Psychological Review
2012, Vol. 119, No. 3, 617–648

0033-295X/12/$12.00 DOI: 10.1037/a0028719

Bayesian Analogy With Relational Transformations

Hongjing Lu, Dawn Chen, and Keith J. Holyoak
University of California, Los Angeles

How can humans acquire relational representations that enable analogical inference and other forms of high-level reasoning? Using comparative relations as a model domain, we explore the possibility that bottom-up learning mechanisms applied to objects coded as feature vectors can yield representations of relations sufficient to solve analogy problems. We introduce Bayesian analogy with relational transformations (BART) and apply the model to the task of learning first-order comparative relations (e.g., *larger*, *smaller*, *fiercer*, *meeker*) from a set of animal pairs. Inputs are coded by vectors of continuous-valued features, based either on human magnitude ratings, normed feature ratings (De Deyne et al., 2008), or outputs of the topics model (Griffiths, Steyvers, & Tenenbaum, 2007). Bootstrapping from empirical priors, the model is able to induce first-order relations represented as probabilistic weight distributions, even when given positive examples only. These learned representations allow classification of novel instantiations of the relations and yield a symbolic distance effect of the sort obtained with both humans and other primates. BART then transforms its learned weight distributions by *importance-guided mapping*, thereby placing distinct dimensions into correspondence. These transformed representations allow BART to reliably solve 4-term analogies (e.g., *larger*:*smaller*::*fiercer*:*meeker*), a type of reasoning that is arguably specific to humans. Our results provide a proof-of-concept that structured analogies can be solved with representations induced from unstructured feature vectors by mechanisms that operate in a largely bottom-up fashion. We discuss potential implications for algorithmic and neural models of relational thinking, as well as for the evolution of abstract thought.

Keywords: analogy, relation learning, generalization, Bayesian models

Supplemental materials: http://dx.doi.org/10.1037/a0028719.supp

One of the hallmarks of human reasoning is the ability to form representations of relations between entities and then to reason about the higher order relations between these relations. Whereas concepts such as *larger* and *smaller*, for example, are first-order relations, potentially derivable by comparing features of individual objects, a relation such as *opposite* is a higher order relation between relations (Gentner, 1983). The capacity to represent and reason with higher order relations has been considered central to human analogical thinking (Gentner, 2010; Halford, Wilson, & Phillips, 2010; Holyoak, 2012).

The development of knowledge about comparative relations provides a clear illustration of these human abilities. By the time they reach school age, children have acquired the ability to accurately assess whether one object (e.g., bear) is "larger" or "smaller" than another (e.g., fox), even under speed pressure (McGonigle & Chalmers, 1984). Moreover, like those of adults (Moyer & Bayer, 1976), children's judgments show a *symbolic distance effect*: The greater the magnitude difference between the two items, the faster the comparison can be made. Such symbolic comparisons are presumably based on stored representations of the perceptual dimensions associated with the individual concepts. A great deal of evidence—particularly, parallels between performance with symbolic and perceptual comparisons—suggests that humans and other species share a basic mechanism for representing continuous quantities on a "mental number line" (Dehaene & Changeux, 1993; Gallistel, 1993; Moyer, 1973). Moreover, rhesus monkeys are capable of learning shapes (Arabic numerals) corresponding to small numerosities (one to four dots), such that the shapes acquire neural representations overlapping those of the corresponding perceptual numerosities (Diester & Nieder, 2007).

These species-general achievements are impressive. However, human children go on to acquire a deeper understanding of comparative relations. For example, they learn that the relations *larger* and *smaller* have a special relationship to each other (a type of antonym). Analyses of corpora of child speech have identified systematic use of such gradable antonyms by children aged 2–5 years (Jones & Murphy, 2005), and experimental studies show that by at least 6 years of age children can use such concepts metaphorically (Gardner, 1974), and are aware that antonyms are contradictory (Glass, Holyoak, & Kossan, 1977). Children even-

Hongjing Lu, Departments of Psychology and Statistics, University of California, Los Angeles; Dawn Chen and Keith J. Holyoak, Department of Psychology, University of California, Los Angeles.

This research was supported by Office of Naval Research Grant N000140810186. John Anderson, Alex Doumas, John Hummel, Alexander Petrov, Dario Salvucci, and Alan Yuille provided valuable comments on an earlier draft of this article. We thank Charles Kemp for sharing Leuven inputs and Mark Steyvers for kindly providing us with topics inputs. MATLAB code for the simulations reported here is included in the supplemental materials.

Correspondence concerning this article should be addressed to Hongjing Lu, Dawn Chen, or Keith J. Holyoak, Department of Psychology, University of California, 405 Hilgard Avenue, Los Angeles, CA 90095-1563. E-mail: hongjing@ucla.edu, sdchen@ucla.edu, or holyoak@lifesci.ucla.edu

tually understand that a pair of concepts like *larger–smaller* is related in basically the same way as the pair *faster–slower*, allowing them to see that such pairs of relations form analogies.

It seems that "something special" happens that enables humans to acquire higher order relational representations. Animals of many taxa have the basic ability to detect and act based on perceptual relations, as exemplified by classic work on relational transposition in rats (Lawrence & DeRivera, 1954), and rudimentary numerical processing is clearly available to many primate and other species (see Gallistel, 1993). Nonetheless, there is a great deal of evidence that the relational capacities of humans exceed those of any other species, perhaps in a qualitative fashion (Penn, Holyoak, & Povinelli, 2008; Povinelli, 2000). The difference has been characterized as a human capacity for *relational reinterpretation*: the ability to transform perceptually grounded relations into explicit relational structures that distinguish the *roles* of relations from the objects that fill them (Doumas & Hummel, 2012), augmented by the additional ability to form higher order relational concepts (e.g., representations of hidden causes, or mental states of others).

From a computational perspective, the challenge is to explain what it might mean for a relation to be reinterpreted or rerepresented into a more explicit and abstract form, and to develop formal models of such a process. How could an inductive system ever get from some initial pool of perceptually available features to more abstract concepts corresponding to higher order relations (e.g., *opposite*), which seem not to be based entirely on the set of perceptual features that provided a starting point? The difficulty of the learning problem is compounded by evidence that children seem to acquire concepts largely from modest numbers of positive examples provided by adults (Bloom, 2000; see Xu & Tenenbaum, 2007).

An important part of the recipe for abstraction may be a pool of innate concepts. For example, Carey (2011) has argued, "There is no proposal I know for a learning mechanism available to nonlinguistic creatures that can create representations of objects, number, agency, or causality from perceptual primitives" (p. 115). But as Carey also argues, constructive mechanisms may operate over some combination of perceptual inputs and preexisting concepts to create new types of mental representations. Part of a learner's innate endowment may be processes that permit various forms of *bootstrapping*, whereby one type of representation is transformed into another. For example, there is evidence that analogical reasoning may play an important role in children's acquisition of natural number (Carey, 2011; Opfer & Siegler, 2007; see also Gentner, 2010; Kurtz, Miao, & Gentner, 2001).

Goals of the Present Article

In the present article we present a new model of the induction of relational representations, Bayesian analogy with relational transformations (BART). In general terms, BART is a computational-level model[1] (Anderson, 1991; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Marr, 1982) that employs bootstrapping to acquire and transform relational representations. We apply BART to the domain of relations related to comparative judgment. Although this is only a special case of the more general problem of relation learning, it is a domain that offers the advantage of a wealth of empirical evidence—behavioral, comparative, developmental, and neural—that can guide theory development.

Our particular focus will be on a restricted but nonetheless realistic subdomain: relations definable over continuous-valued features associated with animal concepts. The basic inputs provided to the model are vectors of feature values for a set of dimensions. Our first goal is to have the model learn a representation of first-order relations such as *larger* and *smaller*, *fiercer* and *meeker*, based on *empirical priors* (i.e., prior knowledge itself acquired by learning simpler concepts from relevant data) coupled with a limited set of positive examples instantiating relations. The use of empirical priors in learning is an example of a simple form of bootstrapping, whereby initial learning of a different or simpler concept provides a useful basis for acquiring more complex concepts. Similar ideas have been exploited in neural network models of learning (e.g., Bao & Munro, 2006; Elman, 1993; but see Rohde & Plaut, 1999). Newport (1990) argued that children's cognitive limitations (e.g., less capacity in working memory) may actually benefit certain aspects of language acquisition. Halford, Wilson, and Phillips (1998) proposed that children are able to learn one-place predicates (e.g., *large*, *small*) prior to two-place relations (e.g., *larger*, *smaller*) because the former require less working memory capacity. Given the strong evidence for this sequential progression in children's concept acquisition (e.g., Smith, 1989), we will focus on the potential use of one-place predicates as the basis for forming empirical priors to facilitate learning of comparative relations.

The use of only positive training examples makes it possible to acquire a stable and context-independent representation of a relation (whereas negative examples can be of many different types, and the learned relational representation will vary depending on which negative examples are encountered). We aim to demonstrate that the acquired representations of relations are generalizable (i.e., can be used to evaluate novel instantiations of the relations) and are sensitive to a basic factor that influences the difficulty of human relational judgments.

Our second goal is to show how these first-order relational representations can be transformed and rerepresented so as to allow the model to evaluate higher order analogy problems of the form A:B::C:D instantiated by the learned relations (e.g., *larger*:*smaller*::*fiercer*:*meeker*, rather than *fiercer*:*slower*). This transformation process is based on what we term *importance-guided mapping*, a subsymbolic form of analogical mapping based on similarity of weights associated with object features. Our overall aim is to provide a proof-of-concept that, for the domain of comparative relations, the capacity to solve structured analogy problems can be acquired by applying basically bottom-up learning mechanisms to raw inputs consisting of object concepts coded as simple feature vectors.

Judgments Based on Comparative Relations

Our target domain, comparative relations, is tied to a rich body of cognitive research. Comparative judgments exhibit a number of

[1] Although BART focuses on the computational level of analysis, its implementation includes assumptions at the level of representation and algorithm.

robust empirical phenomena. The most notable is the semantic distance effect (Moyer, 1973; Moyer & Landauer, 1967). Strong empirical evidence indicates that the long-term memory representation of a relation such as *larger* includes quantitative information that makes the difficulty of comparison decline as the magnitude difference increases. The symbolic distance effect is observed not only with quasiperceptual dimensions such as size, but also with more abstract dimensions such as animal intelligence (Banks, White, Sturgill, & Mermelstein, 1983) and such concepts as adjectives of quality (e.g., *good*, *fair*; Holyoak & Walker, 1976). Although magnitude representations exhibit analog properties, much like an internal number line (e.g., Woocher, Glass, & Holyoak, 1978), magnitude comparisons do not in general depend on visual imagery (Holyoak, 1977). Nonhuman primates also exhibit a distance effect for judgments of numerosity (see Nieder & Miller, 2003, for a review). Given its ubiquity, the distance effect is arguably the primary signature that a learned representation of a comparative relation is psychologically realistic; hence the distance effect will be the first empirical focus in our evaluation of BART. In the General Discussion we consider how BART might be extended to explain additional phenomena involving comparative judgments.

In human children, comparative adjectives emerge as early words in the lexicon, with clear developmental trends (Smith, 1989; Smith & Sera, 1992). In general, children progress from a global sense of similarity and dissimilarity of objects to learning one-place predicates that focus on specific dimensions of individual objects (*big*, *small*), to learning two-place comparative relations between multiple objects (*bigger*, *smaller*). As noted earlier, children eventually detect higher order similarities and differences between comparative relations, coming to understand, for example, that *higher* and *lower* are polar opposites. Less is known about the details of this part of the developmental progression, but presumably a prerequisite for learning a higher order relation approximating *gradable opposite* is to first achieve some degree of mastery with pairs of first-order comparative relations, such as *higher* and *lower*.

The acquisition of relations is intimately related to the development of analogical reasoning ability. A great deal of evidence indicates that children's ability to think analogically changes over the course of cognitive development (e.g., Z. Chen, Sanchez, & Campbell, 1997; Gentner & Toupin, 1986; Holyoak, Junn, & Billman, 1984; Tunteler & Resing, 2002, 2004). The developmental transition toward greater reliance on relational structure has been termed the *relational shift* (Gentner & Rattermann, 1991). The empirical phenomenon of a relational shift is well established, but there has been some debate regarding the developmental mechanisms that may underlie it. Considerable evidence indicates that some changes are maturational, involving increases in working memory capacity (Halford et al., 1998) and inhibitory control (Morrison, Doumas, & Richland, 2011; Richland, Morrison & Holyoak, 2006). However, it is universally accepted that learning new relations is a prerequisite for solving analogy problems based on these relations (Goswami, 1992, 2001). In the present article we focus on relation learning, the most basic mechanism required for analogical reasoning.

Approaches to the Acquisition of Relational Concepts

In recent years a number of different approaches to modeling the induction of relational concepts have been explored, which we will briefly review. We begin by laying out some criteria that we believe are of general importance in evaluating psychological theories of relation learning, including the present model.

1. Choice of inputs: The model should be capable of learning from inputs of realistic complexity that were independently generated. There is certainly much to be gained from exploratory work using small hand-coded inputs, and specifying realistic representations poses many challenges. However, without some tests using independently generated inputs, it is difficult to assess the extent to which a model may owe its successes to the foresight and charity of the modelers. In addition, the model (unless it explicitly assumes that all relational representations are innate) must be able to learn at least some relations from inputs that are *non*relational (e.g., object representations).

2. Learning efficiency: As a psychological model, learning should be achieved on a human time scale as measured by the number of training examples required to produce at least partial success. Given that children seem to be able to acquire preliminary understanding of many concepts from relatively few examples, a model should also be able to demonstrate efficiency by learning from a modest number. Although what "relatively few" means is inevitably vague, our focus will be on what can be learned from up to 100 or 200 positive training examples.

3. Generalization: The model should be able to make accurate relational judgments about novel examples. It is not sufficient to show that the model can learn the training examples as "relational facts"; it must also be able to apply its relational representations productively.

4. Performance difficulty: The difficulty of human relational judgments can be modulated by many factors. To be considered psychological, a model should account for at least some sources of differential difficulty in relational judgments for humans (and/or other animals).

5. Flexible reasoning: Relational knowledge plays an essential role in human reasoning and thinking, in essence providing a deeper source of information about conceptual similarity. Accordingly, the relational representations acquired by the model should be usable (either directly or after some additional learning process) to perform a variety of tasks that require relational reasoning (e.g., solving analogy problems).

These criteria are inherently qualitative rather than quantitative. Alternative assessment metrics could no doubt be advanced, but we have found the above criteria helpful in evaluating previous work on relational learning, as well as the models we test in the present article.

Vector Space Models

There is an extensive literature on automated methods for extracting relations based on the statistics of word or phrase co-occurrence in a large corpus of text. One class of methods, termed *vector space models*, originates from an information retrieval technique of the same name, and uses vectors or matrices in which the value of each element is derived from the frequency of some event, such as the frequency with which a certain word appears in a particular document or phrase (for reviews, see Turney, 2006; Turney & Pantel, 2010). For example, latent semantic analysis (LSA; Landauer & Dumais, 1997) yields vector representations of individual words by applying singular value decomposition to lexical co-occurrence data from a large corpus of text. LSA has proved useful in many applications that require measures of semantic similarity of concepts (Wolf & Goldman, 2003), including modeling the retrieval of story analogs (Ramscar & Yarlett, 2003). However, LSA vectors do not provide any direct basis for identifying abstract relations between concepts (although some modest results have been achieved by exploiting LSA vectors for relation words, such as *opposite*; Quesada, Kintsch, & Mangalath, 2004).

Related machine-learning algorithms have achieved greater success by working directly from co-occurrence data for word combinations found in a large corpus of text (Turney & Littman, 2005). Perhaps the most successful method is latent relational analysis (Turney, 2006), which has been applied to the task of solving Scholastic Aptitude Test verbal analogy problems (e.g., *quart*:*volume*::*mile*:*distance*). The algorithm searches for patterns of words in which the A and B term (and their synonyms) appear (e.g., "quarts in volume"). The frequencies of the various patterns are used to create a vector of relational features for A:B; vectors are similarly formed for potential C:D completions. Cosine similarity is calculated to compare the A:B vector to the corresponding vectors created for various alternative C:D pairs, and the most similar C:D is selected as the analogical completion. Latent relational analysis achieves a level of accuracy on Scholastic Aptitude Test analogy problems comparable to that attained by college students.

Vector space models such as latent relational analysis provide effective machine-learning tools for extracting relational similarity. However, these models operate directly on texts that include relational vocabulary. Our present focus is on learning from inputs based on representations of individual object concepts (including a set of such inputs that is derived from texts by a method similar to LSA).

Hierarchical, Generative Bayesian Models

Perhaps the most ambitious line of work has focused on hierarchical Bayesian models that integrate statistical learning with explicit representations of higher order relational structures (Goodman, Ullman, & Tenenbaum, 2011; Kemp, Perfors, & Tenenbaum, 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For example, Kemp and Tenenbaum (2008) showed how Bayesian techniques can operate on relational structures to learn systems such as hierarchies and linear orderings (see also Kemp & Jern, 2009; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006). In general terms, these hierarchical models are *generative* (MacKay, 2003) in the sense that representations of alternative relational structures are used to predict incoming data, and the data in turn are used to revise probability distributions over alternative structures. The highest level of the structure typically consists of a formal grammar or a set of logical rules that generates a set of alternative relational "theories," which are in turn used to predict the observed data.

Although hierarchical generative models are extremely powerful, the models to date have generally focused on systems of formal relations that have a well-defined logical structure known to the modeler (e.g., hierarchies, rings, or chains). The set of possible relational structures is provided to the system by specifying a grammar that generates them. Since the postulated grammar of relations is not itself learned, the generative approach (although certainly incorporating inductive learning) retains rather strong nativist assumptions.

Neural Network Models

The BART model, like generative Bayesian models, operates at the computational level; however, its emphasis on bottom-up learning and emergence overlaps with the goals of algorithmic approaches to relation learning and analogy based on neural networks (e.g., Gasser & Colunga, 2000; Jani & Levine, 2000; Leech, Mareschal, & Cooper, 2008; Rogers & McClelland, 2008; see McClelland et al., 2010). Our model shares the general aim of seeking emergence of structure from statistical operations over minimally structured inputs, coded as feature vectors.

A standard connectionist approach to learning relational structures has been to create a feed-forward network in which separate pools of input units are used to code features of an object in a role and of a relation. These pools interact via a hidden layer and thereby activate output units representing another filler of the role. Rogers and McClelland (2008) developed a model based on this type of architecture that learns simple propositions (e.g., "a canary can fly"). The model takes a sequence of input–output pairs and over repetitions adjusts the connection weights to learn facts of the form "canary" + "can" → "fly." The Rogers and McClelland model succeeds in capturing a number of important general characteristics of human learning, such as progressive differentiation of concepts and domain-specific feature weighting.

However, models of this sort (including that of Leech et al., 2008, a variation of the same architecture that aimed to account for how children learn to solve simple analogy problems) have not been shown to generalize to dissimilar training items, nor have they been extended to higher order relations. The Leech et al. (2008) model fails on even simple variations of its own training materials. For example, after being trained extensively with the various components required to solve the four-term analogy *apple*:*sliced-apple*::*bread*:*sliced-bread*, the model cannot generalize its knowledge to evaluate *sliced-apple*:*apple*::*sliced-bread*:*bread*, where the roles have been reversed (Holyoak & Hummel, 2008; see also French, 2008; Petrov, 2008).

