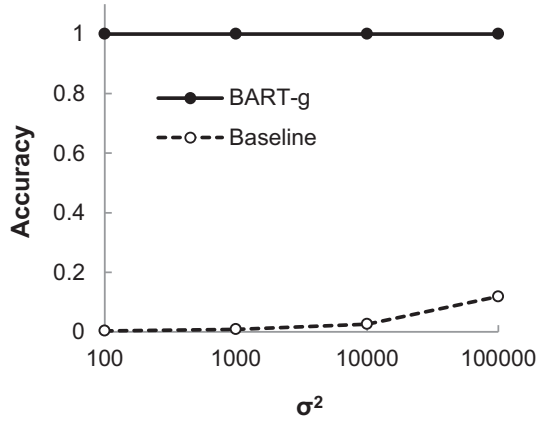

Fig. 6. Mean proportion correct on the transitive inference task for BART-g and the baseline model using topic vectors, as a function of the variance parameter. These results are averaged over 80 learned relational weight distributions.

4. Animal generation task

A second evaluation of the BART-g model involves answering generative questions such as "What is an animal that is smaller than a dog?" Although one might suppose that such questions could be answered by undirected trial and error, we shall see that people's answers are often systematically guided by their representations of the relation and of the animal provided as a cue. We conducted an experiment to characterize the pattern of human responses in an animal generation task, using Amazon Mechanical Turk. In this free-generation study, participants typed responses to queries of the form, "Name an animal that is larger than a dog." They were instructed to enter the first animal that came to mind. Four comparative relations (*larger*, *smaller*, *faster*, and *slower*) and nine cue animals (shark, ostrich, sheep, dog, fox, turkey, duck, dove, and sparrow) were used. At least 50 responses were collected for each of the 36 relation-animal pairs. To minimize learning across trials, we asked each participant to respond to only two queries about a single animal: either *larger* and then *slower*, *slower* and then *larger*, *faster* and then *smaller*, or *smaller* and then *faster*.

Participants were instructed to complete the study only if they were fluent in English. There were 1,147 participants, resulting in a total of 2,294 responses across the 36 queries. We processed the responses by removing articles such as "an," correcting obvious misspellings (e.g., "pidgeon"), and expanding abbreviations (e.g., "hippo"). We removed two of the responses ("dig" and "bow") because it was not clear what animals they were supposed to be.

The same 36 relation-animal cue pairs were presented to BART-g after it had been trained on the relevant relations using either Leuven or topic vectors. For each input representation (Leuven or topic), we used a set of animals having feature vectors in that representation to construct possible responses to the 36 queries. For the Leuven inputs, we

used the 129 animals included in the Leuven dataset. For the topic inputs, we used the set of 168 animals that participants provided as a response at least twice in the entire MTurk study. For each query, BART-g produced a posterior following a multivariate normal distribution for the feature vector of the missing animal using Eq. (4). This distribution was used to derive model predictions of the animal names that could serve as responses to each query in the following way. For each specific animal, we calculated the probability density of its feature vector under the posterior distribution for the feature vector of the missing animal, $P(\mathbf{x}_A|\mathbf{x}_B, R = 1)$. The probability densities calculated for all 129 (for the Leuven inputs) or 168 (for the topic inputs) animals were normalized to produce a discrete probability distribution. These discrete probabilities were then averaged across the 10 runs of the BART relation-learning model. Note that for both the Leuven and topic inputs, the set of animals for which we obtained model predictions included many animals outside the original training set given to the relation-learning model. In other words, the set of animals involved in the generation task included many new animals that had not been encountered by the BART model in the course of acquiring its relational representations.

4.1. Human results for the animal generation task

Table 2 shows examples of human responses, and the complete set of human responses is provided in the Table S1. The human responses appear to be mainly driven by two trends: (a) reporting an animal that is similar to the cue animal and that fits the cue relation (e.g., *cat* for "smaller than a dog"); and (b) reporting a "landmark" animal at an extreme of the continuum (e.g., *turtle* for "slower than a dog"). The landmark animal coupled with the cue animal provides an "ideal" example of the cue relation (i.e., one that maximizes the probability that the relation holds).

Interestingly, these two basic factors—feature similarity and proximity to an ideal—have both been shown to be important in various types of categorization tasks. In general, object categories are based on feature similarity (e.g., Davis, Xue, Love, Preston, & Poldrack, 2014; Rosch & Mervis, 1975), whereas more abstract or relational categories are often based on ideals (e.g., Barsalou, 1983; Goldstone, Steyvers, & Rogosky, 2003; Hampton, 1981). An animal generation question (e.g., "What is an animal larger than a dog?") is cued by a partially instantiated relation, in essence transforming a two-place relation, *larger(x, y)*, into a one-place predicate defining an ad hoc object category, *larger-than-dog(x)*. Accordingly, both the relational ideal and object-oriented feature similarity provide potential constraints to guide generation of answers to such relational queries. The relational ideal has the advantage of guaranteeing generation of a true relational statement. Although feature similarity provides a suboptimal basis for performing the relation generation task, it is highly relevant to many other semantic decisions about category members.

