Concept Learning and Modal Reasoning

Charles Kemp, Faye Han & Alan Jern
Department of Psychology
Carnegie Mellon University

Abstract

Philosophers and linguists have suggested that the meaning of a concept can be represented by a rule or function that picks out examples of the concept across all possible worlds. We turn this idea into a computational model of concept learning, and demonstrate that this model helps to account for two aspects of human learning. Our first experiment explores how humans learn relational concepts such as "taller" that are defined with respect to a context set. Our second experiment explores modal inferences, or inferences about whether states of affairs are possible or impossible. Our model accounts for the results of both experiments, and suggests that possible worlds semantics can help to explain how humans learn and use concepts.

Keywords: concepts and categories; modal reasoning; possible worlds semantics

Knowledge about concepts and categories must support many kinds of operations. Consider simple relational concepts such as "taller" or 'heavier." A learner who has acquired these concepts should be able to use them for *classification*: given a pair of objects, she should be able to pick out the member of the pair that is taller than the other. The learner may also be able to solve the problem of *generation*: for example, she may be able to draw a pair of objects where one is taller than the other. The learner may even be able to use these concepts for *modal reasoning*, or reasoning about possibility and necessity. She may recognize, for example, that no possible pair (x, y) can satisfy the requirement that x is taller than y and that y is taller than x, but that it is possible for x to be taller than y and y to be heavier than x. The three problems just introduced demand increasingly more from the learner: *classification* requires that she supply one or more category labels, *generation* requires that she generate one or more instance of a concept, and *modal reasoning* requires that she make an inference about all possible instances of a concept, including many that have never been observed. This paper describes a formal model of concept learning that helps to explain how people solve all three of these problems, although we focus here on classification and modal reasoning.

Our model relies on possible worlds semantics, an approach that is often discussed by philosophers and linguists (Kripke, 1963; Lewis, 1973) but has received less attention in the psychological literature. The worlds we consider are much simpler than those typically discussed in the philosophical literature, and we focus on problems where each world includes a handful of objects that vary along a small number of dimensions. Figure 1 shows an example where the world under consideration is represented as a solid rectangle, and where three possible worlds are shown as dashed rectangles. We explore the idea that a concept corresponds to a rule represented in a compositional language of thought—for example, the rule in Figure 1 picks out duplicate objects. Given this setup, we explore how concepts can be learned from observing a small number of worlds, and how these concepts can be used to decide whether a statement is possible (true in some possible worlds) or necessary (true in all possible worlds).

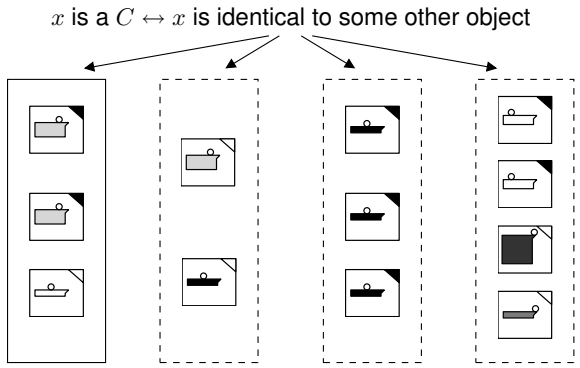

Figure 1: One actual world (solid rectangle) and three possible worlds (dashed rectangles). Each world contains between two and four objects, and the black triangles indicate which objects are instances of concept C. The geometric objects used for this illustration and for our experiments are based on stimuli developed by Kemp and Jern (2009).

Our approach builds on previous accounts of concept learning, and is related most closely to previous rule-based models that rely on logic as a representation language (Nosofsky, Palmeri, & McKinley, 1994; Feldman, 2000; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp & Jern, 2009). Most of these models, however, do not allow the category label of an object to depend on its context, or the world to which it belongs. For example, models of Boolean concept learning (Feldman, 2000) cannot capture relational concepts such as "duplicate," since Boolean logic cannot express rules that rely on comparisons between objects. Previous accounts of relational categorization and analogical reasoning (Gentner, 1983; Doumas, Hummel, & Sandhofer, 2008) often work with richer representation languages, and can therefore capture the idea that the category label of an object may depend on its role within the world (or configuration) to which it belongs. These accounts, however, are limited in another respect. In most cases they are able to compare two or more worlds that are provided as input, but they cannot generate new worlds, or account for inferences that require computations over the space of all possible worlds. In particular, we believe that previous psychological models will not account for the modal inferences that we explore.