A basic problem is that standard neural network models offer no way to represent relational roles. In connectionist networks of relation learning, both objects and relations are coded as distributed patterns of weights on links that serve as conduits for activation passed between units. The learned representations of relations therefore remain implicit, and relational knowledge cannot be accessed in a flexible fashion (cf. Halford et al., 2010). For example, in the Rogers and McClelland (2008) model, the representation of an object (e.g., canary) is inherently linked to a particular pool of relation-specific input units. As a consequence, after training the network that one thing a canary can do is fly, the model (unlike a human) would not be able to infer that one kind of thing that flies is a canary (i.e., make an inference in which *canary* serves as the output rather than input).

Symbolic Connectionist Models

The acquisition of relational structure has been a long-standing concern in the literature on analogical reasoning. Gick and Holyoak (1983) proposed that as a consequence of comparing and mapping one situation to an analogous one in a different content domain (e.g., a military and medical problem), humans can learn relational schemas for more abstract categories. Hummel and Holyoak (1997, 2003) developed a symbolic connectionist model, LISA (learning and inference with schemas and analogies), which is able to form such schemas by comparing and mapping examples. However, LISA's learning algorithm works by recombining preexisting (and hand-coded) relational concepts, rather than by building new relational predicates.

More recently, Doumas, Hummel, and Sandhofer (2008) developed a related model called DORA (discovery of relations by analogy) that addresses the fundamental goal of creating new relational predicates from nonrelational inputs. The basic representational assumptions of DORA are very similar to those of LISA, with both objects and roles of relations represented in a distributed fashion over a pool of semantic units. Relations are explicitly represented by localist units that code individual roles (e.g., *larger* would be coded by units for the larger object and for the smaller one in a pair). Bindings of objects to roles are coded dynamically in working memory by temporal patterns (synchrony in LISA, close asynchrony in DORA), and statically in long-term memory by conjunctive units. Because relations are represented explicitly and independently of their fillers, DORA (like LISA, but unlike classical connectionist models) is able to flexibly generalize relations to new contexts. But like more traditional neural networks, objects and relations are represented in the same basic way (as patterns of weights on links connecting units that code semantic features).

The basic learning algorithm used by DORA is to first compare feature representations of individual objects, creating new predicate units that connect to shared features (e.g., from the objects *elephant* and *bear*, a new one-place predicate connected to the shared feature *large* might be generated). Later, a pair of objects respectively instantiating the one-place predicates *large* and *small* (e.g., *elephant* and *mouse*) might be compared to another pair instantiating these same predicates (e.g., *walrus* and *frog*). With the aid of a comparator operator that can activate the features *more* and *less* based on the specific size values of paired objects, DORA might then generate the two-place predicate *larger*, with its first role connected to *more* and *large* and its second role to *less* and *small*. As additional examples of paired objects are encountered, sequential updating will refine the relational representation, honing in on features that prove to be invariant across examples.

The progression of learning comparative relations in DORA—from objects encoded as features to one-place predicates such as *large*, to two-place relations such as *larger* (that then undergo gradual refinement)—parallels the general developmental sequence identified by Smith (1989; Smith & Sera, 1992) and others in studies of children's acquisition of comparative relations. But although DORA can generate human-like patterns of relation learning, the robustness of its learning algorithm has not been extensively tested. So far DORA has only been tested with small hand-coded representations of objects as inputs, and the relations it learns are coded with features drawn from the same set already provided in these inputs. In particular, DORA assumes that in its inputs, all metric dimensions describing objects (e.g., size, speed) are coded by localist units. The model is also endowed with units representing relational features such as *more* and *less*, and with a comparator that will activate these relational features when given two objects associated with values on the same metric dimension. The model tacitly assumes that all relational predicates are definable by at least one precoded invariant feature (and the modelers ensure that the inputs satisfy this assumption).

Discriminative Bayesian Models

In contrast to the hierarchical, generative Bayesian models discussed above, simpler Bayesian models of category learning (e.g., Anderson, 1991; Fried & Holyoak, 1984) operate in a more bottom-up fashion. An important variant is *discriminative* Bayesian models (MacKay, 2003), which focus on learning the probabilities of categories given features (rather than the probabilities of features given possible categories). Discriminative models have been applied with considerable success to analysis of neural receptive fields in neurophysiology (Rust, Schwartz, Movshon, & Simoncelli, 2005; Victor, 2005) and construction of classification images in psychophysics (Eckstein & Ahumada, 2002; Lu & Liu, 2006). They also provide valuable tools in other complex statistical tasks, such as the recognition of brain states based on neuroimaging data (Bayesian decoding models; see Friston et al., 2008).

A discriminative Bayesian approach to relation learning was developed by Silva, Heller, and Ghahramani (2007), who applied their model to tasks such as identifying classes of hyperlinks between web pages; Silva, Airoldi, and Heller (2007) applied the same model to classifying relations based on protein interactions. Although this model was developed to address applications in machine learning, the general principles can potentially be incorporated into models of human relational learning. The BART model represents such an effort.

One key idea is that a relation can be represented as a function that takes a pair of 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 and then applies the learned representation to classify novel examples. A second key idea is that relation learning can be facilitated by incorporating empirical priors, which are derived via some simpler learning task that can serve as a precursor to the relation learning task. In particular, Silva, Heller, and Ghahramani (2007) explored the usefulness of first teaching the model a general distinction between related and unrelated object pairs and then using the learned representation of the general relation (*related*) as the empirical prior to bootstrap learning of each specific relation of interest. D. Chen, Lu, and Holyoak (2010) incorporated a similar empirical prior into a model for learning abstract semantic relations, such as *synonym* and *antonym*, from features derived by LSA (Landauer & Dumais, 1997).

These models have demonstrated some success in generalization tests involving identifying novel examples of learned relations.[2]

[2] In the present article we refer to this type of test as "relational generalization," whereas it has been called "analogical reasoning" in the machine-learning literature (Silva, Airoldi, & Heller, 2007). The task is indeed closely related to first-order analogical reasoning, in which the relation between A and B concepts (generally objects) is assessed to determine if it is sufficiently similar to the relation between C and D concepts (e.g., Turney, 2006). In contrast, the "analogy" problems described in the present article require second-order analogical reasoning, which is based on the similarity of relations between relations.

However, none of the models attempted to account for systematic sources of difficulty in human relational judgments, nor did they attempt to show that the learned relational representations could in turn be used to reason about higher order relations.

BART: Overview

Choice of Input Representations

BART's inputs are restricted to vectors representing objects, so that all the model's relational knowledge must be acquired from nonrelational inputs. Specifically, we focus on learning comparative relations from feature representations of animal concepts. In accord with the first of the criteria for model evaluation we laid out earlier, we wished to ensure that the inputs we used were not hand-coded by the modelers. We chose three sets of input representations that can be viewed as complementary in their advantages and challenges for testing a learning model.

The first set of inputs can be characterized as simple and transparent (low dimensionality, localist coding of magnitudes). These were feature vectors derived from human ratings of animals on four magnitude continua (size, speed, fierceness, and intelligence; Holyoak & Mah, 1981). No doubt it is oversimplified as a psychological model to assume that each dimension is coded by a single value; nonetheless, there is in fact strong evidence that humans and other primates are equipped with specialized neural circuitry for dealing with approximate magnitude on various dimensions (e.g., Cantlon, Brannon, Carter, & Pelphrey, 2006; Dehaene & Changeux, 1993; Fias, Lammertyn, Caessens, & Orban, 2007; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Piazza, Mechelli, Price, & Butterworth, 2006; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Pinel, Piazza, Le Bihan, & Dehaene, 2004). As a practical matter, the simplicity of the rating-based representations (comparable to that of the hand-coded representations employed by Doumas et al., 2008) will prove helpful in understanding how the model operates.

To assess BART's potential to scale up to learn relations from more complex inputs, we also applied the model to input vectors derived from much more challenging databases (high dimensionality, distributed coding of magnitudes). Our second set, which we will refer to as the "Leuven inputs," was based on norms of the frequency with which participants at the University of Leuven generated features characterizing various animals (De Deyne et al., 2008). Each animal in the norms is associated with a set of frequencies across more than 750 features. Although some features in the Leuven inputs have prima facie relevance to the dimensions of interest to us, none were as direct as the Holyoak and Mah (1981) ratings of specific magnitude dimensions. The Leuven inputs have been successfully used as inputs for a Bayesian model of categorization (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011).

Our third set of inputs was taken from the topics model (Griffiths, Steyvers, & Tenenbaum, 2007). The topics model is broadly similar to LSA (Landauer & Dumais, 1997), taking words in documents as its input and yielding approximate semantic representations of individual words as its output. The topics model uses Bayesian inference to associate each word in the corpus with a set of "topics," which theoretically generate the words. For example, a topic that could loosely be characterized as finance would tend to generate such words as *money*, *savings*, and *bank* (in the sense of financial institution). For each word, a vector (typically of length 300) based on conditional probabilities of each topic given the word can be interpreted as a distributed semantic representation over feature values. Relative to the Leuven inputs, the topics inputs were much more opaque, in that the meaning associated with each topic is generally difficult to characterize; unlike the rating inputs, individual topics do not correspond in any obvious way to the magnitude dimensions underlying the critical comparative.

Vectors based on the Leuven inputs or topics avoid any hand coding of inputs by the modeler. There is thus is no danger that we the modelers have inadvertently planted to-be-discovered relations in the inputs provided to our learning model. Whereas the simple vectors based on human ratings provide magnitude information very directly, the more complex Leuven and topics vectors do not. To preview our computational results, BART achieves near-perfect performance on generalization and analogy tests after learning from the rating vectors, excellent performance using the much larger Leuven inputs, and reliable though imperfect performance based on the yet more complex topics inputs.

Of course, there is no reason to believe that any of these representations directly correspond to the inputs available to human children when they first learn basic relations. The Leuven inputs perhaps come closest, as they include many features of animals that children would likely know. Children have much more direct access to perceptual and motoric features of objects, which can guide relation learning (e.g., Maouene, Hidaka, & Smith, 2008). In addition, children's learning of relations is clearly guided by linguistic cues from adults (e.g., Yoshida & Smith, 2005).

Nonetheless, children surely are faced with considerable complexity in the inputs from which some relations are acquired; hence any plausible model will have to demonstrate robustness. By testing BART with inputs derived from three independent sources, we can have some confidence in the robustness of qualitative aspects of model performance that hold true across all three inputs. For the Leuven and topics inputs, the learning task demands that in a high-dimensional space, BART must infer distributed patterns of features that implicitly code the dimensions over which the model aims to learn relations. In addition, the model must then remap the acquired weight distributions to solve structured analogy problems. The complexity of the learning task would likely be comparable (or greater) for inputs further enriched by perceptual and motoric features. In the General Discussion we consider how the approach used by BART might be extended to operate on such inputs.

Overview of the Operation of BART

Stages. Broadly speaking, BART proceeds in two stages:

First-order relation learning. Given feature vectors corresponding to pairs of objects, the model uses statistical learning to update weights associated with feature dimensions for various comparative relations (e.g., *larger*, *fiercer*), and then uses its learned weights to decide whether novel pairs instantiate a specified relation. As shown in the right-hand plot in Figure 1, BART represents a relation using a joint distribution of weights over object features. Weight distributions code not only first-order statistics (means) but also second-order statistics (variances and

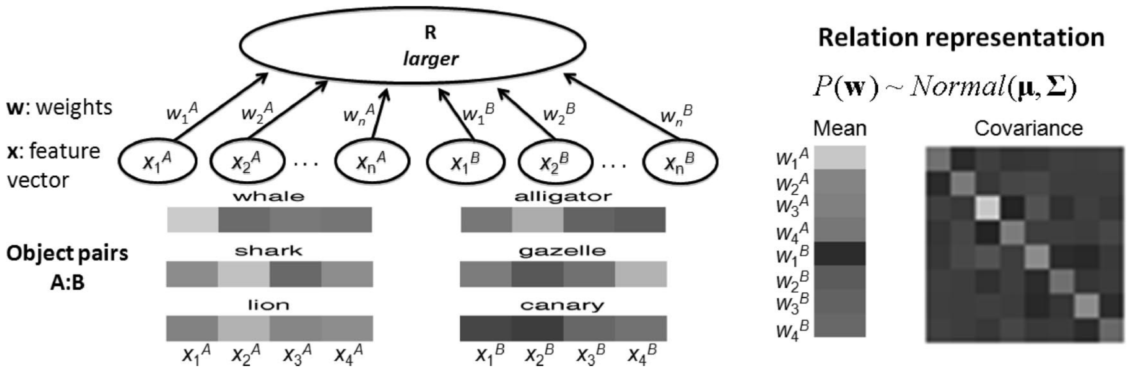

Figure 1. Graphical representation of the general framework for relation learning in Bayesian analogy with relational transformations. (Left) Two objects **A** and **B** in a pair are represented as a vector **x** of *n* features for each object; vector **w** represents the unknown relational weights that define a relation *R*, which is learned by using the training set of examples instantiating this relation (e.g., *whale–alligator*, where the intensity of cells represents feature values on each dimension; light indicates high positive values, dark high negative values). (Right) The relation is represented as the joint normal distribution of weights **w**. The normal distribution is defined with two parameters: the mean weights vector (shown in the mean plot, in which the intensity indicates the values of mean weights) and the covariance matrix of weights including the variance of each weight (diagonal cells in the covariance plot) and the covariances among them (off-diagonal cells in the covariance plot).

covariances) that capture the uncertainty of the estimated weights, as well as interweight correlations.

Importance-guided relation mapping. To evaluate potential analogies between pairs of relations (e.g., *larger*:*smaller*:: *fiercer*:*meeker*), the model rearranges the order of dimensions in acquired weight distributions for the source and target relation pairs to yield transformed relation representations. The transformation is based on an assessment of importance of each dimension in the source pair, and on the correspondence of weight patterns between the source and the target pair.

Learning first-order relations. BART is capable of learning flexibly from any combination of positive and negative examples; however, we focus on learning from positive examples only (as children are able to do; see Bloom, 2000). Importantly, positive examples make it possible to achieve a relatively context-free relational representation, rather than one that varies with the particular types of negative examples included in the training set. In addition, because children often appear to learn useful approximations of concepts from small numbers of examples, we aimed to make learning in BART as efficient as possible, focusing on what the model can learn from a modest number of examples (a range of up to about 200). Also, children's relation learning is clearly guided by linguistic inputs from adults (e.g., Gentner, Anggoro, & Klibanoff, 2011; Yoshida & Smith, 2005). In natural speech to children, comparative relations are given names, such as "larger," which are explicitly connected to positive examples ("the elephant is larger than the hippo"). Accordingly, BART focuses on supervised learning using labeled positive examples.

Bayesian framework. BART learns a first-order relation by estimating the distribution of a corresponding weight vector **w** from a set of training pairs that constitute examples of that relation, as schematized in Figure 1. We adopt a Bayesian framework to learn the probability distribution of $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S})$, where $\mathbf{X_S}$ represents the feature vectors for object pairs in the training set, the subscript **S** indicates 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 vector **w** constitutes the learned relational representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in **X** for relation judgment. Learning a first-order relation is based on estimating the posterior distribution of weights, which can be computed by applying Bayes's rule using the likelihood of the training data and the prior distribution for **w**:

$$P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}) = \frac{P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S})P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S})P(\mathbf{w})}. \quad (1)$$

The likelihood is defined as a logistic function for computing the probability that a pair instantiates the relation, given the weights and feature vectors,

$$P(R = 1 \mid \mathbf{w}, \mathbf{X}) = (1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{X}})^{-1}. \quad (2)$$

This likelihood function has been used in Bayesian logistic regression analysis, and in similar Bayesian models of relation learning described by Silva, Airoldi, and Heller (2007) and Silva, Heller, and Ghahramani (2007). The logistic function is also commonly used in neural networks to introduce nonlinearity into activation functions.

We assume that the prior $P(\mathbf{w})$ in Equation 1 follows a multivariate normal distribution, $P(\mathbf{w}) \sim N(\boldsymbol{\mu}_0, \mathbf{V}_0)$ with a mean of $\boldsymbol{\mu}_0$ and a covariance matrix of $\mathbf{V}_0$. A primary focus of the present article is on the potential role of informative priors in relation learning. The key to efficient statistical learning is a good choice of priors, especially when the learning problem involves high dimensionality. Proposals for priors typically stem from abstract theory (Griffiths & Tenenbaum, 2009; Kemp & Tenenbaum, 2008; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) or analyses of statistics of the natural environment (Geisler, 2008; Griffiths & Tenenbaum, 2006; Lu, Lin, Lee, Vese, & Yuille, 2010; Simoncelli & Olshausen, 2001). Here we explore a variation of what are termed *empirical priors*, which are themselves learned from relevant data, combined with a *hyperprior* for variances of weights.

Empirical priors. BART takes advantage of the potential for inductive bootstrapping, using previously acquired knowledge of simpler concepts to establish empirical priors that guide subsequent learning of more complex concepts. Previous work has explored use of a general relation (*related*) as an empirical prior for learning more specific relations (D. Chen et al., 2010; Silva, Airoldi, & Heller, 2007). Here we consider the potential usefulness of more specific empirical priors tailored to individual relations. There is strong linguistic evidence (across many languages) that two-place comparatives are derived from corresponding relative adjectives either by adding a morpheme (e.g., *large* yields *large* + *er*, termed the *synthetic* form) or by creating a phrase using *more* or *less* (e.g., *intelligent* yields *more intelligent*, termed the *periphrastic* form; see Graziano-King & Cairns, 2005). Psychological evidence also indicates that comparative relations such as *higher* are initially derived from the corresponding one-place predicates (e.g., *high*; see Smith, Rattermann, & Sera, 1988). In choosing the appropriate priors, it seems probable that children are guided by lexical similarities (e.g., *larger* is similar to *large*, *smaller* to *small*). However, to increase the generality of the model, we make the weaker assumption that the learner must infer the most relevant one-place predicate from the actual pairs used as positive training examples for the comparative.

We use one-place predicates as the building blocks for creating empirical priors. To determine which one-place predicate should be used to construct the empirical prior for learning a particular relation, we developed a simple categorization algorithm to select the one-place predicates based on training data. First, we train 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. As schematized in Figure 2, category learning of one-place predicates is conducted with Bayesian logistic regression, with a standard normal distribution for weights (i.e., mean 0 and variance 1) as the prior to infer the weight distribution $P(\mathbf{w_c} \mid \mathbf{X_c})$, in which $\mathbf{w_c}$ indicates the weight vector corresponding to feature dimensions of an object, and $\mathbf{X_c}$ denotes the extreme animals in each group used for category learning.

Second, we employ a simple voting procedure to select the "best" category of one-place predicates based on the training examples for the comparative. For each pair of objects (X^A, X^B) in the training data for relation learning, we compute the probability that each individual object is a member of each category of one-place predicates, obtaining $P(C|X^A)$ and $P(C|X^B)$, respectively. If $P(C|X^A) > P(C|X^B)$ for a pair, a score of 1 is assigned to this category; otherwise, a score of 0 is assigned. These scores are summed over all the pairs of training data. In effect, the procedure for prior selection aims to identify 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 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. Although more sophisticated categorization models could be employed, this simple procedure proved adequate for our present purposes. For the most difficult set of inputs (topics), the method achieved near-perfect selection of the appropriate one-place predicate when given 100 training examples.

Third, the category that yields the highest summed score is selected to set mean weights for the first role of a comparative. Although the model is in effect informed that the relation to be learned involves a comparison of two objects, the basis for the comparison must be learned. 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.[3] For our example, if *large* were to provide the basis for the empirical priors, then the priors for the comparative relation would include the weights of *large* for the first role and the opposite weights for the second role, as shown in Figure 2. If there is a tie in the highest summed score between categories of one-place predicates, we simply take an average of the multiple weight vectors to generate empirical priors.