This tradeoff between reporting animals that are similar to the cue animal and reporting animals that are landmarks for the cue relation (and usually more dissimilar to the

Table 2
Examples of human responses in the animal generation task

Cue Relation	Cue Animal	n[a]	Response Proportions					
larger	Dog	53	Elephant	Horse	Cow	Bear	Lion	Other
			0.26	0.19	0.11	0.08	0.06	0.30
	Sparrow	58	Dog	Elephant	Eagle	Hawk	Bear	Other
			0.19	0.16	0.10	0.07	0.05	0.43
smaller	Dog	65	Cat	Mouse	Rat	Rabbit	Bird	Other
			0.31	0.22	0.17	0.06	0.05	0.19
	Sparrow	58	Mouse	Hummingbird	Ant	Worm	Goldfish	Other
			0.26	0.19	0.09	0.09	0.03	0.34
faster	Dog	65	Cheetah	Horse	Tiger	Cat	Leopard	Other
			0.69	0.08	0.05	0.03	0.03	0.12
	Sparrow	58	Cheetah	Eagle	Hawk	Bee	Lion	Other
			0.53	0.14	0.07	0.03	0.03	0.26
slower	Dog	53	Turtle	Snail	Cat	Pig	Elephant	Other
			0.49	0.08	0.06	0.06	0.04	0.27
	Sparrow	58	Turtle	Sloth	Snail	Dog	Ostrich	Other
			0.29	0.21	0.14	0.03	0.03	0.30

Note. The five most frequent responses are shown for each query. The total proportion of the other responses to each query is shown in the "other" column.
[a]The total number of responses for each query.

cue animal) is captured by the single free parameter in the generative module, σ^2. As explained earlier (see Fig. 3), a low σ^2 results in a response distribution that favors animals similar to the cue animal, whereas a high σ^2 leads to a preference for response animals that are more likely to satisfy the cue relation with respect to the cue animal (i.e., landmark animals for the cue relation).

Another pattern we observed in the human responses is that the responses to each query were often dominated by the most frequent response to that query. The average proportion of the most frequent response to each query relative to the total number of responses, across all 36 queries, was about 0.4. A typical pattern of human responses is displayed in Fig. 7, which shows the response frequencies and proportions (out of 53 total responses) to the query, "Name an animal that is slower than a dog." The most dominant response of *turtle* is followed by a long tail of low-frequency responses. It is difficult to explain exactly why some participants chose these low-frequency responses, especially *baby seal*, *seahorse*, or even *pig* (which was given as an answer by three different participants and tied with *cat* for third place). Fig. 8 shows the pattern of human responses to the query, "Name an animal that is smaller than a dog." Although the proportion of the most frequent response (*cat*) was lower for this query, the distribution once again contains a long tail of low-frequency responses. Therefore, we focused on the most frequent human response to each query when assessing model predictions, although we also report correlations between the model predictions and the entire "noisy" pattern of human responses.

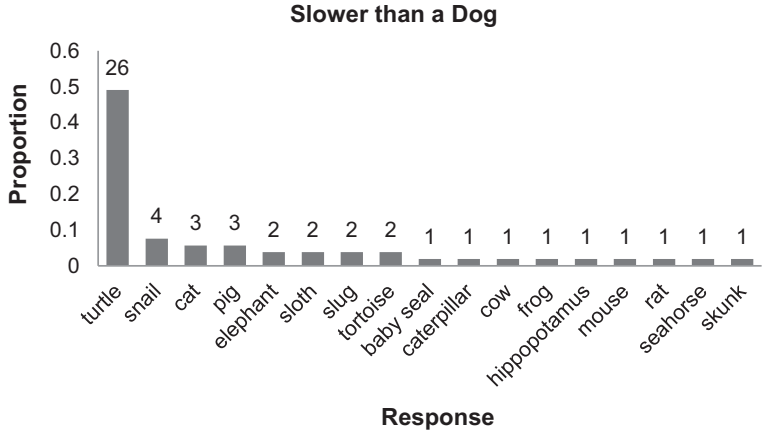

Fig. 7. A typical pattern of human responses in the animal generation task, showing response proportions and frequencies (shown above the bars) for the query, "Name an animal that is slower than a dog." Participants favored the "landmark" response to this query.

4.2. BART-g results for the animal generation task

We evaluated BART-g with respect to its predictions of the most frequent human response as well as the entire pattern of responses to each query. Specifically, we obtained three measures of model performance for each query: (a) the correlation (Pearson's r) between BART-g's predicted probabilities for the entire set of 129 (for the Leuven inputs) or 168 (for the topic inputs) animals and the proportion of participants who named each of these animals as a response; (b) whether BART-g actually gave the highest probability to the most frequent human response; and (c) the rank that BART-g gave to the most frequent human response among the entire set of animals for which we obtained model predictions. That is, we ranked the set of 129 or 168 animals in descending order of their predicted probabilities and examined the rank for the most frequent human response. A lower predicted rank indicates better model performance. To summarize model performance on all 36 queries, we calculated (a) the average correlation between predicted probabilities and observed response proportions; (b) the number of queries for which BART-g gave the highest probability to the top human response (the number of exactly correct predictions); and (c) the median of the ranks that BART-g assigned to the top human response across all queries. We chose the median so that a few outliers would not unduly affect the results (the results were very similar using means).