Although previous accounts of concept learning have not

$$1 \;\; \forall x \; C(x) \leftrightarrow D_i(x) \begin{Bmatrix} = \\ \neq \\ < \\ > \end{Bmatrix} v_k$$

$$2 \;\; \forall x \; C(x) \leftrightarrow \begin{Bmatrix} \exists y \; y \neq x \; \wedge \\ \forall y \; y \neq x \rightarrow \end{Bmatrix} D_i(y) \begin{Bmatrix} = \\ \neq \\ < \\ > \end{Bmatrix} D_i(x)$$

$$3 \;\; \forall x \; C(x) \leftrightarrow \begin{Bmatrix} \exists y \; \exists Q \; y \neq x \; \wedge \\ \exists y \; \forall Q \; y \neq x \; \wedge \\ \forall Q \; \exists y \; y \neq x \; \wedge \\ \forall y \; \exists Q \; y \neq x \rightarrow \\ \forall y \; \forall Q \; y \neq x \rightarrow \\ \exists Q \; \forall y \; y \neq x \rightarrow \end{Bmatrix} Q(y) \begin{Bmatrix} = \\ \neq \\ < \\ > \end{Bmatrix} Q(x)$$

Table 1: Templates used to construct a hypothesis space of rules. An instance of a given template can be created by choosing an element from each set enclosed in braces, replacing each occurrence of D_i with a dimension (e.g. D_1) and replacing each occurrence of v_k with a value (e.g. 1).

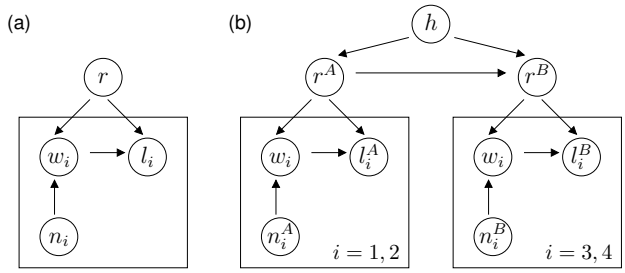

Figure 2: (a) A graphical model that can be used for learning a rule r. World w_i is generated given rule r and count vector n_i which specifies the size of w_i and the number of elements in w_i that must be assigned positive labels by rule r. Label vector l_i indicates which members of w_i are assigned positive labels by rule r. (b) Graphical model for Experiment 2. Rules r^A and r^B capture two concepts that may or may not be different. Hypothesis h indicates whether r^B is identical to r^A or drawn independently from $P(r)$.

focused on modal inferences, these inferences have been studied in several other contexts (Osherson, 1977; Nichols, 2006). Our approach is related most closely to the mental model approach developed by Johnson-Laird and colleagues (Johnson-Laird, 1982; Bell & Johnson-Laird, 1998; Evans, Handley, Harper, & Johnson-Laird, 1999), who argue, for example, that a proposition is rated as possible if a reasoner can construct a mental model (or world) which makes the proposition true. The mental models approach is broadly compatible with our own, and both can be seen as attempts to explore the psychological implications of possible worlds semantics. There are, however, at least two important differences between these approaches. First, the mental model account of modal reasoning has focused almost exclusively on deductive inferences, but we focus on inductive inferences, or inferences to conclusions that do not follow with certainty given the available evidence. Second, mental model approaches are often contrasted with approaches that rely on rules expressed in some compositional language of thought. Our approach relies critically on a representation language, and we show that a syntactic prior over expressions in this language is a useful way to capture human inductive biases.