Hyperprior. Confidence about the empirical priors bootstrapped from object concepts is represented by the variances in the prior distribution of weights. A simple model is to assume the same degree of confidence for all the individual weights in the empirical prior $\boldsymbol{\mu}_0$. Alternatively, confidence may vary from one dimension to another, affording greater flexibility. We adopt the method of automatic relevance determination (MacKay, 1992; Neal, 1996) to define the precisions of the empirical prior using hyperparameters. Specifically, the prior for the *i*th weight in vector $\mathbf{w}$ is assigned in the form of a normal distribution in which the mean is from the empirical prior and the variance is $1/\alpha_i$:

$$P(w_i|\alpha_i) \sim N\left(\mu_{0i}, \frac{1}{\alpha_i}\right), \tag{3}$$

where the value of α_i (also termed precision, the inverse of variance) controls the certainty about mean weight values derived from the empirical prior. Thus increasing α_i values imply greater confidence that w_i is similar to μ_{0i} in the empirical prior. We use a conjugate prior distribution in the form of a gamma distribution for α_i with two hyperparameters, a_0 and b_0, to constrain the precision of each weight:

$$P(\alpha_i) \sim \text{Gamma}(a_0, b_0). \tag{4}$$

Inference algorithm. Although the general framework of the relation learning model is straightforward, the inference step is nontrivial because the calculation of the integral in Equation 1 lacks an analytic solution. A sampling approach is impractically slow for dealing with high feature dimensionality, and hence would unduly limit the generality of the model. Accordingly, as in Silva, Heller, and Ghahramani (2007), we employed the variational method developed by Jaakkola and Jordan (2000) for Bayesian logistic regression to obtain a closed-form approximation to the posterior distribution. Variational methods are a family of methods that transform the problem of interest into an optimization problem

[3] We considered the alternative of forming empirical priors from a combination of two one-place predicates (e.g., *large* and *small* might be used to set priors for *larger*). However, developmental evidence indicates that young children often treat such polar opposites as disjoint, whereas children clearly link the primary one-place predicate to its corresponding comparative (e.g., *large* to *larger*; see Smith et al., 1988).

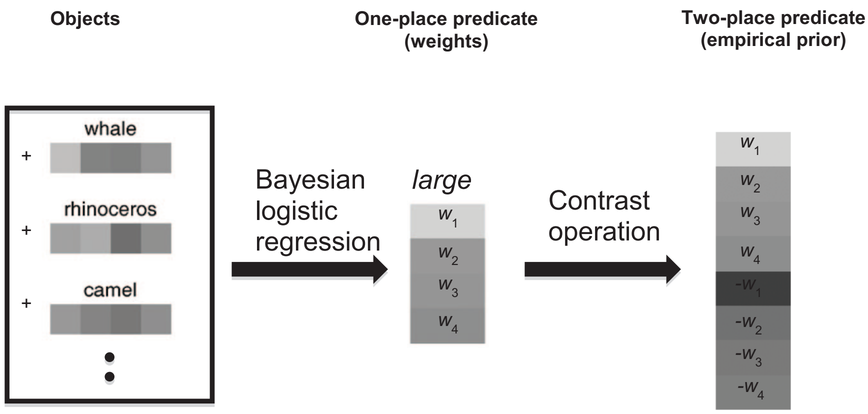

Figure 2. Illustration of the construction of an empirical prior for a comparative relation (*larger*) by bootstrapping from prior learning of weights for a related one-place predicate (*large*), in turn derived from features of individual objects (large animals).

by introducing an extra *variational parameter*, ξ, which is iteratively adjusted to obtain successively improving approximations. The input to the learning model includes training data $\mathbf{X}$, composed of N training pairs and their corresponding relation labels R in which 1 indicates that the pair of words instantiates the relation (positive examples) and -1 indicates it does not (negative examples). The variational updates are applied until convergence or a maximum number of iterations is reached. For learning with an empirical prior, the model starts from the prior mean μ_0 (i.e., bootstrapping from knowledge about the corresponding one-place predicates), and with $\mathbf{V}_0$ assumed to be an identity matrix with variances 1 and covariances 0. On each iteration the variational parameter ξ is updated, along with the mean of the weight vector, $\boldsymbol{\mu}$, and the covariance matrix, $\mathbf{V}$, with the following updating equations:

$$\mathbf{V}^{-1} = \mathbf{V}_0^{-1} + 2\sum_{n=1}^{N} \lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^T$$

$$\boldsymbol{\mu} = \mathbf{V}\left[\mathbf{V}_0^{-1}\boldsymbol{\mu}_0 + \sum_{n=1}^{N} R_n\mathbf{x}_n/2\right]$$

$$\xi_n^2 = \mathbf{x}_n^T[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^T]\mathbf{x}_n, \tag{5}$$

where $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$.

For learning with a hyperprior, the variational method is iteratively applied to update the mean, the covariance matrix, and the hyperparameters, as follows:

$$\mathbf{V}^{-1} = \mathbb{E}_\alpha(\mathbf{A}) + 2\sum_{n=1}^{N} \lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^T$$

$$\boldsymbol{\mu} = \mathbf{V}\left[\mathbb{E}_\alpha(\mathbf{A})\boldsymbol{\mu}_0 + \sum_{n=1}^{N} R_n\mathbf{x}_n/2\right]$$

$$a = a_0 + 1/2$$

$$b_i = b_0 + ((w_i - \mu_{0i})^2 + V_{ii})/2$$

$$\xi_n^2 = \mathbf{x}_n^T[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^T]\mathbf{x}_n, \tag{6}$$

where w_i is the ith element of weight vector $\mathbf{w}$, μ_{0i} is the ith element of empirical prior μ_0, V_{ii} is the ith diagonal element of covariance matrix $\mathbf{V}$, and $\mathbb{E}_\alpha(\mathbf{A})$ is a diagonal matrix with its ith diagonal element given by a/b_i.

Model evaluation on generalization test. To test generalization of the learned relational representation, we conduct a transfer task using new pairs of words, denoted by the subscript T. Given the training pairs $\mathbf{X_S}$ and their labels $\mathbf{R_S}$, the model aims to calculate the posterior predictive probability that a target pair $\mathbf{X}_T$ instantiates the learned relation:

$$P(R_T = 1 \mid \mathbf{X}_T, \mathbf{X_S}, \mathbf{R_S}) = \int_{\mathbf{w}} P(R_T = 1 \mid \mathbf{X}_T, \mathbf{w})P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}). \tag{7}$$

The posterior predictive probability can be approximated with the variational posterior (i.e., the lower bound of the predictive probability), which can be computed in a single pass through the training data set $\{\mathbf{X_S}, \mathbf{R_S}\}$ applying the updating equations as specified in Equation 5. Hence, the probability predicted for a transfer pair (i.e., Equation 7) can be approximated as

$$\log P(R_T = 1 \mid X_T, \mathbf{X_S}, \mathbf{R_S}) = \log(g(\xi_T)) - \frac{\xi_T}{2} + \lambda(\xi_T)\xi_T^2$$
$$- \frac{1}{2}\boldsymbol{\mu}_S^T\mathbf{V}_S^{-1}\boldsymbol{\mu}_S + \frac{1}{2}\boldsymbol{\mu}_T^T\mathbf{V}_T^{-1}\boldsymbol{\mu}_T + \frac{1}{2}\log\frac{|\mathbf{V}_T|}{|\mathbf{V}_S|}, \tag{8}$$

where $\boldsymbol{\mu}_\mathbf{S}$ and $\mathbf{V_S}$ denote the parameters in $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S})$ after learning from the training pairs, and $\boldsymbol{\mu}_T$ and $\mathbf{V}_T$ denote the parameters in $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}, \mathbf{X}_T, R_T = 1)$ found by adding the target pair to the training set.

Higher order relation mapping. In order for any model to have a chance to solve higher order relational analogies (i.e., analogies based on relations between relations), it must first acquire at least approximate representations of the relevant first-order relations. However, as the example in Figure 3

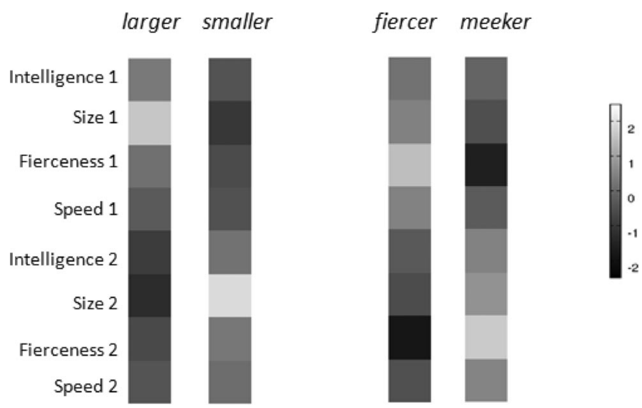

Figure 3. Successful learning of comparative relations is not sufficient to solve four-term analogy problems such as *larger*:*smaller*::*fiercer*:*meeker*, because high (positive or negative) weights are on different dimensions for *larger*+*smaller* versus *fiercer*+*meeker*.

makes clear, successful learning of comparative relations will not in itself guarantee solution of analogy problems such as *larger*:*smaller*::*fiercer*:*meeker*. For example, if we were to compare the learned distributions for *larger*+*smaller* to those for *fiercer*+*meeker*, we would find that the two joint distributions are essentially uncorrelated. Roughly, in the former the size dimension has large (positive or negative) weights while the other dimensions have weights near 0, whereas in the latter the fierceness dimension has large weights and the rest have weights near 0. Without any mechanism to map different salient dimensions to one another, any implicit similarity would remain hidden.

To solve higher order analogy problems, BART employs an algorithm for importance-guided mapping. In general terms, the algorithm aims to find a mapping between the dimensions for the A:B relation and those for the C:D relation that minimizes a distance measure defined over the weight distributions. Because the full search space for this correspondence problem scales exponentially with the number of dimensions, we employ a greedy search algorithm (Friston et al., 2008), a type of procedure designed to make locally optimal choices with the hope of approximating the global optimum. More specifically, the algorithm develops a one-to-one mapping between dimensions sequentially on the basis of the overall "importance" of dimensions. In essence, the algorithm minimizes correspondence errors for more important dimensions at the possible cost of greater errors for less important dimensions.

In more detail, we assume that an analogy problem in the form A:B::C:D is evaluated by first focusing on the relation in the source (A:B), and then determining how well the target relation (C:D) maps to A:B. The algorithm prioritizes dimensions in proportion to their importance in A:B. Specifically, the mapping algorithm first searches for a dimension in C:D that is most similar to the most important dimension in A:B; it then searches for a dimension that maps to the second most important dimension in A:B among the remaining pool of dimensions in C:D, and so on until each dimension in A:B is mapped to a unique dimension in C:D. Qualitatively, BART aims to map important dimensions in A:B to dimensions in C:D that influence relation classification in an analogous way.

Figure 4 schematizes the algorithm for importance-guided mapping. Intuitively, *larger*+*smaller* and *fiercer*+*meeker* are alike in that each has a key important dimension (size and fierceness, respectively). Moreover, importance has a clear numerical definition based on the absolute magnitudes of weights (normalized by their variances). To evaluate an analogy in the form A:B::C:D (e.g., *larger*:*smaller*::*fiercer*:*meeker*), the model first assesses the importance of each dimension for A:B and then reorders the dimensions (and transforms the distributions of mean weights) accordingly. The transformed representation of A:B can be ob-

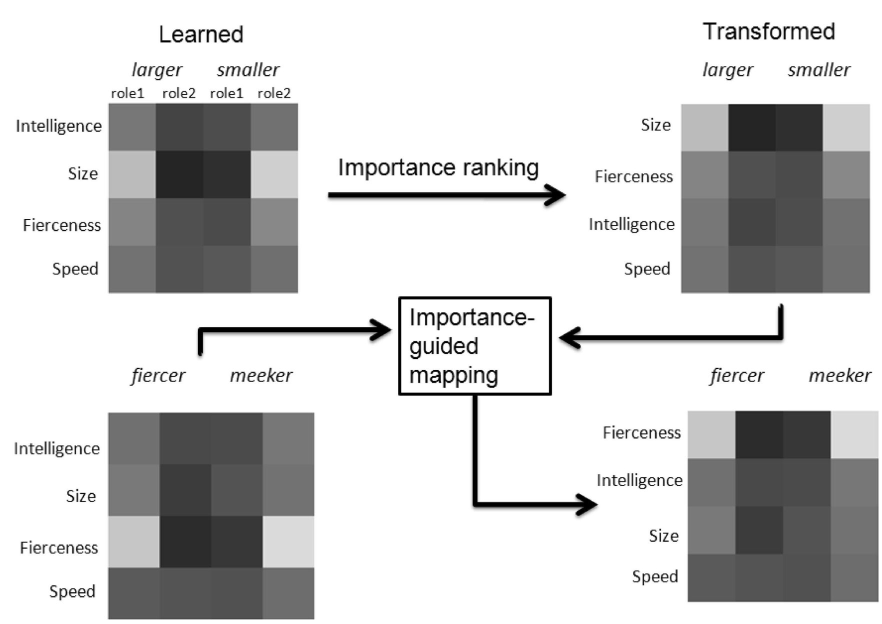

Figure 4. Illustration of importance-guided mapping for solving an analogy problem.

tained in three steps: (a) compute the normalized weights using mean weight values divided by their standard deviations; (b) sum up the absolute values across the two roles in each of the two relations in A:B to get an importance score for each feature dimension; (c) rank-order the dimensions (maintaining consistency across the two roles of both relations) based on the importance index.

Next, for each dimension in A:B, BART selects the dimension in C:D with the most similar pattern of weight distributions. Here we take advantage of a natural property of multivariate normal distributions. The marginal distribution over a subset of multivariate normal random variables can be obtained by dropping the irrelevant variables (the variables that one wants to marginalize out) from the mean vector and the covariance matrix. The marginal weight distributions for each feature dimension can therefore be easily calculated for A:B and C:D, respectively. Then the similarity of marginal distributions is evaluated by computing a distance measure between two distributions.[4] The *J-divergence distance* is employed to maintain the symmetric property of a distance measure by summing up two Kullback–Leibler (KL) divergences (Cover & Thomas, 2006),

$$D(p, q) = KL(p \| q) + KL(q \| p), \tag{9}$$

where p and q denote two distributions, and

$$KL(p \| q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx.$$

The advantage of using normal distributions is that it becomes possible to solve analytic expressions for the distance measure with the means and covariance matrices of the two normal distributions,

$$D(p, q) = \frac{1}{2}(\mu_p - \mu_q)^T (V_p^{-1} + V_q^{-1})(\mu_p - \mu_q) + \frac{1}{2} tr[V_p^{-1} V_q + V_q^{-1} V_p - 2I_d], \tag{10}$$

where $tr[\cdot]$ denotes the matrix trace.

Finally, having transformed the C:D distribution to reflect its mapping to A:B, BART uses the overall J-divergence distance between the two transformed distributions as its measure of how well the C:D relation matches that of A:B. For our example, the transformed representations will identify size as the most important dimension for *larger+smaller*, and then select fierceness as the dimension for *fiercer+meeker* that has the most similar mean weight distribution to that of size for *larger+smaller*. The resulting transformed distributions will map size to fierceness, thereby contributing to a lower overall J-divergence distance (i.e., higher similarity) for A:B::C:D than for a C:D′ foil such as *fiercer+slower*.

Note that BART evaluates relational analogy problems without forming an explicit representation of a higher order relation such as *opposite*. Rather, BART estimates the degree of match between the C:D and A:B relations under the assumption that similar relations (whatever they may be) will generate lower J-divergence distance between the two mean weight distributions based on the correspondences produced by importance-guided mapping. In the General Discussion we consider how an extension of BART might go on to acquire explicit representations of higher order relations.

Tests of BART Using Rating Inputs

Inputs. The rating vectors used as inputs to BART were based on norms reported by Holyoak and Mah (1981), collected for use in a study of symbolic magnitude comparisons. Holyoak and Mah had 25 undergraduates rate the subjective magnitude of each of 80 animal names on four continuous dimensions: size, speed, fierceness, and intelligence. Ratings were made on a 9-point Likert scale, with a rating of 9 indicating maximum magnitude. Magnitude norms for each dimension were then derived by successive interval scaling (Bock & Jones, 1968). This method provides a simultaneous normalization of the responses to each item across the nine response categories, yielding what can be interpreted as an interval scale. The resulting values (which for each dimension correlated .99 with mean ratings) were normalized to range from 0 through 10. A few examples are shown in Table 1 (see Holyoak & Mah, 1981, Table 1, p. 200, for the entire set of norms). Because the ratings reflect subjective magnitude differences, the norms incorporate the typical nonlinear relationship between subjective and objective magnitudes (e.g., the norms indicate that the difference in subjective size between a goldfish and a cat is roughly the same as that between a deer and a hippopotamus). Of the 80 animals in the norms, topics representations were available for 77, and all simulations were based on this subset. Intercorrelations among the four dimensions across the 77 animals were moderate, ranging from .38 (size with speed) to .60 (size with fierceness).

Each of the 77 words thus initially corresponded to a vector of four continuously valued features, with all values being nonnegative. However, the logistic likelihood function used by BART is designed to map values between negative and positive infinity onto the outcome variable, with the value of 0 serving as the natural midpoint of the input scale. Accordingly, we centered the rating vectors by a linear transformation, subtracting from each value the mean value for that dimension across all 77 words. Thus the feature values in the vectors used as inputs to BART included both negative and positive values with means of 0.

In our tests, both training and test items were created by randomly selecting pairs of animals and concatenating their rating vectors. Thus, each input vector had two components: the four features of Animal 1 and the four features of Animal 2. To ensure that the differences in magnitudes between animals in a pair were likely to be distinguishable by humans, we constrained all training

[4] A simpler variant of importance-guided mapping is based on just maximum a posteriori probability (MAP) estimates (i.e., mean weights) rather than on entire covariance matrices. Indeed, in exploring all three data sets reported in the present article, we have found that the simpler variant yields virtually the same performance as the version based on the full covariance matrix. The covariance matrix plays important roles in guiding the acquisition of the MAP estimates during learning, and aids in relational generalization, but apparently is not essential in subsequent analogical processing. In the present article we describe the complete version of importance-guided mapping for the sake of computational generality, but the simpler MAP variant may well be more psychologically realistic (see General Discussion). The MATLAB code provides the simpler variant as an option.

Table 1
Examples of Ratings of Animals on Four Dimensions of Magnitude

Animal	Size	Fierceness	Intelligence	Speed
Alligator	5.46	8.88	3.67	5.03
Cow	6.52	3.95	3.35	4.59
Flea	0.00	2.52	0.24	3.65
Goldfish	1.91	1.35	1.45	4.18
Moose	7.04	5.98	3.96	6.29
Mouse	2.41	3.08	3.34	5.02

Note. Adapted from "Semantic Congruity in Symbolic Comparisons: Evidence Against an Expectancy Hypothesis," by K. J. Holyoak and M. A. Mah, 1981, *Memory & Cognition, 9,* p. 200. Copyright 1981 by Springer. Adapted with permission.

and test pairs to be based on animals differing by at least 0.50 on the relevant dimension. Under this criterion, over 2,000 animal pairs were available as positive examples for each to-be-learned relation.

Training. For the purpose of generating empirical priors, the 20 animals that were "greatest" and "least" on each dimension were first used to train BART to classify each of the eight possible one-place predicates (i.e., *large*, *small*, *fierce*, *meek*, etc.). For this initial phase of learning, the priors on all weights were set to standard normal distributions (i.e., means of 0, variances of 1, covariances of 0).