Participants in the animal generation task seemed to favor "landmark" responses for some of the four tested relations (especially *faster* and *slower*, and *larger* to a lesser extent), whereas they seemed to prefer responses based on similarity to the cue animal for other relations. Accordingly, BART-g's variance parameter was chosen separately for each relation in order to mimic the varied response strategies (landmark or similarity) that participants tended to use for different relations. For each of the four relations, we chose the variance parameter so as to maximize the number of queries for which

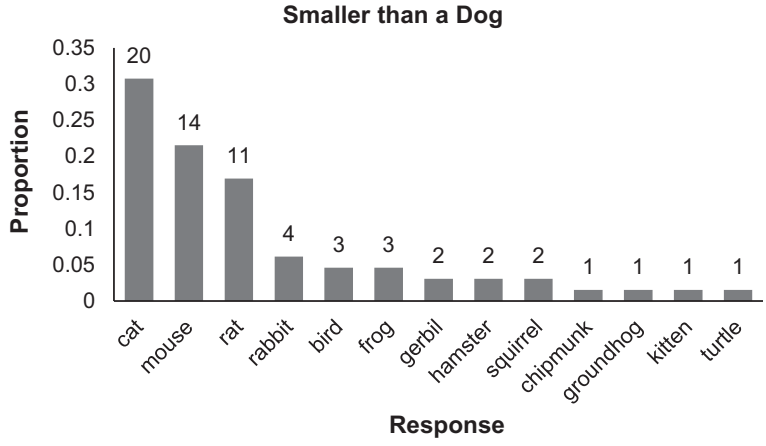

Fig. 8. The pattern of human responses for the query, "Name an animal that is smaller than a dog." Participants favored responses based on similarity to the cue animal for this query.

the model correctly predicted the top human response. When the number of correct predictions was the same for different values of the variance parameter, we broke the tie by choosing the variance that produced a lower median rank. After fitting the variance parameter for a relation, the same value was used for all cue animals for that relation.

We compared the BART-g model with two alternative, simpler models. The first alternative model simply used the prior term in Eq. (3), $P(\mathbf{x}_A|\mathbf{x}_B)$, and thus considered only the similarity of each possible response to the cue animal. The second alternative model made a decision based on the likelihood term, $P(R = 1|\mathbf{x}_A, \mathbf{x}_B)$, for each of the animals, and thus cared only about the probability that a possible response satisfies the cue relation with respect to the cue animal.

4.2.1. Results for Leuven inputs

BART-g's variance parameter was chosen from the values 1, 5, 10, 50, and 100. The best-performing variances were 50, 10, 10, and 100, respectively, for *larger*, *smaller*, *faster*, and *slower*. For *larger*, *smaller*, and *slower*, the chosen variances reflect the general patterns of "landmark" versus "similarity" responses for these relations. The relatively small value of 10 for *faster* is due to the fact that the Leuven dataset does not include *cheetah*, the landmark animal for the *faster* relation and the most popular human response to all of the *faster* queries. Accordingly, for the purpose of evaluating the models, given the Leuven inputs, the second most frequent human response was considered to be the dominant response. For many of the queries, the second most frequent human response was more similar to the cue animal, resulting in a smaller chosen variance.

4.2.1.1. Number of correct predictions: BART-g correctly predicted the top human response for 13 of the 36 queries (with the caveat noted above concerning the absence of *cheetah* in the Leuven set), which is impressive considering that there were 129 animals

from which to choose for each query. In fact, the probability of correctly predicting the top response for at least 13 of the 36 queries by choosing uniformly at random from the 129 animals is only 7.14×10^{-19}.

We can further evaluate BART-g's performance by comparing its behavior to that of a typical participant in our study. Although each participant answered only two of the 36 possible queries, we can estimate the total number of queries for which we would expect the average participant to provide the most frequent response. As mentioned previously, the average proportion of the most frequent response to each query across all 36 queries was about 0.4. Therefore, we would expect a typical participant to provide the top response for $36 \times 0.4 = 14.4$ queries. In comparison, BART-g generated the top response for 13 of the 36 queries. The model thus agreed with the dominant response of the entire set of human participants about as often as would be expected for a typical individual participant.

In contrast, the alternative model that uses only the prior term (the "prior" model) correctly predicted the top response for only one of the 36 queries ("smaller than a dog," to which the top response was *cat*), and the likelihood model made only four correct predictions (for one *faster* query and three *slower* queries). The probabilities of getting at least one correct and at least four correct by random chance are about 0.24 and 1.74×10^{-4}, respectively.

BART-g correctly predicted the top response for two *larger* queries, one *smaller* query, one *faster* query, and all nine *slower* queries. Note that predicting *turtle* as the top human response to all nine *slower* queries required an impressive feat of generalization on the model's part, because *turtle* was not in the original training set given to the BART relation-learning model.

4.2.1.2. Median ranks: Across all 36 queries, the median of the ranks that BART-g assigned to the top human responses was 8.5. In comparison, the median rank was 71.5 for the prior model and 11.5 for the likelihood model. Fig. 9 shows the breakdown of these results for the four comparative relations, with the median ranks displayed above the bars. For easier comparison with the topic inputs, in which the models considered a different total number of animals, the *y*-axis shows the median rank as a fraction of the total number of animals considered (129 in this case). Note that a lower median rank fraction indicates a better fit between model predictions and human responses. As shown in Fig. 9, the prior model performed poorly for all four relations, and the likelihood model tended to perform slightly worse than the BART-g model.