In learning two-place relations, we tested two models. The first model was a version of BART that selected empirical priors for means of weights based on one-place predicates as described earlier (e.g., the mean weights for *large* might be selected to provide the priors for the first role of *larger*, with the second role set by replicating the weights for *large* as a contrast). Priors for variances were set to 1, and those for covariance were set to 0. (Because the learning task with rating inputs proved to be extremely easy for BART, a hyperprior was not used in these simulations.) For comparison, a baseline model simply used uninformative priors (standard normal distributions). We trained and tested BART on each of the eight comparative relations involving the animal ratings (*larger*, *smaller*; *fiercer*, *meeker*; *smarter*, *dumber*; *faster*, *slower*). If one assumes that training examples are randomly sampled from the same population, the solutions for polar-opposite relations (e.g., *larger* and *smaller*) would be expected to converge at asymptote with symmetrical weight distributions (i.e., distributions with weights reversed between the two roles). This result was clearly obtained, so we will only report generalization results for the four "greater" relations. However, the analogy results are based on learned representations of all relations ("lesser" as well as "greater").

Generalization performance.

Basic tests. On each run, we trained the model on some number (1–100) of randomly selected pairs that constituted positive examples of the target relation (and satisfied the minimum difference criterion). All the remaining pairs in the pool (both positive and negative examples) were then used as test pairs. For test pairs, negative examples were created by simply reversing the "correct" order of the two animals for the target dimension. The number of test pairs that instantiated a relation was always equal to the number that did not instantiate it (since they involved the same animals in reverse order).

A test pair that instantiated the relation was counted as correct if its posterior predictive probability of being an example of the relation was greater than .5, whereas a test pair that did not instantiate the relation was counted as correct if its predicted probability was less than .5. This criterion assumes that the model is unbiased. When trained solely with positive examples, it is plausible that a learning model might develop an overall bias favoring a "yes" response. Based on signal detection theory, sensitivity after correcting for possible bias can be measured with the A_z measure (Dorfman & Alf, 1969), which calculates the area under the receiver operating characteristic (ROC) curve. For the ratings data, the criterion of .5 in fact proved to be optimal prior to reaching ceiling accuracy in generalization performance, indicating that BART's generalization decisions were unbiased within this range. Accordingly, we will simply report percent correct.

All reported results are based on the average performance over 100 runs, each of which randomly selected a set of training pairs from the pool. Figure 5 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Learning was very successful for all relations. The BART model with empirical priors generalized moderately accurately after a single training example (mean of 71% correct across all relations), and reached 96% correct after 20 training trials. BART's learned representations of one-place predicates thus provided effective empirical priors for the two-place comparative relations. The baseline model with uninformative priors (means of 0) started at a substantially lower level of performance (mean of 59%) and required about twice as many training examples (40) to reach 95% accuracy. After 100 training examples, both models converged at near-perfect accuracy (99% correct) in generalization. These results demonstrate that at least when magnitude information is transparently coded in small input vectors, BART

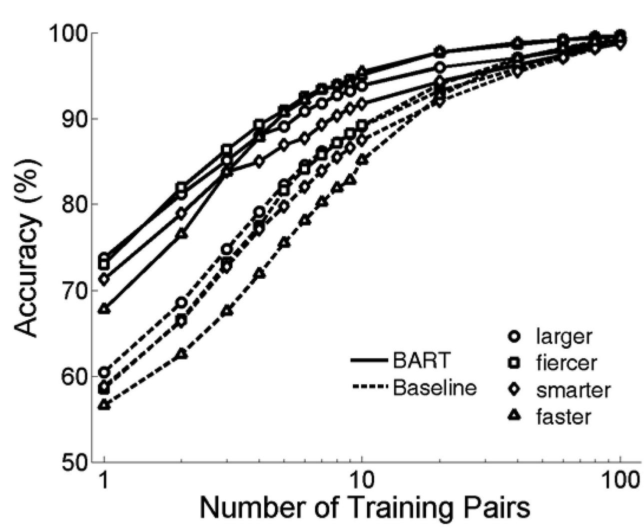

Figure 5. Accuracy in the generalization task with rating vectors as a function of the number of training examples for the four comparative relations (log scale). Solid lines indicate the performance of Bayesian analogy with relational transformations (BART) with the empirical prior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

can learn comparative relations very efficiently from a modest number of positive examples, especially when guided by empirical priors.

To determine whether the relational representations acquired by BART yield the ubiquitous symbolic distance effect obtained for comparative judgments by humans, we examined how BART's probability estimates (using the full model with empirical priors) relate to the rated subjective distance between each test pair of animals on the dimension of interest (i.e., size, fierceness, intelligence, or speed). Distance effects are generally revealed in reaction time paradigms. Although BART does not provide a process model of speeded judgments, standard models of reaction time (e.g., Link, 1990) would predict that reaction time as a measure of judgment difficulty will have an inverse monotonic relationship to the log ratio of posterior probabilities that each ordering of a pair fits the indicated relationship (e.g., for a pair such as *elephant–horse*, a positive log ratio will indicate that elephant is larger than horse, with the predicted difficulty of the discrimination decreasing as the log ratio becomes increasingly positive).

Distances were grouped into five bins based on interitem distance in ratings on the relevant continuum (i.e., animals very similar in size fell in Bin 1, animals maximally different in size fell in Bin 5). Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0 to 10: Bin 1 (distances between 0.5 and 2), Bin 2 (distances 2–4), Bin 3 (distances 4–6), Bin 4 (distances 6–8), and Bin 5 (distances 8–10). Figure 6 plots the log ratio of the predicted posterior probability for each positive test pair compared to the predicted probability for the reversed pair as a function of distance between the pair after learning based on 40 training pairs, averaged across the four comparative relations. Consistent with a symbolic distance effect, the log ratio increases with distance.

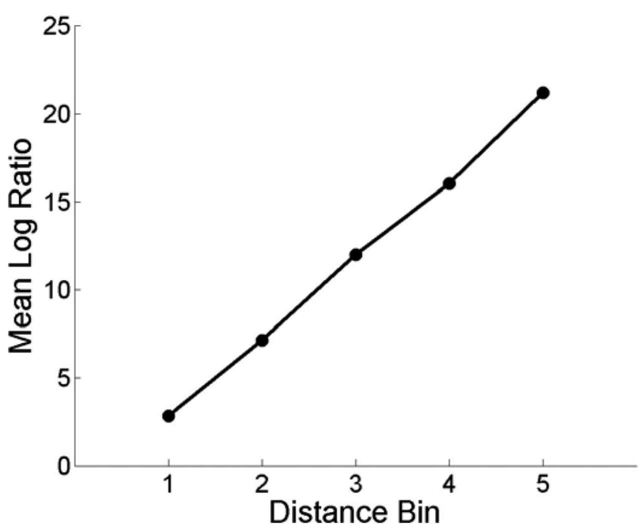

Figure 6. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (rating inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0 to 10: Bin 1 (distances between 0.5 and 2), Bin 2 (distances 2–4), Bin 3 (distances 4–6), Bin 4 (distances 6–8), and Bin 5 (distances 8–10). Results are collapsed over the four continua.

Generalization beyond the range of training examples. The basic generalization tests described above always involved test pairs that had not been shown during training. We also performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

1. One test introduced pairs in a distance range outside that used in training. Using empirical priors set in the same manner as described previously, we trained BART on the relation *larger* based on 40 positive examples drawn randomly from the first three distance bins only (e.g., 40 pairs of animals exhibiting small or moderate size differences for *larger*). We then tested the model's generalization performance at each of the five distance bins, using all possible animal pairs excluding the training pairs. We again obtained a monotonic increase in mean log ratio across all levels of distance: 2.80, 7.12, 12.00, 16.07, and 21.13 for Bins 1–5, respectively. The model thus assessed pairs of animals with large size differences (Bins 4–5) as the best positive examples of *larger*. BART's acquired representation of *larger* was sufficiently robust and flexible as to enable very accurate generalization to novel test pairs exhibiting size differences greater than the range presented during training.

2. Another series of generalization tests varied the magnitudes of the individual training and test objects. For this purpose all the animals were sorted into four roughly equally sized groups based on their value on the relevant dimension in the Holyoak and Mah (1981) norms, such that animals in Group 1 have the lowest values and animals in Group 4 have the highest values. We then trained the model with 100 examples based on pairs of the form [4, 1]. In other words, the first animal is drawn from Group 4 and the second animal is drawn from Group 1 (i.e., for learning *larger*, the first animal is very large and the second animal is very small). The generalization test included all and only pairs of the form [3, 2] (i.e., middle-sized animals). BART's performance was similar across the four "greater" relations with an overall accuracy of 91%, indicating very successful generalization.

3. Because pairs of the form [3, 2] are necessarily close in magnitude, a generalization test that includes only pairs of the form [3, 2] is inherently more difficult than one composed of pairs formed from all groups. For comparison with Test 2, we also trained the model in the usual way (positive examples formed from all groups) and then tested it on only pairs of the form [3, 2]. BART performed similarly across the four "greater" relations on this test as well, achieving an overall accuracy of 99% after 100 training examples. In comparison, the 91% accuracy obtained in Test 2 is somewhat lower, indicating that restricting the magnitude range of the training items impaired generalization to some extent.

4. Another test involved training with 100 pairs of the form [2, 1] (i.e., pairs of small animals) and testing with those of the form [4, 3] (i.e., pairs of large animals), or the reverse (training on [4, 3], then testing on [2, 1]). This test is inherently difficult because the training items are drawn from a restricted range and the test items are drawn from a different restricted range (and, moreover, are very close in magnitude). Averaged across the two variations, generalization accuracy was 65%, 79%, 77%, and 81% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus transfer from one extreme on a continuum to the other was reliable, although imperfect.

5. A final test ensured that the animals (not just pairs) used during training and testing did not overlap by selecting a random half of the animals for training and then testing on all pairs formed by the remaining animals. After 100 training examples, BART achieved 98% overall accuracy, indicating very successful generalization to animals not encountered during training.

Analogy performance. To test BART's ability to solve higher order analogy problems using its acquired relational representations, we constructed problems based on the comparative relations. If the model is able to implicitly learn relations between relations, then its standard training on the four sets of paired comparatives should allow it to solve analogies based on two distinct higher order patterns, which we will gloss as "same–extreme" (e.g., the relationship of *larger* to *fiercer*, or *smaller* to *meeker*; see Clark, 1970, for a discussion of the polarity of comparative relations) and "opposite" (e.g., the relationship of *larger* to *smaller*, or *fiercer* to *meeker*). Table 2 gives examples of five types of four-term analogy problems that can be constructed by pairing one of the two higher order relations with various foils, using the first-order relations acquired by BART. The first types are based on same–extreme, with the foil being either an opposite pair (Same–O) or a pair of relations at the opposite extreme of their respective dimensions (Same–OE). The other three types are based on opposite. The foil could be split across two dimensions (Opp–S), reverse polarity on a dimension (Opp–R), or involve a conflict (Opp–C) in which one relation in the foil was in fact identical to one of the A:B terms. In such problems the analogical answer C:D has to overcome the misleading featural identity of the D′ term in the C:D foil to the B term in A:B. Except for Same–OE problems, the C:D′ (or C′:D) foils always share one word with the analogical C:D completion.

For the first four types, chance performance would be 50% if the A:B and/or C:D relations had not been acquired. J-divergence, like other proposed measures of relational similarity that have been used to model human judgments (e.g., Goldstone, 1994; Taylor & Hummel, 2009), is sensitive to featural as well as relational overlap. For Opp–C analogies, expected performance in the absence of relation learning would therefore be 0%, because the featural overlap based on the word shared by A:B and the foil C:D′ would cause the foil always to be selected as more similar to A:B. The Opp–C conflict set thus provided an especially challenging test of BART's ability to solve higher order analogies based on its learned relational representations, directly pitting relational against featural similarity. Similar designs have been employed in studies of human analogical mapping both with adults (e.g., Markman & Gentner, 1993) and children (Richland et al., 2006).

To test BART's capacity to make analogical inferences, we created sets of each of the five types (see Table 2 for the number of each type). BART's assessment was counted as correct (i.e., as an analogical response) if the calculated J-divergence distance was lower for the analogical C:D pair than for the nonanalogical foil, C:D′ (or C′:D). Figure 7 shows the performance of BART and the baseline model (both using the identical algorithm for importance-guided mapping) on the five types of analogy problems. Although both models performed extremely well after learning from ratings inputs, BART achieved slightly higher success after fewer training examples. The advantage of BART over the baseline model in efficiency of learning to solve analogy problems is most apparent for type Opp–C. After three training examples, BART begins to show more accurate performance than the baseline model, and the two models do not converge in their performance until after 60 training examples, at which point both models achieve essentially perfect performance on all problem types.

These results demonstrate that the algorithm for importance-guided mapping is in fact capable of solving structural analogies based on learned representations of first-order relations, implicitly finding correspondences between nonidentical dimensions based on the importance-guided mapping operation with marginal weight distributions. Moreover, the empirical priors proved to be effective in establishing relational distributions that support analogical reasoning, especially in competition with featural similarity.

In summary, the tests with ratings vectors provide a first demonstration that BART is able to learn relational representations that pass five critical tests: (a) learning from nonrelational inputs; (b) learning with high efficiency; (c) generalizing to new examples of first-order relations; (d) capturing a key source of differential difficulty in human performance, symbolic distance; and (e) supporting structured analogical reasoning.

Tests of BART Using Leuven Inputs

Inputs. We next applied BART to the much more challenging problem of learning comparative relations from high-dimensional input representations based on the Leuven database (De Deyne et al., 2008). As noted earlier, these norms are based on the frequency with which participants generated features characterizing 129 animals, for 759 features. To make the results as comparable as possible to those obtained with the ratings inputs, we used the subset of 44 animal names from the Holyoak and Mah (1981) norms that were also included in the Leuven database. Although this subset was substantially smaller than the 77 animals used in the simulations based on

Table 2
Examples of Analogies Based on the Relations "Same–Extreme" and "Opposite," With Various Types of Foils

Analogy test type	No. of examples	Target	Foil
Same–O	48	larger:fiercer::smarter:faster	larger:fiercer::smarter:stupider
Same–OE	48	smaller:stupider::meeker:slower	smaller:stupider::faster:fiercer
Opp–S	48	faster:slower::smarter:stupider	faster:slower::smarter:meeker
Opp–R	24	larger:smaller::smarter:stupider	larger:smaller::stupider:smarter
Opp–C	48	fiercer:meeker::smarter:stupider	fiercer:meeker::smarter:meeker

Note. Same–O = same–extreme (opposite as foil); Same–OE = same–extreme (opposite–extreme as foil); Opp–S = opposite (split pair as foil); Opp–R = opposite (reversed as foil); Opp–C = opposite (conflict as foil).

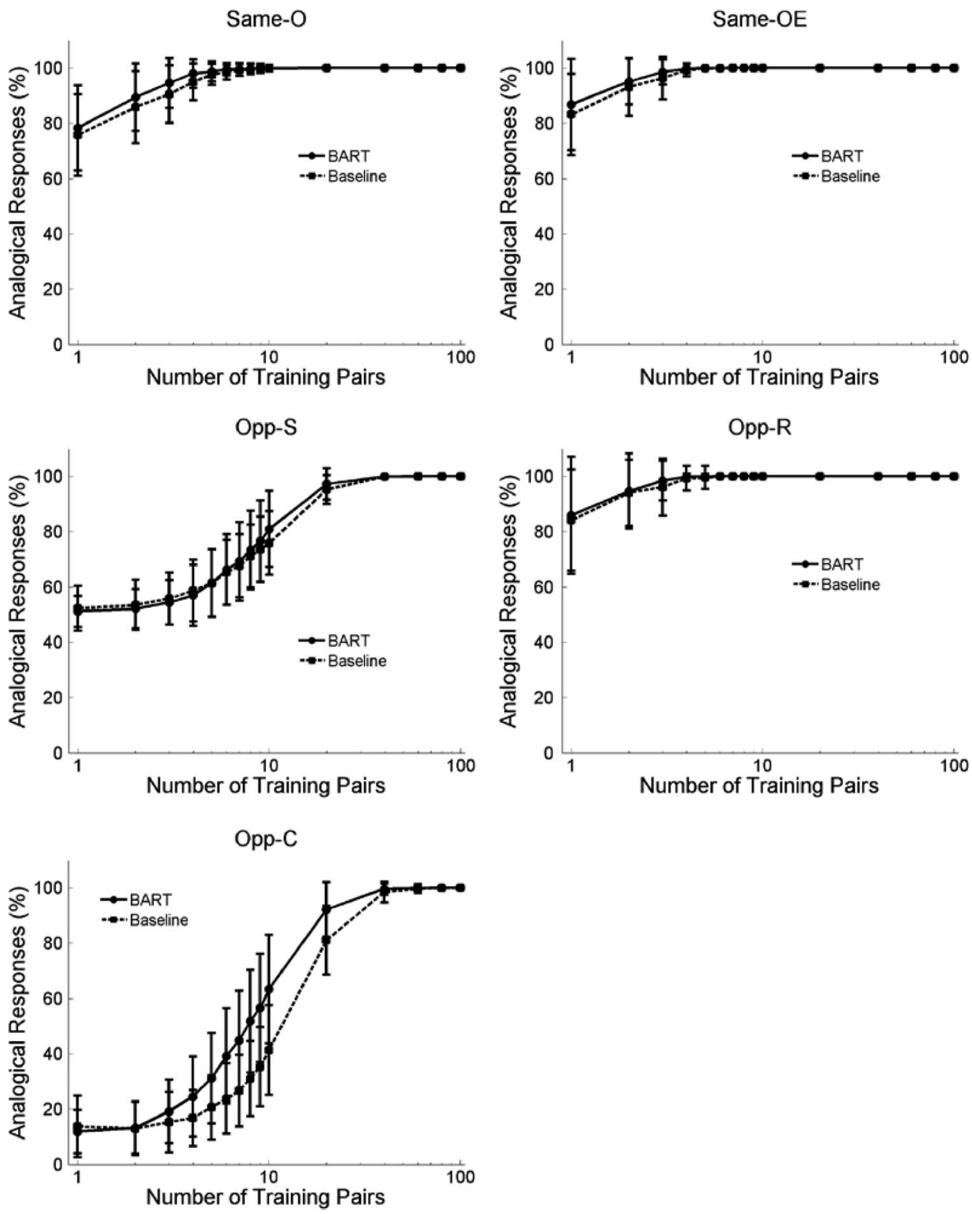

Figure 7. Proportion of analogical responses as a function of the number of training examples (log scale) with rating inputs for the five types of analogy problems. Solid lines present the results for Bayesian analogy with relational transformations (BART) with empirical prior; dashed lines present results for baseline model (with uninformative prior). Error bars indicate 1 standard deviation (results based on 100 runs). Same–O = same–extreme (opposite as foil); Same–OE = same–extreme (opposite–extreme as foil); Opp–S = opposite (split pair as foil); Opp–R = opposite (reversed as foil); Opp–C = opposite (conflict as foil).

ratings, it was still large enough to generate a pool of over 750 pairs for training and generalization tests.

To construct input representations, we followed the procedure used by Kemp, Chang, and Lombardi (2010, p. 219), who used the Leuven norms to estimate the probability of each feature conditional on each animal (see their Equation 8, top). We multiplied the computed probability by 100 to make the magnitude range roughly comparable to that of the rating inputs. Because the values as described so far are based on probabilities, they necessarily are nonnegative. As noted in connection with the rating vectors, to optimize the scale for the logistic likelihood function, it is desirable to center the vectors by a linear transformation. Accordingly, we subtracted from each feature the mean value of that feature across all 129 animals in the Leuven norms. The feature values in the vectors used as inputs to BART therefore included both negative and positive values, with means near 0.

To reduce the size of the search space, we focused on the most important dimensions. Specifically, we summed the feature vectors for the 44 animals and identified the 50 dimensions that yielded the largest sums (after dropping one dimension, "is small," that was clearly redundant with another, "is big"). By using just these 50 most important dimensions to form vectors for each word, the total size of the vector for each word pair was fixed at 100.