4.2.1.3. Correlations: Our analyses focused on the dominant human responses, which were relatively stable. However, we also applied the models to the complete set of human responses. Across all 36 queries, the average correlation (Pearson's r) between predicted probabilities and observed response proportions for the entire set of 129 Leuven animals was 0.31 for BART-g, 0.04 for the prior model, and 0.19 for the likelihood model. The response frequencies for human responses following the dominant response were very low (see Figs. 7 and 8), so it is not surprising that correlations for the entire set of human responses were low for all models. However, BART-g outperformed the alternative

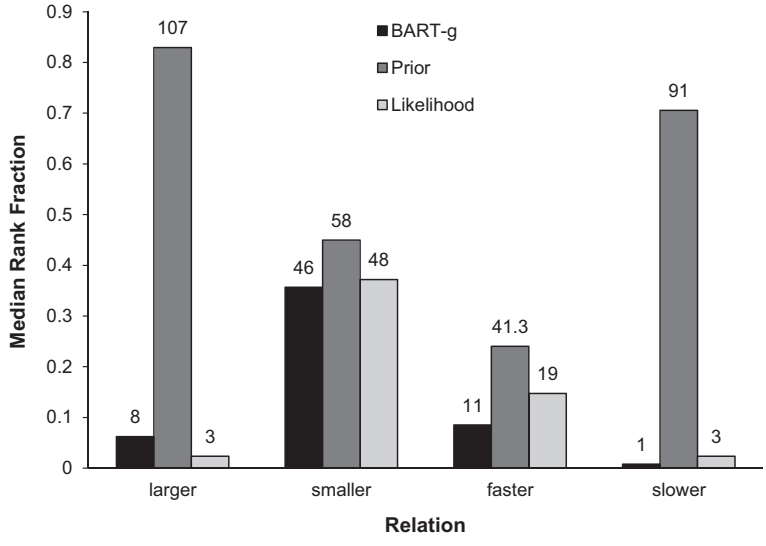

Fig. 9. Median ranks for the top human responses assigned by the different models using Leuven inputs, broken down by relation. The *y*-axis shows the median rank as a fraction of the total number of animals considered by the models. The actual median ranks (out of 129 animals) are shown above the bars. Note that lower values indicate better performance.

models not only across all 36 queries but also for each of the four relations, as shown in Fig. 10. Note that the variance parameter was not specifically chosen to maximize this correlation measure, but rather to maximize the number of queries for which the model correctly predicted the top human response. The pattern of results for the complete item set is similar to that obtained using the median ranks of the top human responses. Overall, these results indicate that the BART-g model, which considers both feature similarity to the cue animal and the likelihood of satisfying the cue relation with respect to the cue animal, predicts the pattern of human responses more accurately than models that consider only one of these factors.

4.2.2. Results for topic inputs

BART-g's performance using topic inputs was compared with two alternative models. One of these was the likelihood model that we tested for the Leuven inputs. The second alternative model, analogous to the prior model in the previous section, calculated a probabilistic quantity that represents word association strength in the topic model (Griffiths et al., 2007, p. 221):

$$P(w_2|w_1) = \sum_z P(w_2|z)P(z|w_1). \tag{8}$$

For a given cue animal word, w_1, we calculated $P(w_2|w_1)$ for each of the 168 possible response animal words (w_2) using all 300 topic dimensions (z) obtained from the topic

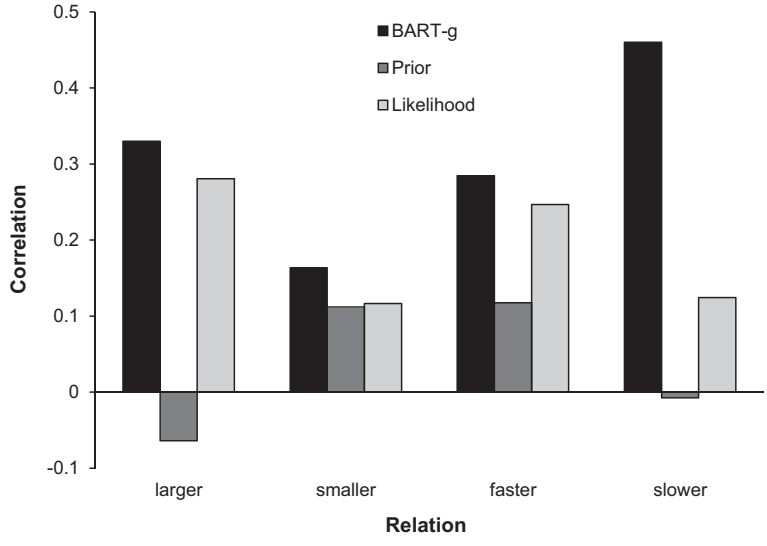

Fig. 10. Correlations (Pearson's *r*) between the models' predicted probabilities and observed response proportions for all 129 animals using Leuven inputs, broken down by relation.

model. This method yielded a predicted probability for each possible response animal corresponding to the semantic association strength from the cue animal word to the response animal word, which can be (but is not always) based on feature similarity between the two animals.