Training. The basic training regimen was very similar to that employed with the rating vectors. To create empirical priors, we again selected 20 animals close to each of the two extremes on each of the four dimensions of interest. These included all the extreme animals included in the subset of 44 for which Leuven vectors were available (the number ranging from 8 to 15 across the eight sets). We augmented this core group with additional animals from the entire Leuven set of 129 animals that we judged to be close to the relevant extremes, thus bringing the total number of animals in each set to 20. Insofar as some of the animals used to train one-place predicates may not have been the most extreme, and many were not included in the subset of 44 used to train relations, this procedure for selecting positive examples for learning empirical priors would be expected to make successful relation learning more challenging.

The search space for the Leuven representations was much larger relative to that for the ratings inputs used previously. Accordingly, we aimed to improve the stability of the estimates for empirical priors by increasing the number of examples. Given that the set of positive examples available for each one-place predicate was necessarily constrained, we augmented the training pool by including negative examples. To learn *large*, for example, BART was given both 20 positive examples (i.e., 20 large animals) and 20 negative examples (i.e., 20 small animals). As in the case of our simulations with rating vectors, direct training on each comparative relation (e.g., *larger*) was still based solely on positive examples.

To help cope with the greater complexity of the learning problem with high dimensionality, we used a hyperprior to increase BART's representational flexibility. On the basis of a preliminary search of the parameter space, we set the values of the hyperparameters (a_0, b_0) to be 5 and 1, respectively. We found that allowing BART to use the hyperprior (with hyperparameters fixed for all simulations) tended to improve its generalization performance by about 2 percentage points relative to using the standard covariance matrix (the procedure used in the simulations with rating inputs), and significantly improved accuracy in certain analogy tests. For comparison, we also tested the same baseline model as that used with ratings vectors (i.e., Bayesian logistic regression with standard normal distributions as uninformative priors).

Generalization performance.

Basic tests. All reported results are based on the average performance over 10 runs, each of which randomly selected a set of training pairs from the pool. Figure 8 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Not surprisingly, given the greatly increased dimensionality of the learning problem, the level of performance was lower overall than was obtained with the rating vectors. However, the full BART model, with empirical priors on mean weights and a hyperprior on variances, achieved substantial generalization (about 80%–95% accuracy for the four "greater" relations after 100 training examples). The baseline model showed much weaker generalization performance, achieving only about 60%–70% accuracy overall after 100 training examples.

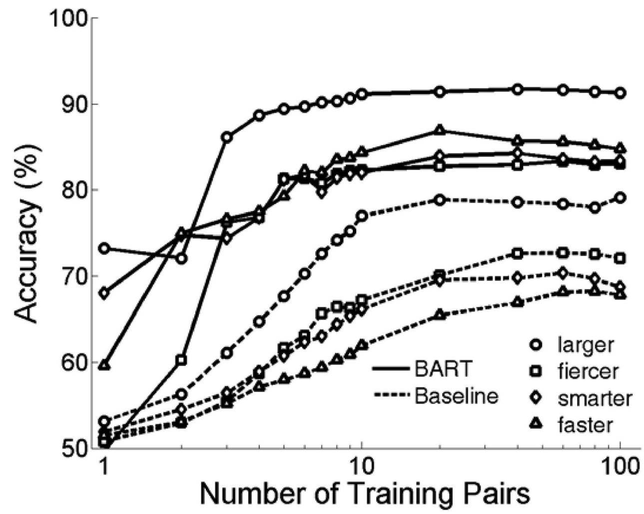

Figure 8. Accuracy in the generalization task with Leuven inputs as a function of the number of training examples (log scale) for the four comparative relations. Solid lines indicate the performance of Bayesian analogy with relational transformations (BART) with the empirical prior and hyperprior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

We also explored how BART's generalization performance changed with more extended training. Whereas BART appeared to be unbiased when trained with up to 80–90 examples, a further increase in the number of training examples led to a bias toward "yes" responses. This type of response bias leads to reduced accuracy if a fixed decision criterion is used. Accordingly, we computed the A_z measure, which is more robust to response bias (Dorfman & Alf, 1969). Generalization performance as measured by A_z continued to improve slightly with increased numbers of training examples. After 700 training examples, both BART and the baseline model achieved an A_z value of about .95.

To examine whether the relational representations that BART derives from Leuven vectors yield the distance effect obtained for comparative judgments by humans, we examined how BART's generalization performance relates to the rated subjective distance between each test pair of animals on the dimension of interest (as measured with the Holyoak & Mah, 1981, norms). Figure 9 plots the mean log ratio of predicted probabilities for positive versus negative test pairs as a function of distance on the relevant dimension between the two animals in a pair (after learning from 100 training examples). Because only 44 of the animals in the Holyoak and Mah (1981) norms are included in the Leuven data set, we used four distance bins instead of five. The log ratio of posterior probabilities increased monotonically with distance. Thus, the relational representations that BART acquired from Leuven inputs clearly yield a symbolic distance effect.

Generalization beyond the range of training examples. As in the case of the simulations based on ratings, we performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

1. Training with 100 examples from distance Bins 1–2 only and testing on all four distance bins. We obtained a monotonic increase in mean log ratio across all levels of distance: 1.02, 2.30,

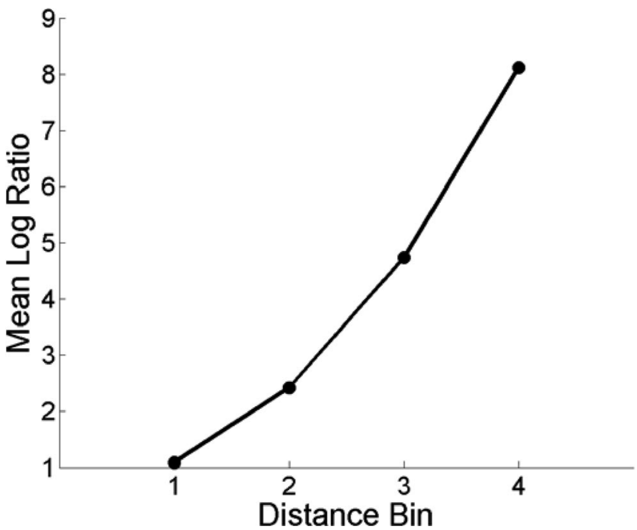

Figure 9. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (Leuven inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms: Bin 1 (distances between 0.5 and 1.5), Bin 2 (distances 1.5–3), Bin 3 (distances 3–5.5), and Bin 4 (distances 5.5–10). Results are collapsed over the four continua.

4.51, and 8.00 for Bins 1–4, respectively. These results again demonstrate that the model assessed pairs of animals with large size differences, in Bins 3 and 4, as the best positive examples of *larger*.

2. Training with 100 pairs of the form [4, 1] and testing on all pairs of the form [3, 2]. BART's accuracy was 89%, 67%, 73%, and 92% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization performance based on a restricted set of training inputs.

3. Standard training with 100 pairs formed from all groups and testing on all pairs of the form [3, 2]. BART achieved accuracies of 87%, 69%, 77%, and 87% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus for the Leuven inputs, restricting the training pairs to those of the form [4, 1] (Test 2) had minimal negative impact on generalization performance.

4. Training with 100 pairs of the form [2, 1] and testing on all pairs of the form [4, 3], or the reverse. Averaged across the two variations, generalization accuracy was 85%, 73%, 77%, and 54% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization across magnitude extremes for the first three relations.

5. No overlap between training and test animals. This test was performed with 60 training examples because the Leuven subset included only 44 animals in total. BART achieved accuracies of 92%, 79%, 84%, and 85% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization to animals not encountered at all during training.

Content of learned weight distributions. To convey a sense of the content that BART used to learn comparative relations from the Leuven inputs, Figure 10 depicts typical mean weights for the four "greater" relations that the model acquired using 100 training examples. For each relation the 50 dimensions are ordered by importance. Several qualitative observations are of interest. First, the representations are clearly contrastive, with positive (light) weights associated with important weights on the first role and negative (dark) weights associated with the second role, or vice versa. Second, among the more important weights, the positive value is predominantly associated with the first role. This is the type of relational information that indicates to BART that these comparatives are in fact oriented toward the "greater" extremes of their respective continua.

Third, the representations are highly distributed. For each relation, upward of 20 corresponding dimensions (i.e., 40 weights) show clear contrasts between the two roles. Unlike the rating vectors, in which a single dimension provided a localist code for each continua, the Leuven vectors lack any single dimension that suffices to define any comparative relation. To take the most salient example, one might have supposed that "is big" would be sufficient to predict relative size. In fact, although this dimension is indeed the single most important predictor of which object is larger, it is far from sufficient. The Leuven dimensions were derived from the frequencies with which participants generated features, rather than from a continuous rating scale of the sort used to create the Holyoak and Mah (1981) norms. Accordingly, in the Leuven data set, animals for which size is a salient dimension (often in reference to a subcategory) tend to have higher feature values for "is big." Based on a comparison of feature values on that dimension alone, the Leuven data set indicates that, for example, an eagle is larger than a hippopotamus, a seagull is larger than a horse, and a cow is the same size as a pelican. However, BART is able to flexibly integrate weakly predictive information provided by dozens of individual dimensions to successfully learn and generalize the comparative relations.

Analogy performance. The distributed nature of the relation representations acquired from the Leuven inputs posed a strong test of BART's algorithm for importance-guided mapping. Although this algorithm was extremely successful when applied to localist representations derived from the rating data, it was far from obvious whether it would also be effective with distributed representations. We tested BART and the baseline model on the five types of analogy problems in the same manner as for the rating inputs. The results are shown in Figure 11. The overall level of performance is lower than was obtained when the models were trained with ratings inputs, which is not surprising given the much greater complexity of the Leuven inputs. Indeed, it was far from obvious that the algorithm for importance-guided mapping would work at all when applied to representations of first-order relations that are highly distributed over many dimensions, rather than being localized on one critical dimension (as was the case for the simulations based on rating vectors; see Figure 4). As indicated in Figure 10, the first-order relations acquired from the Leuven vectors generally involved at least 40 reliable predictor variables working together.

In fact, the performance of the BART model on the analogy tests was excellent, achieving essentially perfect accuracy after 100 training trials on four problem types, and about 90% on Opp–C problems. For the Same–O, Same–OE, and Opp–R analogy types, the baseline model does not catch up to BART until after 40 training pairs. For the Opp–S and Opp–C analogy types, performance of the baseline model hovers around chance (50% and 0%, respectively) even after 100 training trials. For the latter problem types, we explored the impact of more extended training on analogy performance for the baseline

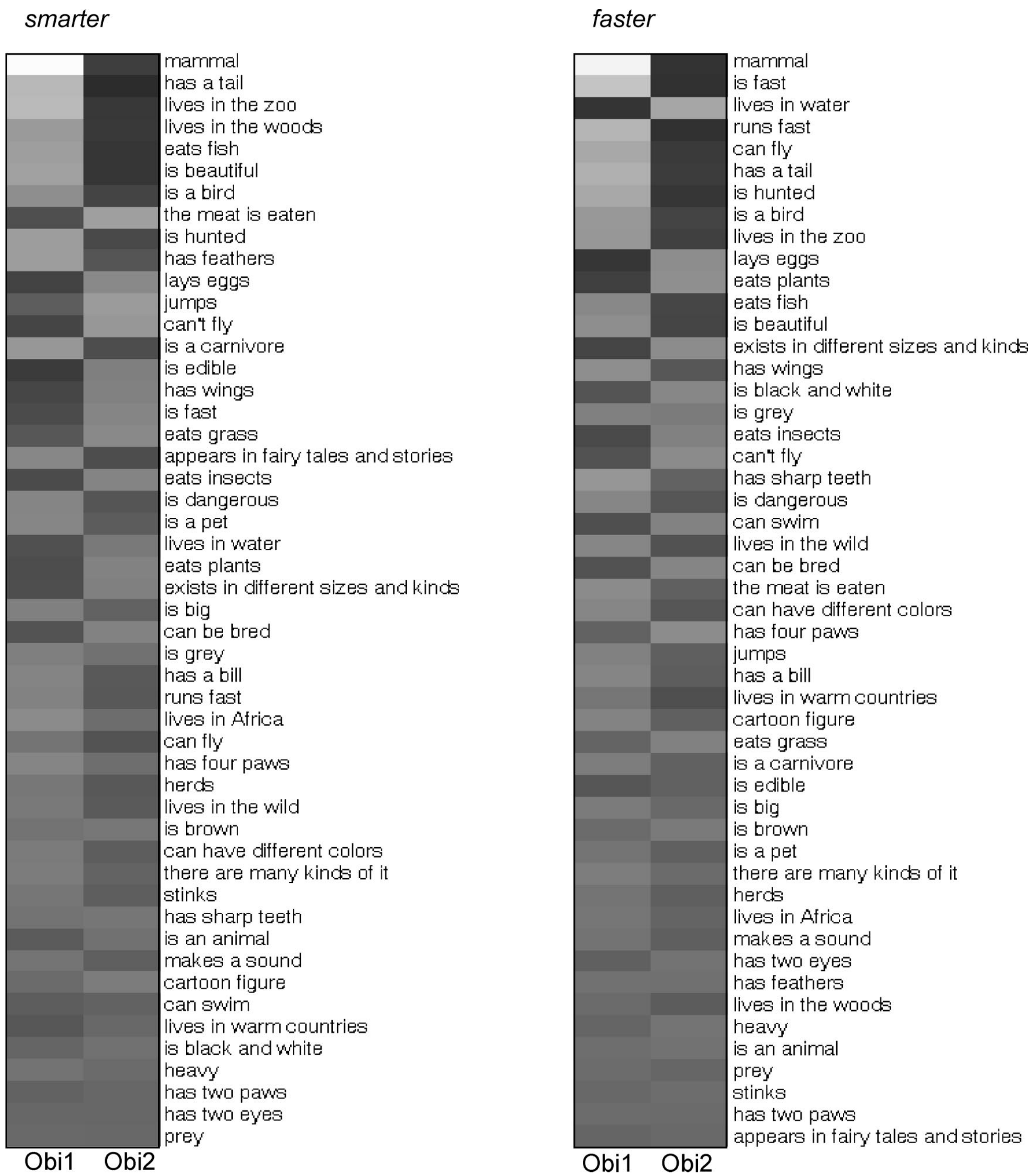

Figure 10. Illustration of mean weights for four relations learned from 100 training examples with Leuven inputs. For each relation, the weights on 50 dimensions (based on the specified query) are rank-ordered by importance. The intensity of cells represent weight values on each dimension (light indicates high positive values, dark indicates high negative values). The first column corresponds to weights on features of the first object, and the second column corresponds to weights on features of the second object.

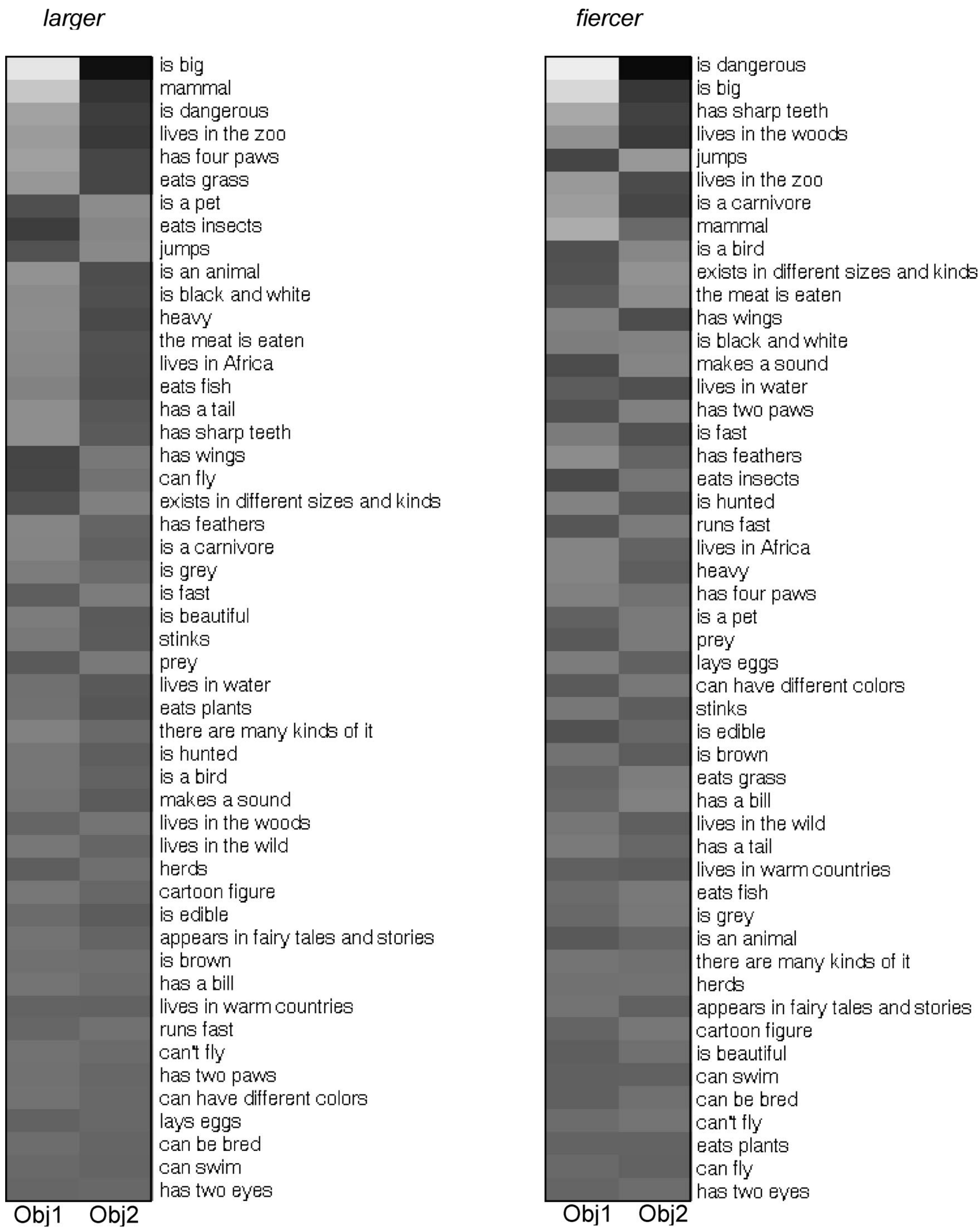

Figure 10 (continued).

model. Even after 700 training examples, the baseline model still lagged behind BART by about 9 percentage points on Opp–S problems and 46 percentage points on Opp–C problems. In sum, when faced with high-dimensional inputs based on Leuven inputs, BART was able to achieve substantial success in solving structured analogy problems, with its informative priors playing a decisive role.