For the topic inputs, the variance parameter for the generative model was chosen from the values 100, 500, 1,000, 5,000, and 10,000. The best-fitting variances selected were 10,000, 1,000, 10,000, and 500, respectively, for *larger*, *smaller*, *faster*, and *slower*. These variance values are reasonable for *larger*, *smaller*, and *faster* given their respective response patterns. As we will see, BART-g performed the worst on the *slower* queries, though still better than both of the alternative models.

4.2.2.1. Number of correct predictions: BART-g correctly predicted the top human response for 15 of the 36 queries using the topic inputs. The probability of making at least 15 correct predictions by random chance when there are 168 animals from which to choose for each query is about 2.07×10^{-24}. In contrast, both the likelihood model and the model based on word association correctly predicted the top human response for only one of the 36 queries (one *faster* query and one *smaller* query, respectively), for which the corresponding chance probability is about 0.19. BART-g correctly predicted the top response for two *larger* queries, three *smaller* queries, all nine *faster* queries, and one *slower* query. Of particular note, predicting that *cheetah* would be the top human response to all nine *faster* queries required the model to generalize beyond the set of animals it encountered when learning the comparative relations (as *cheetah* was never used in the training pairs).

4.2.2.2. Median ranks: The median predicted rank for the top human response across all 36 queries was 7 for the BART-g model, 24 for the model based on word association, and 29 for the likelihood model. Fig. 11 shows the breakdown of these results for the four relations. BART-g outperformed the two alternative models for all four relations.

4.2.2.3. Correlations: Across all 36 queries, the average correlation (Pearson's *r*) between predicted probabilities and observed response proportions for the entire set of 168 animals was 0.34 for BART-g, 0.12 for word association, and 0.17 for the likelihood model. As shown in Fig. 12, BART-g outperformed both alternative models on *smaller* and *faster*, and the word association model on *slower*. Overall, these results indicate that BART-g accounts for the human data more successfully than either simple word association or consideration of the relation alone. Table 3 summarizes all the model results on the animal generation task for both Leuven and topic inputs.

5. General discussion

5.1. Generative inferences from a bottom-up model of relation learning

The present findings provide evidence that a bottom-up model of relation learning, designed to make discriminations between positive and negative examples of relations (Lu et al., 2012), can be extended to yield generative inferences. These inferences can involve

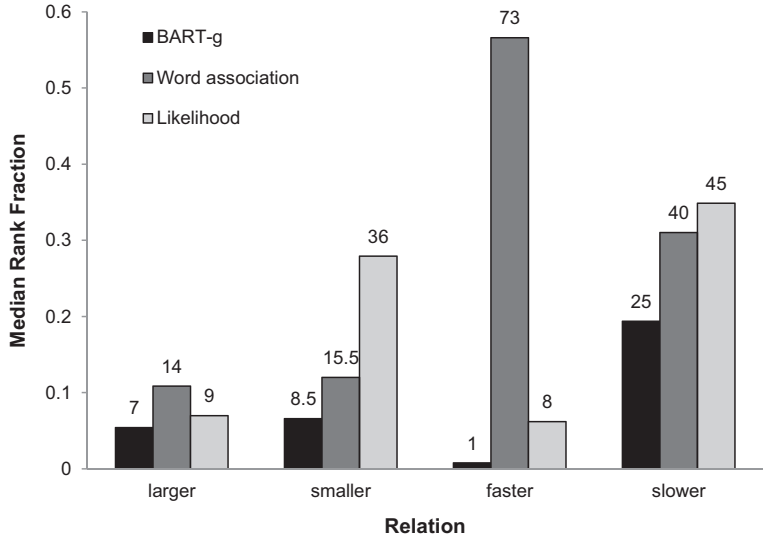

Fig. 11. Median ranks for the top human responses assigned by the different models using topic inputs, broken down by relation. The *y*-axis shows the median rank as a fraction of the total number of animals considered by the models. The actual median ranks (out of 168 animals) are shown above the bars. Note that lower values indicate better performance.

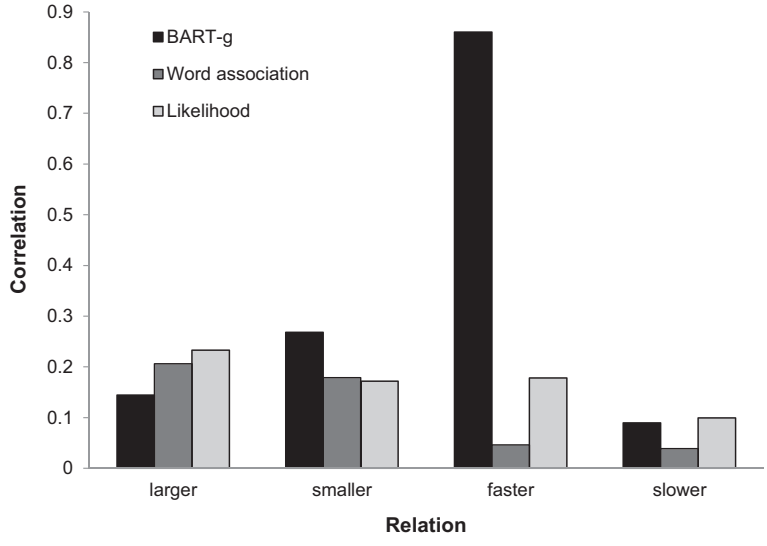

Fig. 12. Correlations (Pearson's *r*) between the models' predicted probabilities and observed response proportions for all 168 animals using topic inputs, broken down by relation.

relations between either hypothetical (in the case of transitive inference) or actual (in the case of animal generation) objects. The BART-g model thus constitutes an existing proof: It is possible to learn relations from non-relational (and independently generated) inputs in a bottom-up manner and to use the resulting relational representations not only to evaluate whether a stated relation does or does not hold, but also to generate novel instantiations of the relation. This is an important demonstration, as the capacity to make generative inferences using relations provides a key prerequisite for abstract relational thought.