Tests of BART Using Topics Vectors

Inputs. We also applied BART to the yet more challenging problem of learning comparative relations from input representations taken from the topics model (Griffiths et al., 2007). Whereas the ratings vectors has clear localist codes for the critical continua (size, fierceness, etc.), and the Leuven vectors included some

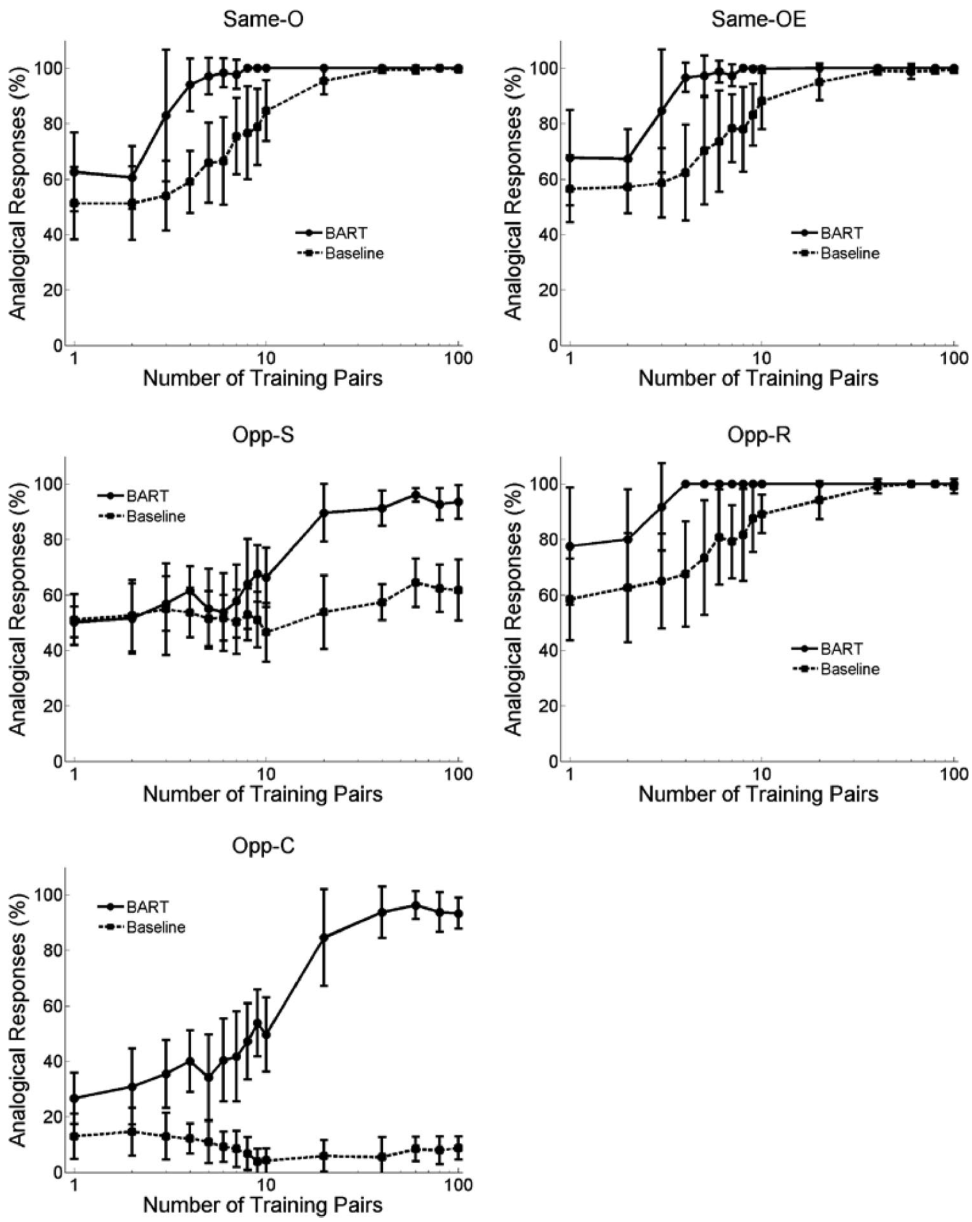

Figure 11. Proportion of analogical responses as a function of the number of training examples (log scale) with Leuven inputs for the five types of analogy problems. Solid lines present results for Bayesian analogy with relational transformations (BART) with empirical prior and hyperprior; dashed lines present results for baseline model (uninformative prior). Error bars indicate 1 standard deviation (results based on 10 runs). Same–O = same–extreme (opposite as foil); Same–OE = same–extreme (opposite–extreme as foil); Opp–S = opposite (split pair as foil); Opp–R = opposite (reversed as foil); Opp–C = opposite (conflict as foil).

features that were transparently related to them, the more opaque topics vectors did not provide any dimensions that were transparently relevant for learning comparative relations. To make the results as comparable as possible to those obtained in the previous simulations, we used the same set of 77 animal names taken from the Holyoak and Mah (1981) norms that were used in the simulations based on rating inputs.

The topics representations we used for these words were derived from the output of the topics model based on the tasaALL data base (see Griffiths et al., 2007). This output consists of 24 samples with 300 topics each for 26,243 unique words. Each sample consists of a word-by-topic matrix, in which the entry in the *i*th row and *j*th column is the number of times that the *i*th word appeared in the corpus and was assigned to the *j*th topic. We performed several preprocessing operations to create the vectors used as the immediate inputs to BART. First, we added a smoothing parameter of 0.01 to each entry in the matrix. We then derived a feature vector for each word, in which each feature value corresponds to the conditional probability of a corresponding topic given that word. This value is simply the joint frequency of the topic and word (an entry in the matrix) divided by the frequency of that word (the sum of a row in the matrix). Although the frequencies of specific words for each topic vary somewhat across the 24 samples, it was clear from inspection that the same 300 topics appeared in the same order across all 24 samples, indicating that the topics solutions are robust. Accordingly, a single feature vector for each word was calculated simply by averaging its feature vectors across all samples. We multiplied the computed probability by 100 to make the magnitude range roughly comparable to that of the rating inputs.

If all 300 dimensions of each word vector were used, each word pair vector would have 600 dimensions. However, for any individual word, most feature values are close to 0 (reflecting the fact that most of the 300 topics are irrelevant for any particular word). Dimensions that yield feature values at or near 0 for all words of interest (the animal names) will be useless in subsequent relation learning, and are likely to introduce noise that will impede any learning algorithm given the sheer size of the search space and limited number of training examples. To focus on the most important dimensions (those for which animals names tend to have nonzero probabilities), we summed the feature vectors for all 77 animals and identified the 50 dimensions that yielded the largest sums. By using just these 50 most important topics dimensions to form vectors for each individual word, the total size of the vector for each word pair was reduced to 100 (the same dimensionality as for the Leuven vectors).

Because the feature values as described so far are based on probabilities, they are necessarily nonnegative. Accordingly, we subtracted from each feature the mean value of that feature across all 26,243 word vectors. The feature values in the vectors used as inputs to BART therefore included both negative and positive values, with means near 0.

Training. The basic training regimen was identical to that employed with the Leuven vectors (except all learning was based solely on animals from the Holyoak & Mah, 1981, norms). The same hyperprior parameters were used. (Hyperpriors improved generalization performance by about 3 percentage points overall, with more significant improvement for certain analogy tests.) For comparison, we again tested the same baseline model as that used with both rating and Leuven inputs (i.e., Bayesian logistic regression with standard normal distributions as uninformative priors).

Generalization performance.

Basic tests. All reported results are based on the average performance over 10 runs, each of which randomly selected a set of training pairs from the pool. Figure 12 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Not surprisingly, given the vastly greater opacity of topics representations, the level of performance was considerably lower overall than was obtained with the rating or Leuven vectors. However, the full BART model, with empirical priors on mean weights and a hyperprior on variances, achieved substantial generalization (about 70%–80% accuracy for the four "greater" relations after 100 training examples). These results indicate that even when magnitude information is not coded in any clear way in the inputs, BART can learn useful representations of comparative relations from positive examples. The baseline model showed much weaker generalization performance, starting at chance (50%) and peaking at a mean of 67% accuracy after about 80 training examples.

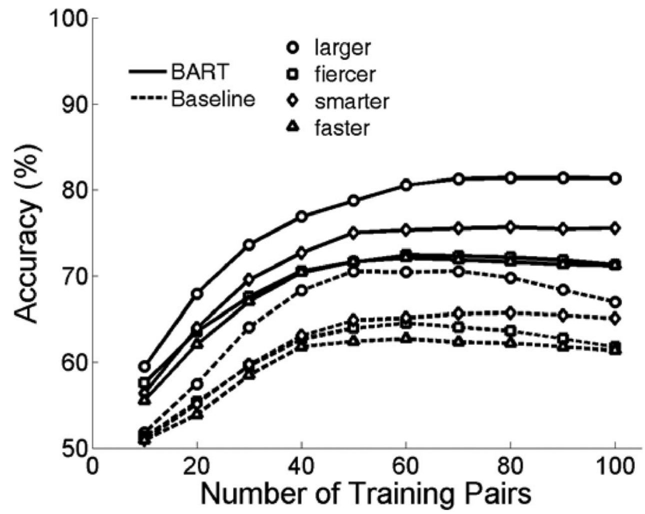

Figure 12. Accuracy in the generalization task with topics vectors as a function of the number of training examples for the four comparative relations. Solid lines indicate the performance of Bayesian analogy with relational transformations (BART) with the empirical prior and hyperprior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

We also explored how BART's generalization performance changed with more extended training. Whereas BART appeared to be unbiased when trained with up to 80–90 examples, a further increase in the number of training examples led to a bias toward "yes" responses. This type of response bias leads to reduced accuracy if a fixed decision criterion is used. (Note the slight decline in accuracy apparent in Figure 12 after 70 training examples, especially for the baseline model.) Generalization performance as measured by A_z continued to improve slightly with increased numbers of training examples. But even after 2,000 training examples, overall generalization performance as measured by A_z was higher for BART (.87) than for the baseline model (.82).

To examine whether the relational representations that BART derives from topics vectors yield the distance effect obtained for comparative judgments by humans, we again examined how BART's generalization performance relates to the rated subjective distance between each test pair of animals on the dimension of interest (as measured with the Holyoak & Mah, 1981, norms). Figure 13 plots the mean log ratio of predicted probabilities for positive versus negative test pairs as a function of distance on the relevant dimension between the two animals in a pair (after learning from 100 training examples), using the same distance bins as were used to test the model with rating vectors. The log ratio again increased monotonically with distance. Thus, the relational representations that BART acquired from topics inputs clearly yield a symbolic distance effect.

Generalization beyond the range of training examples. As in the case of the simulations based on ratings and Leuven vectors, we performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

1. Training with 100 examples from distance Bins 1–3 only and testing on all five distance bins. We obtained a monotonic increase in mean log ratio across all levels of distance: 0.78, 2.05,

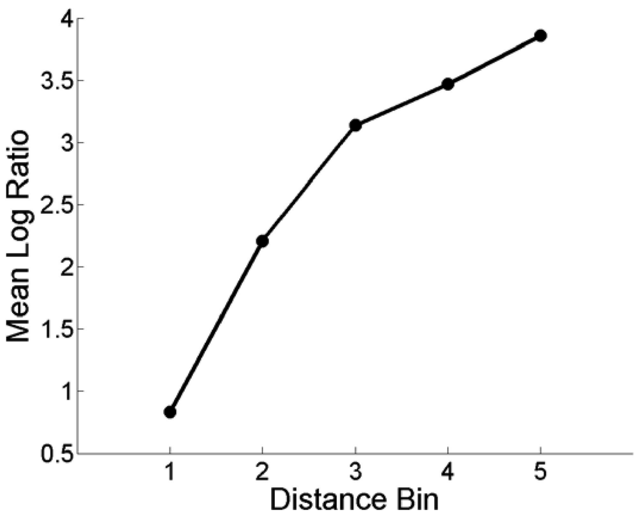

Figure 13. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (topics inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms (see Figure 6). Results are collapsed over the four continua.

2.86, 3.29, and 3.61 for Bins 1–5, respectively, extending the similar pattern obtained with ratings and Leuven inputs.

2. Training with 100 pairs of the form [4, 1] and testing on all pairs of the form [3, 2]. BART's accuracy levels were 75%, 45%, 54%, and 44% for *larger*, *fiercer*, *smarter*, and *faster*, respectively.

3. Standard training with 100 pairs formed from all groups and testing on all pairs of the form [3, 2]. BART's accuracy levels were 87%, 54%, 56%, and 54% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus for topics inputs, Tests 2 and 3 indicate that generalization performance was weak for close midrange pairs, especially when the range of the training pairs was restricted to those of the form [4, 1] (Test 2).

4. Training with 100 pairs of the form [2, 1] and testing on all pairs of the form [4, 3], or the reverse. Averaged across the two variations, generalization accuracy was 49%, 61%, 63%, and 62% respectively for *larger*, *fiercer*, *smarter*, and *faster*, indicating modest performance for the latter three relations.

5. No overlap between training and test animals with 100 training examples. BART achieved accuracies of 71%, 65%, 68%, and 63% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating moderately successful generalization to animals not encountered during training.

Content of learned weight distributions. We examined representative solutions that BART generated in learning representations for the comparative relations based on topics inputs. These solutions were even more distributed than those obtained with Leuven inputs (see Figure 10), with around 30 dimensions (i.e., about 60 weights) distinguishing the two roles for each comparative. The two roles were generally contrastive (i.e., weights associated with the two roles took on opposite signs). However, unlike those based on ratings and Leuven vectors, the topics solutions did not clearly distinguish the "greater" and "lesser" poles of individual continua. That is, rather than having predominantly high positive weights on the first role and high negative weights on the second, the pattern of weight polarity was more mixed. As we will show, the lack of a clear distinction between "greater" comparatives (*larger*, *fiercer*, etc.) and "lesser" ones (*smaller*, *meeker*, etc.) had negative consequences for BART's ability to solve some specific types of analogy problems based on topics inputs.

In addition to being more highly distributed, the topics solutions proved to be much more opaque than the Leuven solutions. Indeed, the term *topic* (which suggests an overall semantic theme) seems like a misnomer when applied to the feature dimensions that loaded highly for the various continua. Rather than having a clear semantic interpretation, each topic can really only be characterized by the list of words associated with it. For example, the topic most strongly predictive that the first object was larger than the second was highly associated with words for body parts (e.g., *blood*, *body*, *heart*, *cells*). Of course, this is only one of about 30 topics that collectively drove the decision as to which animal is larger. Thus BART was able to learn and generalize representations of comparatives from topics inputs with moderate success, even though the underlying features were subsymbolic.

Analogy performance. Given that the topics representations for comparatives were highly distributed and semantically opaque, it is not surprising that using them to solve higher order analogy problems proved to be challenging. We tested BART and the baseline model on the five types of analogy problems in the same manner as for the ratings and Leuven inputs. The results for up to 300 training examples are shown in Figure 14. The overall level of performance is lower than was obtained when the models were trained with ratings or Leuven inputs. Nonetheless, performance of the BART model on the analogy tests after learning from topics inputs was quite good. After 300 training examples, the performance level of BART was essentially perfect for Same–O problems, at about 90% accuracy for Opp–S problems, and 80% for Opp–C problems. BART performed less well on the Opp–R problems (about 70%) and the Same–OE problems (about 60%).

BART's lower performance on the latter two types of problems reflects that fact that in each case the correct answer can only be discriminated from the foil on the basis of polarity. Thus in the Same–OE type, the foil is a pair of relations at the opposite pole from the A:B pair, and in the Opp–R type the foil has the polarity reversed relative to A:B (e.g., if *larger*:*smaller* is the A:B term and *fiercer*:*meeker* is the correct C:D choice, the foil might be *meeker*:*fiercer*). As discussed earlier, for any pair of polar opposites BART implicitly identifies the "greater" relation as that for which the first role has predominantly positive weights on the relevant dimensions, whereas the "lesser" relation is that for which the first role has predominantly negative weights. As noted above, the topics inputs were much less clear than the rating or Leuven inputs, providing a mix of dimensions that were positively and negatively weighted as indicants of, for example, *larger* and *smaller*. In other words, topics inputs did not clearly establish "which end is up" for the various continua, making it difficult for BART to use polarity information as its sole basis for selecting an analogical completion.

Somewhat paradoxically, the baseline model actually was more accurate than BART for the problem types where polarity information was critical (Same–OE and Opp–R). The reason is that the baseline model (without empirical priors or hyperpriors) estimated

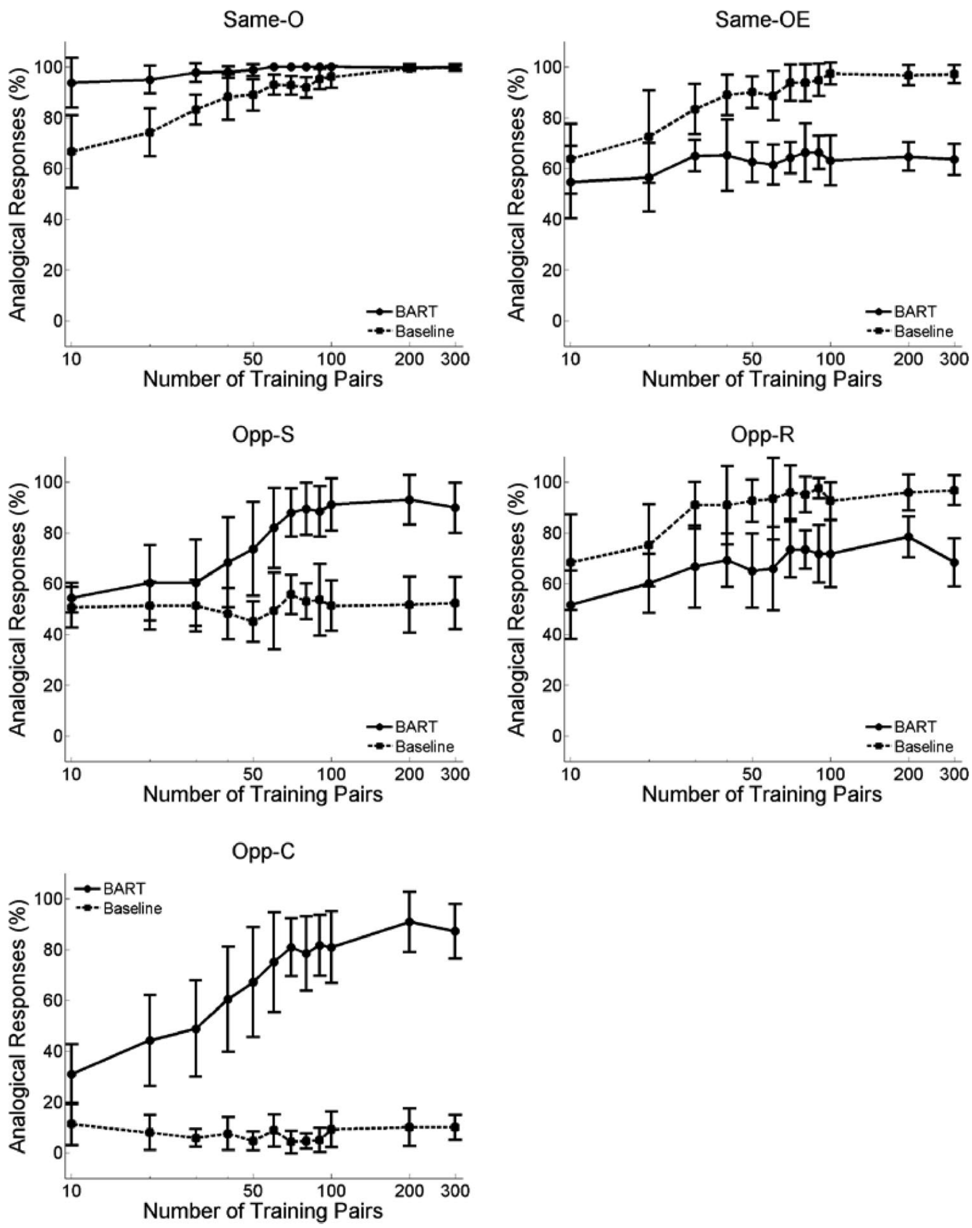

Figure 14. Proportion of analogical responses as a function of the number of training examples with topics input for five types of analogy problems. Solid lines present results for Bayesian analogy with relational transformations (BART) with empirical prior and hyperprior; dashed lines present results for baseline model (uninformative prior). Error bars indicate 1 standard deviation (results based on 10 runs). Same–O = same–extreme (opposite as foil); Same–OE = same–extreme (opposite–extreme as foil); Opp–S = opposite (split pair as foil); Opp–R = opposite (reversed as foil); Opp–C = opposite (conflict as foil).

higher variances on weights, which increased sensitivity to the small difference between the same-pole correct choice and the opposite-pole foil. Intuitively, BART "knew too much," viewing, for example, a pair like *meeker–fiercer* as a competitive foil to *fiercer–meeker* when seeking a match to *larger–smaller*, since all these pairs instantiate contrasting relations. By contrast, the baseline model could detect no apparent relationship between *meeker–fiercer* and *larger–smaller*, so was more likely to favor the correct option as the analogical completion.

More broadly, however, the performance of the baseline model across the entire set of analogy tests was dismal (see Figure 14). As was the case for the Leuven inputs, the baseline model failed completely on the Opp–S problem type (near chance level of 50%) and the Opp–C type (near chance level of 0%). Even after 2,000 examples, performance of the baseline model lagged 33 percentage points behind BART on Opp–S problems and 90 percentage points on Opp–C problems. Thus even though the baseline model achieved modest success in relational generalization with both

Leuven and topics inputs, it was unable to use its learned representations to reliably solve higher order analogy problems.

It may seem surprising that priors continued to play a critical role in analogy performance even after 2,000 trials, as the general rule for Bayesian models is that priors are eventually swamped by data. However, the fact that learning in our simulations was based solely on positive examples may have made priors especially potent. Although the model was expected to learn a specific comparative relation, such as *larger*, a finite set of positive examples is likely to be consistent with multiple possible relations (e.g., *both animals*, *both physical objects*). The contrastive priors provided BART with a strong push in the direction of comparative relations, whereas the baseline model, with its uninformative priors, might sometimes have acquired weights consistent with other possible relations exhibited by the positive training examples. Consequently, the patterns of weight distributions acquired by the baseline model likely were more variable from one comparative to another, impairing its performance on higher order analogy problems.

General Discussion

Summary

The work described in this article addresses the issue of whether and how relational representations that can support structured reasoning may be learnable from nonrelational inputs. Using comparative relations as a model domain, we performed a series of computational experiments based on BART, a Bayesian model of relation learning and analogical inference. To relate our findings to the general manner in which children appear to acquire concepts, we focused on learning from labeled positive examples of each target relation. BART incorporates a key representational assumption: A relation is represented by a weight distribution that serves to assess the probability that a pair of objects exemplifies it. BART proceeds in two basic stages. First, guided by empirical priors on mean weights and hyperprior on variances, the model uses Bayesian inference to update its weight distribution based on training examples. Second, the model uses importance-guided mapping to transform its learned weight distributions and then calculate the distance between pairs of relations, thereby assessing the validity of higher order analogies based on the implicit relations "both same" (*higher*:*fiercer*::*smarter*:*faster*) and "opposite" (*higher*:*lower*::*fiercer*:*meeker*).

When trained and tested with items based on small object vectors derived from human ratings of subjective magnitudes, BART achieved near-perfect accuracy both in generalizing to new examples of relations and in assessing higher order analogies based on its acquired relations. The high-dimensional and more complex Leuven and topics vectors posed a far greater computational challenge, as no invariant features are apparent, and learning depends on acquiring distributed representations over dozens of feature dimensions. When Leuven vectors were used as inputs, generalization accuracy was in the range of 90% accuracy after 100 training examples; with topics vectors, accuracy was in the range of 70%–80%. Thus BART showed substantial generalization ability after learning from the more complex inputs. Moreover, for all three types of inputs, BART was able to generalize to a completely new set of animals than those used in training.

When tested on higher order analogy problems based on the relations "same–extreme" and "opposite," the algorithm for importance-guided mapping yielded near-perfect performance via BART's learned relational representations, for both ratings and Leuven inputs. BART's analogy performance was also quite strong for topics inputs (except that topics inputs did not provide information that could clearly distinguish the "greater" versus "lesser" pole of each magnitude continua). In contrast, a baseline model with uninformative priors showed substantially weaker generalization and analogy performance for both Leuven and topics vectors (even though it was provided with the identical algorithm for importance-guided mapping).

For all three types of inputs, the relational representations learned by BART provided a qualitative account of the symbolic distance effect (Moyer, 1973). The degree of difficulty of making relative judgments on a dimension (as indexed by the model's estimates of log posterior probability ratio for a target relation) varied inversely with the magnitude difference between the two items in a pair. In addition, BART demonstrated the capacity to generalize outside the range of magnitude distances provided in the training set. Even when trained on animal pairs exhibiting small or medium size differences, the model was most confident when generalizing to novel pairs exhibiting large size differences. The model is thus consistent with evidence that people can distinguish "ideal" from "most typical" exemplars (Kittur, Hummel, & Holyoak, 2006). BART in effect defines the ideal exemplar of a *larger* relation as the pair with the largest size difference (e.g., "a dinosaur is larger than a flea"), even though a pair like "a fox is larger than a dove" would be more typical of the observed instances (i.e., closer to their central tendency).

Overall, the simulations reported here thus show that BART was able to pass five critical tests that can be posed for any model of relation learning. We have shown that the model can (a) learn first-order relations from complex nonrelational inputs that were independently generated; (b) learn with high efficiency; (c) generalize to classify novel relational examples; (d) capture a major factor (symbolic distance) that affects difficulty of human comparative judgments; and (e) use its learned relational representations to solve higher order analogy problems. No previous model of relation learning has met all these criteria.

Comparison With Previous Approaches

There have been many previous computational models of various aspects of analogical reasoning, which have been classified as symbolic, connectionist, and various hybrids (French, 2002; Gentner & Forbus, 2011). BART's capabilities appear painfully limited when compared to those of the state-of-the-art analogy models. To date, its most advanced accomplishment is to solve simple four-term analogy problems, whereas other models can perform much more complex feats involving analog retrieval, mapping, inference, and schema formation. Even in the restricted domain of four-term analogy problems, BART cannot compete with state-of-the-art machine-learning models (e.g., Turney, 2006).

However, BART is focusing on very different issues than those addressed by most previous analogy models. It attempts to answer the basic question, How might relational representations be created? The operation of BART provides a well-specified computational model of the type of relational rerepresentation that seems to underlie the power of human thinking (Penn et al., 2008). More generally, the model provides a concrete example of how new representations can be acquired by forms of inductive bootstrapping: first the use of empirical priors to jump-start relation learning, and then the use of importance-guided mapping to transform and compare relational representations. BART illustrates the centrality of analogical bootstrapping in learning relations (cf. Carey, 2011; Gentner, 2010).

The idea that functionally defined importance provides a key pragmatic constraint on mapping has a long history (Holyoak, 1985), and BART exemplifies a very basic mechanism by which importance can be defined quantitatively and used to place nonidentical dimensions into correspondence. Unlike previous computational models of mapping, BART finds mappings between features at a subsymbolic level (identifying systematic correspondences between distributed patterns in a high-dimensional weight space), rather than between explicit predicates. Subsymbolic mapping processes of this sort may underlie various types of implicit analogical transfer (e.g., Day & Goldstone, 2011).

The present findings provide a proof-of-concept that, for the domain of comparative relations, a capacity for structured relational reasoning can potentially emerge from bottom-up learning based on unstructured inputs. Particularly, in the case of the topics vectors, the inputs were created without any human guidance that might have tailored them to the relational learning task. If we consider the operation of the topics model itself (Griffiths et al., 2007) in conjunction with BART, the representations that the latter model used to assess structured analogy problems can be traced back to the raw statistics of covariation among words in texts. The to-be-learned relations did not correspond to any specific dimensions created by the topics model. In broad strokes, we have shown that BART can solve structured analogy problems, albeit simple ones, using relational representations the model learned itself from unstructured inputs that were independently generated. No previous model of analogy has demonstrated a comparable capability, which arguably is the essential precursor to a satisfying account of complex human analogical reasoning.

The most appropriate comparisons for BART are with previous models of relation learning. As a discriminative Bayesian model, BART is most directly related to the model developed by Silva, Airoldi, and Heller (2007; Silva, Heller, & Ghahramani, 2007; see also D. Chen et al., 2010). Like the model of Silva and colleagues, BART uses empirical priors to bootstrap relation learning; however, in BART the empirical priors are relation specific, and can be used together with a hyperprior on variances. BART is able to use training examples to automatically select the one-place predicate best suited for generation of priors on mean weights for a new comparative relation. The algorithm for importance-guided mapping is a key innovation that enables BART to move beyond relational generalization to the more challenging task of solving higher order analogy problems based on its learned representations.

It is instructive to relate the core concepts and mechanisms instantiated in BART to those that underlie other approaches to relation learning. We will focus on three central aspects of the BART model. These are (a) exploiting empirical priors, (b) representing relations as weight distributions, and (c) allowing role-dependent operations on representations. With these in mind, we can draw comparisons and contrasts with three general approaches reviewed earlier.

Comparison to hierarchical generative models. If we take the structure-learning model of Kemp and Tenenbaum (2008) as an example, the generative approach is broadly similar to BART in the use of statistical learning over distributions. The two approaches also converge in denying that explicit representations of relations could be acquired by a complete tabula rasa. However, the models make different assumptions about what knowledge or capacities the learner brings to the task. Kemp and Tenenbaum's model is endowed with a grammar that can generate candidate structures over which statistical learning can be applied. BART does not come equipped with a comparable grammar of relations. Rather, it comes with a suite of operations that it can use to create and transform representations (in particular, the ability to select and use empirical priors to initialize the representation of a to-be-learned relation, and the ability to perform importance-guided mapping and subsequent relational transformations).

The generative and discriminative approaches to learning relations may well prove to be complementary. An important question for future research is whether a relational representation of the type acquired by BART might be transformed into a generative model, a step likely to be necessary in order to achieve the full range of human-like relational capabilities.

Comparison to neural network models. Both generative and discriminative Bayesian models are similar to neural network models in their emphasis on statistical learning as a major contributor to the acquisition of knowledge. Discriminative models such as BART are perhaps somewhat closer to the spirit of neural network models, emphasizing the emergence of knowledge from bottom-up processing of data provided by the environment. However, the weight distributions over feature vectors that BART uses to code relations capture more information than do the representations created by typical neural network models. Weight distributions code not only first-order statistics (means) but also second-order statistics (variances and covariances) that capture uncertainty about weights and interweight correlations (a property shared by "deep learning nets"; Salakhutdinov & Hinton, 2009).

Yet, paradoxically, BART's relational representations are also explicit and structured in ways that representations in a distributed neural network are not. Most basically, BART's weight distributions respect the integrity of distinct roles (e.g., the roles of the larger versus smaller member of a pair of objects). The internal structure of relations in BART is presumably inherited from the output of the perceptual system, which codes objects as individuals. A relation such as *larger* is learned from pairs of objects, and hence is structured as a pair of roles. In contrast to the model of Rogers and McClelland (2008), for example, an individual relation in BART has a distinct identity (e.g., the weight distribution for *larger* is different from that for *fiercer*, and also from the complementary relation *smaller*). Because relational representations in BART are isolable from one another, they can be compared and systematically transformed. The properties of the weight distributions employed by BART are thus qualitatively and quantitatively different from those of the weight matrices used in classical neural networks.

Comparison to symbolic connectionist models. Among the algorithmic models of relation learning, BART has most in common with the class of symbolic connectionist models, such as

LISA (Hummel & Holyoak, 1997, 2003) and DORA (Doumas et al., 2008; see also Halford et al., 1998). These models, like BART, assume that relational representations are structured in terms of roles, thereby escaping the fundamental limitations of conventional neural network models. Another important similarity is that BART, like DORA, exploits the potential for bootstrapping from initial learning of one-place predicates to learning comparative relations.

In general terms, both DORA and BART aim to learn relations using bottom-up mechanisms based on detection of covariation among the objects that fill relational roles. DORA emphasizes learning from unlabeled examples, whereas BART focuses on learning from labeled examples (either positive or negative, although the former are assumed to be more common in child language acquisition). DORA, extending earlier proposals concerning schema induction (Gick & Holyoak, 1983), learns relations by taking the intersection of the feature representations of multiple examples. In comparison to a regression algorithm of the sort used by BART, the logical intersection operator appears to be too strict (a single exception may cause a feature to be dropped from the representation of a relation). The method is ill-suited for learning concepts defined by distributed representations over features that are only probabilistically predictive, as was the case for Leuven and topics vectors. By using Bayesian inference, BART is able to learn probabilistic representations of relations from positive examples, without requiring any strictly invariant features, while simultaneously factoring in the influence of prior knowledge. As noted earlier, DORA has only been tested on hand-coded inputs that include invariant features over which to-be-learned relations can be defined. DORA (like LISA) is not designed to map nonidentical features to one another, as the mapping algorithm is restricted to mapping predicates, rather than object features.

Most basically, BART's representational scheme for relations distinguishes it from symbolic as well as nonsymbolic connectionist models. In both varieties of neural network models, relations and objects have generally been represented in a common format (as distributed sets of units interconnected by weighted associations). In symbolic-connectionist models, which focus on explicit relational representations, both relations and objects have been represented in terms of units for semantic features: either separate pools of feature units for the two types of entities (LISA; Hummel & Holyoak, 1997, 2003; see also Halford et al., 1998) or a single pool (Doumas et al., 2008). BART introduces a very different representational assumption. Whereas objects are represented by a vector of features, first-order relations are represented as weight distributions. In BART, relations (weight distributions) and objects (feature vectors) constitute distinct but connected representational elements. This representational distinction is critical for BART's ability to acquire relational representations by statistical learning.

BART's representational assumptions may suggest an important way in which algorithmic symbolic-connectionist models can be refined. The use of temporal synchrony for role binding gives rise to inherent capacity limits, related to the number of distinct temporal phases that can be interleaved without significant overlap of firing for each phase. Given established limits on neural firing rates, this "relational bottleneck" has been estimated at four to six distinct phases (Cowan, 2001; Hummel & Holyoak, 1997). If role bindings are coded by synchronizing the neural code for a role and its filler, as the LISA model assumes, then this limit translates directly into four to six concurrently active role bindings, or two to three complete propositions, a number that appears plausible for adult humans. But as noted earlier, models that use temporal firing patterns as a dynamic code for role bindings in active memory can only synchronize representations that can be kept distinct despite firing together. LISA's use of neural synchrony therefore depends on defining separate pools of features for objects and relations.

In order to model the learning of features of relations from features of objects, the DORA model (Doumas et al., 2008) assumes instead that relations and objects are defined over a single pool of semantic features and that role bindings are coded by asynchrony of firing for a role and its filler. This shift in representational assumptions means that DORA requires twice as many distinct temporal phases as LISA to represent the same number of role bindings. In effect, DORA's estimate of the capacity of human working memory is half the value predicted by LISA. Doumas et al. (2008, pp. 30–31) suggest that asynchrony may be required only for relation learning and not for relational inference. However, it is unclear how inferences could be made reliably if roles and their fillers were coded on the same pool of features and yet allowed to fire in synchrony (e.g., the distinction between "elephants are big" and "elephants are gray" would seem to be lost).

BART's assumption that relations (more generally, predicates) and objects rely on distinct types of representations (weight distributions versus feature vectors) goes between the horns of this dilemma, providing a potential basis for a LISA-like system of binding by synchrony that nonetheless is capable of relation learning. That is, the dynamic form of coding a role binding might involve the synchronous activation of a role (a distinct subset of the weight distribution for a relation) and the feature vector for its filler. The role (weight distribution) and its filler (feature vector) would not be confusable even when synchronized, because each would constitute a distinct representational type. An algorithmic implementation based on BART's form of relation representation would thus yield the same estimate for the capacity of working memory as does LISA.

Potential Extensions

Acquiring more detailed developmental data. For the present project, we created a microworld in which a learner (the BART model and variations on it) must learn several comparative relations defined over a set of animal concepts, using inputs consisting of feature vectors, and then must draw higher order analogies based on the acquired relational representations. In general terms, we constrain the task in ways that seem consistent with comparable relation learning by children (modest numbers of largely positive examples, acquiring one-place predicates prior to true relations). But we acknowledge that our microworld is not the one that children actually encounter. Children do not learn *larger* and other comparative relations from animals only, and we lack detailed knowledge of the inputs children actually have available. At most, realistic inputs resemble those that BART receives in that they are also based (at least in part) on sets (likely quite large) of features associated with individual objects.

Because of the idealized nature of our microworld, empirical assessment of the models was largely qualitative (which is rather counterintuitive, since BART generates detailed learning curves). We hope that future tests of computational models of relation learning can be informed by more detailed empirical evidence regarding the inputs children use to acquire relations, the trajectory of children's learning for specific relational concepts, and the linkage between relational generalization and the ability to reason by analogy.

Extensions to richer inputs. As we emphasized at the outset, the representations that serve as inputs to children learning relations are undoubtedly richer than those we provided to BART in the present set of simulations. Children learn from more direct perceptual experience, including motoric feedback from their own actions. For example, Maouene et al. (2008) showed that the age of acquisition for basic English verbs (e.g., *kiss*, *hug*, *kick*) is related to the nature of their association (for adults) with body parts (e.g., the mouth versus the hands and arms). As another example, work on action recognition has identified certain "signature movements," such as a punch, that have a special status in rapid identification of types of threatening actions (van Boxtel & Lu, 2011, 2012). Such cues (in conjunction with adult speech) very likely provide a significant part of the inputs available to children as they learn verbs corresponding to basic actions. Realistic inputs are likely to involve greater structure than the "flat" vectors used in the present article, including various types of higher order features (Regier, 1996; Regier & Carlson, 2001). Future research should explore the use of learning algorithms that can create and exploit hierarchical structure in their inputs.

Role of empirical priors in relation learning. The simulations reported here demonstrate that representations of one-place predicates can provide very useful empirical priors to facilitate learning of the corresponding two-place relations. Knowledge about a one-place predicate such as *large* can be learned from a set of single objects (e.g., *elephant*), whereas learning the relation *larger* requires joint processing of pairs of objects (e.g., *elephant* and *bear*). On the basis of Halford's (1993; Halford et al., 2010) assumption that capacity increases over the course of cognitive and neural development, and that attending to two objects requires greater capacity than attending to one, it follows that children will tend to learn one-place predicates prior to multiplace relations (which have at least two roles), in accord with developmental evidence (Smith, 1989).

However, this developmental pattern does not necessarily imply that learning specific one-place predicates (e.g., *large*, *small*) is a strict prerequisite for learning a related two-place predicate (e.g., *larger*). At least for ratings and Leuven vectors, the baseline model with uninformative priors was able to learn comparative relations and achieve substantial generalization performance when given an adequate number of training examples. As long as we assume sufficient working memory to hold two items, BART can proceed to learn a two-place predicate directly, regardless of whether it has already acquired corresponding one-place predicates. It is an open empirical question whether children necessarily learn one-place relative adjectives as a prerequisite to learning two-place comparative adjectives (cf. Halford et al., 2010). More generally, however, many multiplace predicates (e.g., *opposite*) do not seem to naturally decompose into simpler one-place predicates.

The further exploration of empirical priors will be especially important in attempting to extend the current approach to other types of relations besides comparatives (see Jurgens, Mohammad, Turney & Holyoak, 2012). As the pool of potential empirical priors grows larger and more varied, more sophisticated algorithms for prior selection may prove useful. For example, prior selection may involve a hierarchical process, winnowing options based on general types of relations (e.g., varieties of sameness versus contrast).

It should be emphasized that the concept of empirical priors is considerably more general than the idea of using one-place predicates to guide learning of related two-place relations. Relation learning can also potentially be bootstrapped by previously learned relations (e.g., a perceptually based comparative such as *larger* might facilitate subsequent acquisition of a more abstract comparative such as *smarter*). Yet, more generally, the entire process of analogical reasoning can be viewed as a sophisticated use of empirical priors, in which the source analog is used to impose priors to guide learning about the target (Holyoak, Lee, & Lu, 2010).