The model's ability to make transitive inferences based on relations it has learned from examples in a bottom-up fashion illustrates the potential power of the discriminative approach to relation learning. Note that BART-g is not endowed with any notion of what a "transitive and asymmetric" relation is (though like a 6-year-old child, it *is* endowed with sufficient working memory to integrate two relations as premises). Rather, it simply uses its learned comparative relations to imagine possible object triads, and without exception concludes that the inference warranted by transitivity holds in each such triad. The model thus approximates "logical" reasoning by systematically searching for counterexamples (and failing to find any), akin to a basic mechanism postulated by the theory of mental models (Johnson-Laird, 2008; see also Holyoak & Glass, 1975). The fact that BART-g achieves error-free performance in the tests of transitive inference is especially impressive given that its inductively acquired relational representations are most certainly fallible (e.g., the BART model makes errors in judging which of two animals close in size is the larger; see Lu et al., 2012). It turns out that imperfect representations of comparative relations, acquired by bottom-up induction, can be sufficiently robust as to yield reliable quasi-deductive transitive inferences.

Table 3
Summary of model results on the animal generation task

		Leuven Inputs			Topic Inputs		
		BART-g	Prior	Likelihood	BART-g	Word Association	Likelihood
Number correct	Overall	13	1	4	15	1	1
	larger	2	0	0	2	0	0
	smaller	1	1	0	3	1	0
	faster	1	0	1	9	0	1
	slower	9	0	3	1	0	0
Median rank (and fraction)	Overall	8.5 (0.07)	71.5 (0.55)	11.5 (0.09)	7 (0.04)	24 (0.14)	29 (0.17)
	larger	8 (0.06)	107 (0.83)	3 (0.02)	7 (0.05)	14 (0.11)	9 (0.07)
	smaller	46 (0.36)	58 (0.45)	48 (0.37)	8.5 (0.07)	15.5 (0.12)	36 (0.28)
	faster	11 (0.09)	31 (0.24)	19 (0.15)	1 (0.01)	73 (0.57)	8 (0.06)
	slower	1 (0.01)	91 (0.71)	3 (0.02)	25 (0.19)	40 (0.31)	45 (0.35)
Correlation (Pearson's *r*)	Overall	0.31	0.04	0.19	0.34	0.12	0.17
	larger	0.33	−0.06	0.28	0.14	0.21	0.23
	smaller	0.16	0.11	0.12	0.27	0.18	0.17
	faster	0.28	0.12	0.25	0.86	0.05	0.18
	slower	0.46	−0.01	0.12	0.09	0.04	0.10

BART-g's capacity to make transitive inferences with specific relations, such as *larger*, may provide a first step toward acquiring a more general capacity to make such inferences. Many theorists have suggested that children and adults make transitive inferences by mapping stated premises onto a one-dimensional ordered mental array (e.g., Halford, 1993; Hummel & Holyoak, 2001; Huttenlocher, 1968). Halford (1993) hypothesized that such a mental array may originate from some basic perceptual ordering (e.g., for objects varying in size) that is learned at an early age, to which other relations may be mapped. Although BART-g does not create a mental array, its outputs could be adapted to generate small sets of items ordered by their magnitudes on a dimension (much like the simpler BARTlet model described by Chen et al., 2014), which could be mapped onto an array. Importantly, whether or not a novel relation is transitive must always be established empirically. For a brief period of time, young children tend to overgeneralize transitivity to non-transitive relations (e.g., if told that a boy loves a girl and the girl loves a dog, a child may infer that the boy must love the dog; Kuczaj & Donaldson, 1982). The implicit test of transitivity embodied in BART-g could be used to assess whether or not the inferences generated by mapping to an ordered array actually hold for a novel relation, thereby helping to correct relational overgeneralization.

In the animal generation task, BART-g achieved moderate success in modeling human response patterns by attempting to maximize both similarity to the cue animal and the probability that the cue relation is satisfied, performing better than models that consider just one of these factors alone. Human answers to questions that require relational completions (e.g., "What is an animal smaller than a dog?") are neither random, nor solely guided by word associations, nor solely guided by the likelihood of satisfying the

relation. Rather, people (and BART-g) appear to integrate similarity information (cf. Ward, 1994) with likelihoods of relations to generate constrained and systematic answers to what might appear to be open-ended questions.