Learning higher order relations. Although the present version of BART does not create explicit representations of higher order relations such as *opposite*, it does appear to set the stage for this possibility. In evaluating higher order analogies, the model is implicitly sensitive to whether A:B and C:D both instantiate some version of *opposite*. By assessing the distance between transformed weight distributions, BART shows how representations of different relations can be compared with one another. To create explicit higher order representations, an extension of the model could treat these transformed weight distributions in a manner analogous to feature vectors, recursively applying its statistical learning procedures to acquire higher order weights that capture the commonalities between pairs of first-order relations such as *larger*:*smaller* and *fiercer*:*meeker* (i.e., a representation of *opposite*). In moving from first-order to higher order relations in this manner, an extension of BART would in effect rerepresent first-order relations as derived feature vectors, which can then serve as inputs to a learning process that yields representations of higher order relations. This basic move—treating learned weights as derived features—provides a potential avenue to allow the development of hierarchical relational systems. It also provides a possible answer to the inductive puzzle we raised at the outset: How does the mind acquire concepts that cannot be defined in terms of features bound to perception?

Toward an algorithmic model. Although we have developed BART as a computational-level model of how relations might be learned and transformed to solve higher order analogy problems, it should be possible to incorporate the basic ideas into algorithmic models. As suggested above, the fact that BART creates separate (but linked) representations for relations and their fillers is compatible with synchrony-based models of the symbolic-connectionist variety (Hummel & Holyoak, 1997, 2003). More generally, it is useful to distinguish those aspects of BART that depend on role-governed operations from those that do not. Importantly, the inductive process that updates weight distributions based on training examples is not directly dependent on roles. The feature vectors associated with the two objects being compared are simply concatenated. BART's weight distributions can be viewed as a type of "attention weights" that reflect the importance of each dimension for accurate classification of relations. Learning models based on the idea of attention weights have been applied to object categorization and perceptual learning (Nosofsky, 1985; Petrov, Dosher, & Lu, 2005).

BART would require a more complex learning algorithm to acquire distributions of weights (rather than simply mean weights) based on supervised learning (cf. Salakhutdinov & Hinton, 2009). A psychologically realistic learning model would have to accommodate sequential training inputs. Although the version of BART we have described operates on all training data at once, we have in fact also implemented a variant that uses sequential updating. (In general, regression models can operate in either batch or sequential fashion.) The sequential version produces very similar results after roughly 100 training examples. Thus, although a full model of sequential learning would require additional theoretical work, there is reason to be optimistic that such a model is possible (see Lu, Rojas, Beckers, & Yuille, 2008, for a sequential model of causal learning). For example, it is conceivable that the brain in effect implements some kind of variational method based on tacit assumptions about the form of neural distributions.

Although the core learning model (updating of weight distributions) is not role governed, BART does operate on roles (a) to establish empirical priors that guide acquisition of relations and (b) to perform importance-guided mapping based on the learned representations of relations. These operations depend on the manipulation of structured knowledge, a capacity that is arguably specific to humans (Penn et al., 2008). Interestingly, neither of these operations appears to depend on the full covariance matrix for weight distributions. Rather, the mean weights (MAP estimates) may suffice (see Footnote 4). At a neural level, it is more plausible that summary statistics such as MAP estimates could be transmitted to downstream brain regions, rather than the full covariation matrix. It seems plausible that early neural areas are sensitive to intercorrelations among neural firing patterns, which encode covariance information (Aertsen, Gerstein, Habib, & Palm, 1989; Cohen & Kohn, 2011; Cohen & Maunsell, 2009; Kohn & Smith, 2005; Nirenberg & Latham, 2003), whereas higher level areas instead respond to broader temporal patterning, such as synchrony (Siegel, Donner, & Engel, 2012; Uhlhaas & Singer, 2010).

The operations of BART can thus be viewed as demarcating major points along an evolutionary continuum in relational processing and representation. The basic capacity to code approximate magnitudes so as to enable comparative judgments is common across many species. Some primates, including the rhesus monkey, have a limited ability to attach arbitrary symbols to small magnitudes (Diester & Nieder, 2007), and can also learn alternative first-order relations defined over a common continuum (e.g., selecting the larger or else the smaller of two numerosities in response to a discriminative cue; Cantlon & Brannon, 2005). Roughly, these species-general capabilities correspond to the modest ability of our baseline model, starting with uninformative priors, to learn weight distributions that support comparative judgments, allowing generalization to novel pairs.

However, the capacity to learn weight distributions only sets the stage for acquiring explicit relational representations. BART has the additional capacity to treat weight distributions as structured representations with multiple roles. These explicit representations of first-order relations can then be made available to symbolic processes capable of comparison and rudimentary analogical mapping, thereby enabling a variety of bootstrapping operations. BART uses roles to guide selection of empirical priors, thereby greatly increasing the efficiency of relation learning. After first-order relations have been acquired, BART is able to make structured analogical inferences by mapping dimensions based on their functional influence on relation discrimination, as opposed to their literal identity. An extension of the model that treats weights as derived features could potentially go on to discover the higher order commonalities shared by first-order relations defined over different dimensions, thereby acquiring explicit representations of higher order relations such as *opposite*. These symbolic capabilities, perhaps specific to humans, may depend on multiple subregions of the prefrontal cortex (particularly the rostrolateral portion; for a review, see Knowlton & Holyoak, 2009). The capacity for role-governed operations may thus represent a late evolutionary development that has allowed humans to attain their unique capacity for abstract thought.

References

Aertsen, A. M., Gerstein, G. L., Habib, M. K., & Palm, G. (1989). Dynamics of neuronal firing correlation: Modulation of "effective connectivity." *Journal of Neurophysiology, 61,* 900–917.

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

Banks, W. P., White, H., Sturgill, W., & Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgments. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 560–582. doi:10.1037/0096-1523.9.4.560

Bao, J., & Munro, P. (2006). Structural mapping with identical elements neural network. In *Proceedings of the 2006 International Joint Conference on Neural Networks* (pp. 870–874). Piscataway, NJ: IEEE Computational Intelligence Society. doi:10.1109/IJCNN.2006.246776

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice.* San Francisco, CA: Holden-Day.

Cantlon, J. F., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 16507–16511. doi:10.1073/pnas.0506463102

Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. A. (2006). Functional imaging of numerical processing in adults and 4-yr-old children. *PLoS Biology, 4,* e125. doi:10.1371/journal.pbio.0040125

Carey, S. (2011). Précis of *The Origin of Concepts. Behavioral and Brain Sciences, 34,* 113–124. doi:10.1017/S0140525X10000919

Chen, D., Lu, H., & Holyoak, K. J. (2010). Learning and generalization of abstract semantic relations: Preliminary investigations of Bayesian approaches. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32th Annual Conference of the Cognitive Science Society* (pp. 871–876). Austin, TX: Cognitive Science Society.

Chen, Z., Sanchez, R. P., & Campbell, T. (1997). From beyond to within their grasp: The rudiments of analogical problem solving in 10- and 13-month-olds. *Developmental Psychology, 33,* 790–801. doi:10.1037/0012-1649.33.5.790

Clark, H. H. (1970). The primitive nature of children's relational concepts. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 269–278). New York, NY: Wiley.

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience, 14,* 811–819. doi:10.1038/nn.2842

Cohen, M. R., & Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience, 12,* 1594–1600. doi:10.1038/nn.2439

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). New York, NY: Wiley.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi:10.1017/S0140525X01003922

Day, S. B., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 551–567. doi:10.1037/a0022333

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40,* 1030–1048. doi:10.3758/BRM.40.4.1030

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5,* 390–407. doi:10.1162/jocn.1993.5.4.390

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology, 5,* e294. doi:10.1371/journal.pbio.0050294

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology, 6,* 487–496. doi:10.1016/0022-2496(69)90019-4

Doumas, L. A. A., & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 52–66). New York, NY: 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. doi:10.1037/0033-295X.115.1.1

Eckstein, M. P., & Ahumada, A. J., Jr. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision, 2*(1), Article i. doi:10.1167/2.1.i

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48,* 71–99. doi:10.1016/0010-0277(93)90058-4

Flas, M., Lammertyn, J., Caessens, B., & Orban, G. A. (2007). Processing of abstract ordinal knowledge in the horizontal segment of the intraparietal sulcus. *Journal of Neuroscience, 27,* 8952–8956.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences, 6,* 200–205. doi:10.1016/S1364-6613(02)01882-X

French, R. M. (2008). Relational priming is to analogy-making as one-ball juggling is to seven-ball juggling. *Behavioral and Brain Sciences, 31,* 386–387. doi:10.1017/S0140525X0800455X

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.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, H., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *NeuroImage, 39,* 181–205. doi:10.1016/j.neuroimage.2007.08.013

Gallistel, C. R. (1993). A conceptual framework for the study of numerical estimation and arithmetic reasoning in animals. In S. T. Boysen & E. J. Capaldi (Eds.), *The development of numerical competence: Animal and human models* (pp. 211–223). Hillsdale, NJ: Erlbaum.

Gardner, H. (1974). Metaphors and modalities: How children project polar adjectives onto diverse domains. *Child Development, 45,* 84–91. doi:10.2307/1127753

Gasser, M., & Colunga, E. (2000). Babies, variables, and relational correlations. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 160–165). Mahwah, NJ: Erlbaum.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59,* 167–192. doi:10.1146/annurev.psych.58.110405.085632

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

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34,* 752–775. doi:10.1111/j.1551-6709.2010.01114.x

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development, 82,* 1173–1188. doi:10.1111/j.1467-8624.2011.01599.x

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *WIREs Cognitive Science, 2,* 266–276. doi:10.1002/wcs.105

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). London, England: Cambridge University Press. doi:10.1017/CBO9780511983689.008

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10,* 277–300. doi:10.1207/s15516709cog1003_2

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38. doi:10.1016/0010-0285(83)90002-6

Glass, A. L., Holyoak, K. J., & Kossan, N. E. (1977). Children's ability to detect semantic contradictions. *Child Development, 48,* 279–283. doi:10.2307/1128911

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 3–28. doi:10.1037/0278-7393.20.1.3

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review, 118,* 110–119. doi:10.1037/a0021336

Goswami, U. (1992). *Analogical reasoning in children.* Hillsdale, NJ: Erlbaum.

Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 437–470). Cambridge, MA: MIT Press.

Graziano-King, J., & Cairns, H. S. (2005). Acquisition of English comparative adjectives. *Journal of Child Language, 32,* 345–373. doi:10.1017/S0305000904006828

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14,* 357–364. doi:10.1016/j.tics.2010.05.004

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211–244. doi:10.1037/0033-295X.114.2.211

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17,* 767–773. doi:10.1111/j.1467-9280.2006.01780.x

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. doi:10.1037/a0017201

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

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,* 803–830. doi:10.1017/S0140525X98001769

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences, 14,* 497–505. doi:10.1016/j.tics.2010.08.005

Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology, 9,* 31–51. doi:10.1016/0010-0285(77)90003-2

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 59–87). New York, NY: Academic Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak

& R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234–259). New York, NY: Oxford University Press.

Holyoak, K. J., & Hummel, J. E. (2008). No way to start a space program: Associationism as a launch pad for analogical reasoning. *Behavioral and Brain Sciences, 31,* 388–389. doi:10.1017/S0140525X08004585

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development, 55,* 2042–2055. doi:10.2307/1129778

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. doi:10.1037/a0020488

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition, 9,* 197–204. doi:10.3758/BF03202335

Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior, 15,* 287–299. doi:10.1016/0022-5371(76)90026-8

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466. doi:10.1037/0033-295X.104.3.427

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264. doi:10.1037/0033-295X.110.2.220

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing, 10,* 25–37. doi:10.1023/A:1008932416310

Jani, N. G., & Levine, D. S. (2000). A neural network theory of proportional analogy-making. *Neural Networks, 13,* 149–183. doi:10.1016/S0893-6080(99)00106-9

Jones, S., & Murphy, M. L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics, 10,* 401–422. doi:10.1075/ijcl.10.3.06jon

Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)* (pp. 356–364). Montreal, Quebec, Canada: Association for Computational Linguistics.

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica, 133,* 216–233. doi:10.1016/j.actpsy.2009.11.012

Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 943–951). Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10,* 307–321. doi:10.1111/j.1467-7687.2007.00585.x

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 10687–10692. doi:10.1073/pnas.0802631105

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 1, pp. 381–388). Palo Alto, CA: AAAI Press.

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2006). Ideals aren't always typical: Dissociating goodness-of-exemplar from typicality judgments. In R. Son & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 429–434). Hillsdale, NJ: Erlbaum.

Knowlton, B. J., & Holyoak, K. J. (2009). Prefrontal substrate of human relational reasoning. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 1005–1017). Cambridge, MA: MIT Press.

Kohn, A., & Smith, M. A. (2005). Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *Journal of Neuroscience, 25,* 3661–3673. doi:10.1523/JNEUROSCI.5106-04.2005

Kurtz, K. J., Miao, C.-H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10,* 417–446. doi:10.1207/S15327809JLS1004new_2

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211

Lawrence, D. H., & DeRivera, J. (1954). Evidence for relational transposition. *Journal of Comparative and Physiological Psychology, 47,* 465–471. doi:10.1037/h0063235

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Relational priming: A developmental and computational perspective on the origins of a cognitive skill. *Behavioral and Brain Sciences, 31,* 357–378. doi:10.1017/S0140525X08004469

Link, S. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology, 34,* 2–41. doi:10.1016/0022-2496(90)90010-7

Lu, H., Lin, T., Lee, A., Vese, L., & Yuille, A. L. (2010). Functional form of motion priors in human motion perception. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 23, 1495–1503). Cambridge, MA: MIT Press.

Lu, H., & Liu, Z. (2006). Computing dynamic classification images from correlation maps. *Journal of Vision, 6*(4), Article 12. doi:10.1167/6.4.12

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 195–188). Austin, TX: Cognitive Science Society.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. doi:10.1037/a0013256

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation, 4,* 415–447. doi:10.1162/neco.1992.4.3.415

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge, England: Cambridge University Press.

Maouene, J., Hidaka, S., & Smith, L. B. (2008). Body parts and early-learned verbs. *Cognitive Science, 32,* 1200–1216. doi:10.1080/03640210802019997

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431–467. doi:10.1006/cogp.1993.1011

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Freeman.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348–356. doi:10.1016/j.tics.2010.06.002

McGonigle, B., & Chalmers, M. (1984). The selective impact of question form and input mode on the symbolic distance effect in children. *Journal of Experimental Child Psychology, 37,* 525–554. doi:10.1016/0022-0965(84)90075-4

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science, 14,* 516–529. doi:10.1111/j.1467-7687.2010.00999.x

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics, 13,* 180–184. doi:10.3758/BF03214124

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8,* 228–246. doi:10.1016/0010-0285(76)90025-6

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215,* 1519–1520. doi:10.1038/2151519a0

Neal, R. M. (1996). *Bayesian learning for neural networks.* New York, NY: Springer-Verlag.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14,* 11–28. doi:10.1207/s15516709cog1401_2

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron, 37,* 149–157. doi:10.1016/S0896-6273(02)01144-3

Nirenberg, S., & Latham, P. E. (2003). Decoding neuronal spike trains: How important are correlations? *Proceedings of the National Academy of Sciences of the United States of America, 100,* 7348–7353. doi:10.1073/pnas.1131895100

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

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55,* 169–195. doi:10.1016/j.cogpsych.2006.09.002

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. doi:10.1017/S0140525X08003543

Petrov, A. A. (2008). Relational priming plays a supporting but not leading role in adult analogy-making. *Behavioral and Brain Sciences, 31,* 392–393. doi:10.1017/S0140525X08004627

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review, 112,* 715–743. doi:10.1037/0033-295X.112.4.715

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44,* 547–555. doi:10.1016/j.neuron.2004.10.014

Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research, 1106,* 177–188. doi:10.1016/j.brainres.2006.05.104

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53,* 293–305. doi:10.1016/j.neuron.2006.11.022

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron, 41,* 983–993. doi:10.1016/S0896-6273(04)00107-2

Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works.* New York, NY: Oxford University Press.

Quesada, J., Kintsch, W., & Mangalath, P. (2004). Analogy-making as predication using relational information and LSA vectors. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (p. 1623). Austin, TX: Cognitive Science Society.

Ramscar, M., & Yarlett, D. (2003). Semantic grounding in models of analogy: An environmental approach. *Cognitive Science, 27,* 41–71. doi:10.1207/s15516709cog2701_2

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception. *Journal of Experimental Psychology: General, 130,* 273–298. doi:10.1037/0096-3445.130.2.273

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94,* 249–273. doi:10.1016/j.jecp.2006.02.002

Rogers, T. T., & McClelland, J. L. (2008). Précis of *Semantic Cognition: A Parallel Distributed Processing Approach. Behavioral and Brain Sciences, 31,* 689–714. doi:10.1017/S0140525X0800589X

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72,* 67–109. doi:10.1016/S0010-0277(99)00031-1

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron, 46,* 945–956. doi:10.1016/j.neuron.2005.05.021

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. In D. van Dyk & M. Welling (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (pp. 448–455). Cambridge, MA: Journal of Machine Learning Research.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition, 120,* 1–25. doi:10.1016/j.cognition.2011.02.010

Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience, 13,* 121–134. doi:10.1038/nrn3137

Silva, R., Airoldi, E., & Heller, K. A. (2007). *Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning* (Tech. Rep. GCNU TR 2007-001). London, England: University College London, Gatsby Computational Neuroscience Unit.

Silva, R., Heller, K. A., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (pp. 500–507). Cambridge, MA: Journal of Machine Learning Research.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24,* 1193–1216. doi:10.1146/annurev.neuro.24.1.1193

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. 146–178). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511529863.008

Smith, L. B., Rattermann, M. J., & Sera, M. (1988). "Higher" and "lower": Comparative interpretations by children. *Cognitive Development, 3,* 341–357. doi:10.1016/0885-2014(88)90020-2

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24,* 99–142. doi:10.1016/0010-0285(92)90004-L

Taylor, E. G., & Hummel, J. E. (2009). Finding similarity in a model of relational reasoning. *Cognitive Systems Research, 10,* 229–239. doi:10.1016/j.cogsys.2008.09.004

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. doi:10.1126/science.1192788

Tunteler, E., & Resing, W. C. M. (2002). Spontaneous analogical transfer in 4-year-olds: A microgenetic study. *Journal of Experimental Child Psychology, 83,* 149–166. doi:10.1016/S0022-0965(02)00125-X

Tunteler, E., & Resing, W. C. M. (2004). Age differences in patterns of spontaneous production of strategies for analogy problems among five- to eight-year-old children. *Educational and Child Psychology, 21,* 74–88.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics, 32,* 379–416. doi:10.1162/coli.2006.32.3.379

Turney, P. D., & Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning, 60,* 251–278. doi:10.1007/s10994-005-0913-1

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37,* 141–188. doi:10.1613/jair.2934

Uhlhaas, P. J., & Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience, 11,* 100–113. doi:10.1038/nrn2774

van Boxtel, J. J. A., & Lu, H. (2011). Visual search by action category. *Journal of Vision, 11*(7), Article 19. doi:10.1167/11.7.19

van Boxtel, J. J. A., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. *PLoS ONE, 7*(5), e37085. doi:10.1371/journal.pone.0037085

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience, 8,* 1651–1656. doi:10.1038/nn1607

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, & Computers, 35,* 22–31. doi:10.3758/BF03195494

Woocher, F. D., Glass, A. L., & Holyoak, K. J. (1978). Positional discriminability in linear orderings. *Memory & Cognition, 6,* 165–173. doi:10.3758/BF03197442

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245–272. doi:10.1037/0033-295X.114.2.245

Yoshida, H., & Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science, 16,* 90–95. doi:10.1111/j.0956-7976.2005.00787.x

Received September 8, 2011
Revision received April 24, 2012
Accepted April 25, 2012 ■