Although addressed to very different tasks, there are interesting connections between BART-g and the PFC/BG working memory model (PBWM) developed by Kriete, Noelle, Cohen, and O'Reilly (2013). Both models are based on statistical learning and demonstrate that statistical learning can achieve reasonable performance for certain types of relations and some relatively simple forms of generative tasks. However, both modeling projects acknowledge that this type of computational mechanism only provides a first step in learning to bootstrap acquisition of relation representations for more complex relations and a broader range of generative inferences. The architecture of PBWM exemplifies a possible direction for the future development of BART, with the aim of providing the capability to discover role components that enable the formation of a hierarchical structure to connect object features and relational representations. For generative tasks, BART-g is limited by its lack of an explicit role representation, which would be a critical addition in order to move from comparative relations to other semantic relations. Nonetheless, BART-g benefits from the semantic richness of inputs derived from the topic model. Unlike the PBWM model, BART-g does not need to be trained with all possible words used in the test phase.

5.2. Limitations and future directions

BART-g is able to generate completions that form true comparative relations (e.g., generating *dog* as an animal larger than *cat*), and it can make one-shot transitive inferences about hypothetical instances. An apparent limitation, however, is that the model would not be able to make transitive inferences that are counterfactual in nature. For example, suppose that after learning animal sizes in the usual way, the model were asked to assume *cat is larger than dog* and *mouse is larger than cat* as premises. These premises are clearly false, and the model would have no way to treat them as true; hence, it would be unable to satisfy the prerequisite for assessing the transitive inference *mouse is larger than dog.* Rather, the model would simply determine that the putative conclusion (like the stated premises) is false. To the best of our knowledge, human performance with such counterfactual inferences based on comparatives has never been examined experimentally. However, work on syllogistic reasoning has shown that adults have some ability to reason with counterfactual premises, although they have difficulty overcoming belief biases (e.g., deciding that a conclusion known to be false nonetheless constitutes a valid inference; Evans, Barston, & Pollard, 1983). In general, people have difficulty making deductive inferences when they are unable to form an integrated mental model of the premises (Oakhill, Johnson-Laird, & Garnham, 1989); BART-g is entirely stymied in such cases. Perhaps the model could be augmented with inhibitory mechanisms that could suppress prior knowledge when it is necessary to reason counterfactually.

Although the extension of bottom-up relation learning to enable basic generative inference is an important theoretical advance, the more general project of modeling human inference abilities based on relational learning is still in its infancy. An obvious limitation of

BART-g is that it has only been tested in the very limited domains of comparative relations defined over a set of animals. As a step toward overcoming this limitation to comparatives, Eq. (5) could be extended in two ways. First, we could introduce a nonlinear kernel, $K(x_A, x_B)$, to map feature values to a higher dimensional space before applying the logistic function in this derived space. This method has been widely used in machine learning applications to capture nonlinear patterns in the data. Second, the logistic function in the equation could be replaced by a more sophisticated generative model to bind the features in the two related objects (cf. Jern & Kemp, 2013; Tenenbaum et al., 2011).

More generally, an important direction for future research is to extend models of learning and inference to a broader class of relations (e.g., Roy, Kemp, Mansinghka, & Tenenbaum, 2007). Comparative relations between generic types of entities can be defined by information that is intrinsic to the objects being compared. A number of important abstract relations have this property (e.g., superordinate, antonym, synonym). However, most relational predicates involve additional information *extrinsic* to the objects being related (e.g., spatial relations, action verbs, causatives). In particular, a great deal of modeling work has investigated learning and inference with causal relations (e.g., Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Lu, Rojas, Beckers, & Yuille, 2015; for a review, see Holyoak & Cheng, 2011), and some progress has been made in integrating causal learning with analogical inference (Holyoak & Lee, in press; Holyoak, Lee, & Lu, 2010). However, it remains unclear how models of causal learning relate to models of relation learning in general. More generally, it is essential to develop models of relation learning that operate on richer input representations, rather than solely using feature vectors for individual objects.

Although the BART model takes a discriminative (bottom-up) approach to relation learning, it may be possible to integrate discriminative models with more top-down generative models. As Jern and Kemp (2013) have argued, discriminative and generative models each seem especially suitable for particular types of tasks and stimuli. Jern and Kemp pointed out that any discriminative model can be augmented by an algorithm for generating instances, but that the most obvious such algorithm (based on random sampling from the full feature space) is prohibitively inefficient. The algorithm for instance generation incorporated into BART-g does not involve such random sampling. Rather, it is based on distributional assumptions about the features of types of objects, in which respect it is similar in spirit to the sampling approach incorporated into a generative model by Jern and Kemp (in which samples of exemplars are drawn from category distributions). However, rather than implementing a sampling algorithm, the BART-g algorithm is implemented using a variational method, allowing direct computation of generative inferences. BART-g thus illustrates how a discriminative model of relation learning can potentially merge with the generative approach to relational inference. More generally, relational knowledge initially acquired by a discriminative model can potentially provide a pool of relational schemas, which can in turn be used in a top-down fashion to guide further learning.

In order to account for human-level analogical reasoning, a model must be able to make generative inferences based on the integration of multiple relations (Halford, Wilson, Andrews, & Phillips, 2014; Halford et al., 1998; Waltz et al., 1999). Many current analogy models (for a review, see Gentner & Forbus, 2011) are able to reason with

complex systems of relations, and hence can account for basic phenomena of human analogical reasoning that remain beyond the reach of BART-g. However, none of these models have been tested on their ability to reason with relations that were acquired from independently generated, non-relational inputs. We believe it is important to establish that a computational path exists from non-relational inputs of realistic complexity to the acquisition of explicit relations, and beyond that, to the adult human capacity to reason generatively on the basis of complex relational representations.

Acknowledgments

We thank Airom Bleicher for helping us to conduct the animal generation study on Amazon Mechanical Turk, Charles Kemp for sharing Leuven inputs, Mark Steyvers for making the topic model code available, and Peter Gordon for providing us with a pre-processed version of the Wikipedia corpus. Preparation of the paper was supported by grant BCS-135331 from the National Science Foundation and grant N000140810186 from the Office of Naval Research.

Notes

1. Throughout this paper, we use the term "generative" to refer to inferences that require partial construction of a proposition (e.g., supplying an answer to a question such as, "What is an animal larger than a dog?"). In this usage, generative inferences contrast with *discriminative* judgments that involve evaluation of a fully stated proposition (e.g., "Is a bear larger than a dog?"). In the cognitive science literature, the terms "generative" and "generativity" are often used in a broader sense (e.g., in connection with the apparent systematicity of language and thought). We do not claim that "generative" inferences of the sort on which we focus here necessarily exhibit generativity in the broader sense.
2. We used different ranges of σ^2 for the Leuven and topic inputs because they are scaled differently. Across the animals in the ratings dataset, the mean variance among the 50 Leuven features is 0.79, whereas the mean variance among the 52 topics features is 147.24.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, *4*, 185–211.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27–54.

Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, *34*, 7472–7484.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030–1048.

Doumas, L. A. A., & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 52–66). New York: Oxford University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1–43.

Doumas, L. A. A., Morrison, R. G., & Richland, L. E. (2009). The development of analogy: Working memory in relational learning and mapping. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3133–3138). Austin, TX: Cognitive Science Society.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*, 295–306.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 234–257.

Gentner, D. (1977). If a tree had a knee, where would it be? Children's performance on simple spatial metaphors. *Papers and Reports on Child Language Development*, *13*, 157–164.

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science*, *2*, 266–276.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). Cambridge, UK: Cambridge University Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Glass, A. L., Holyoak, K. J., & Kossan, N. E. (1977). Children's ability to detect semantic contradictions. *Child Development*, *48*, 279–283.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition*, *31*, 169–180.

Goswami, U. (1995). Transitive relational mappings in 3-and 4-year-olds: The analogy of Goldilocks and the Three Bears. *Child Development*, *66*, 877–892.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *38*, 264–272.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Halford, G. S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology*, *16*, 65–93.

Halford, G. S. (1992). Analogical reasoning and conceptual complexity in cognitive development. *Human Development*, *35*, 193–217.

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.

Halford, G. S., Wilson, W. H., Andrews, G., & Phillips, S. (2014). *Categorizing cognition: Conceptual coherence in the foundations of psychology*. Cambridge, MA: MIT Press.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–831.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*, 497–505.

Hampton, J. A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition*, *9*, 149–156.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*, 504–507.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York: Oxford University Press.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Holyoak, K. J., & Glass, A. L. (1975). The role of contradictions and counterexamples in the rejection of false sentences. *Journal of Verbal Learning and Verbal Behavior*, *4*, 215–239.

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development*, *55*, 2042–2055.

Holyoak, K. J., & Lee, H. S. (in press). Inferring causal relations by analogy. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning*. New York: Oxford University 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.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197–204.

Hsu, A. S., & Griffiths, T. L. (2010). Effects of generative and discriminative learning on use of category variability. In R. Catrambone & S. Ohllson (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 242–247). Austin, TX: Cognitive Science Society.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. In M. L. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 279–305). Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, *75*, 550–560.

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian logistic regression: A variational approach. *Statistics and Computing*, *10*, 25–37.

Jern, A., & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, *66*, 85–125.

Johnson-Laird, P. N. (2008). Mental models and deductive reasoning. In L. Rips & J. Adler (Eds.), *Reasoning: Studies in human inference and its foundations* (pp. 206–222). Cambridge, UK: Cambridge University Press.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797–2822.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences, USA*, *110*(41), 16390–16395.

Kuczaj, S. A., & Donaldson, S. A. (1982). If the boy loves the girl and the girl loves the dog, does the boy love the dog? The overgeneralization of verbal transitive inference skills. *Journal of Psycholinguistic Research*, *11*, 197–206.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*, 266–282.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617–648.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2015). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science*, *39*, 1–36. doi:10.1111/cogs.12236

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–982.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587.

Merritt, D. J., & Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (*Homo sapiens*) and rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology*, *125*, 227–238.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, *8*, 228–246.

Oakhill, J. V., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117–140.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109–130; discussion 130–178.

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

Roy, D. M., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2007). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems*, *19*, 1185–1192.

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

Silva, R., Heller, K., & Ghahramani, Z.. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (pp. 500–507). Cambridge, MA: Journal of Machine Learning Research.

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 147–177). Cambridge, UK: Cambridge University Press.

Smith, L. B., Gasser, M., & Sandhofer, C. M. (1997). Learning to talk about the properties of objects: A network model of the development of dimensions. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *Advances in the psychology of learning and motivation*. Vol. 36 (pp. 219–255): *Perceptual learning*. San Diego, CA: Academic Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*, 1–40.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1. Human responses in the animal generation task.