The next section introduces our computational model and we then evaluate it in two behavioral experiments. Our first experiment demonstrates that our model helps to explain how people learn simple relational concepts such as "duplicate." Our second experiment focuses on modal reasoning, and we ask participants to learn two concepts then to decide whether statements about these concepts are possible or impossible.

A possible worlds account of concept learning

Philosophers and psychologists often distinguish between two aspects of a concept's meaning—its extension and its intension. The extension of a concept is the set of objects in the actual world that are instances of the concept. For example, the extension of the concept C in Figure 1 includes the two members of the actual world that are marked with black triangles. The extension of a concept may be the empty set—for example, in Figure 1, the extension of the concept "black objects" is the empty set, since none of the objects in the actual world is black.

Two concepts may have the same extension in a given world but may nevertheless be different. For example, the concepts of unicorns and griffins have the same extension in our world—the empty set—but intuition suggests that these concepts are nevertheless different. One reason why the concepts must be different is that we can imagine possible worlds which include both unicorns and griffins, and where the extensions of these concepts are different. The intension of a concept includes the information that allows us to pick out its extension in every possible world—for example, knowing that unicorns have horns might help to decide which objects in a given world qualify as unicorns. Formal approaches to semantics often formalize the intension of a concept as a function that picks out its extension in every possible world, and we adopt this perspective here. These functions could be represented in many different ways, and our working hypothesis is that intensions correspond to rules that are mentally represented in a language of thought. For example, the rule in Figure 1 serves to pick out an extension (i.e. the set of Cs) in each possible world.

The rule in Figure 1 is primarily expressed in English, but the representation language that we will explore is a simple version of predicate logic. Table 1 shows three templates that can be used to generate the rules we consider. In all cases, $C(\cdot)$ is the concept of interest, and each possible rule can be used to decide whether an object x is an instance of C. D_i represents the ith dimension and v_k represents the kth value along a dimension: for example, if D_1 is the dimension of color and value v_1 along this dimension indicates black, then the rule $\forall x \, C(x) \leftrightarrow D_1(x) = v_1$ indicates that x is

a C if and only if x is black. The language includes four Boolean connectives: and ($\wedge$), or ($\vee$), if ($\rightarrow$), and if and only if ($\leftrightarrow$). The language also includes four relations for comparing values along the dimensions ($=, \neq, <, >$), and incorporates quantification over the objects within a world ($\forall y$) and over dimensions ($\forall Q$). The three templates in Table 1 capture all simple rules with at most one instance of quantification over objects and at most one instance of quantification over dimensions. Our hypothesis space of rules includes all instances of the templates in Table 1 along with all conjunctions that can be generated by combining up to three of these instances. Note that some rules have identical intensions—in other words, they pick out the same set of objects in each possible world. We strip out rules with identical intensions, and include only the shortest rule for any given intension.

We now develop a probabilistic framework for learning and using intensions. Suppose that C is a concept of interest, and that rule r captures the intension of C. Suppose that a learner observes one or more worlds w_i, and let l_i be a binary label vector that indicates which objects in world w_i are instances of C. Given these observations, the posterior distribution $P(r|w_{1:k}, l_{1:k})$ will represent an ideal learner's beliefs about the rule r after observing k worlds and label vectors for these worlds. We work with the posterior distribution induced by the graphical model in Figure 2a.

We assume that rule r is generated from a prior $P(r)$ that favors simple rules, and use a description-length prior $P(r) \propto \lambda^{|r|}$, where $|r|$ is the number of symbols in rule r. For all applications of the model we set $\lambda = 0.5$. To generate a world w_i, we first generate a vector n_i that specifies the total number of objects in the world and the number that are instances of C, then sample uniformly at random among all worlds that are consistent with n_i and r. Our assumption that w_i is generated given n_i is a natural extension of previous accounts of concept learning which assume that the examples observed are randomly sampled from the set of positive instances (Tenenbaum & Griffiths, 2001). For simplicity, we assume that worlds have at most four objects, and use a uniform prior $P(n_i)$ over all count vectors that satisfy this constraint. Finally, we assume that each label vector l_i is deterministically generated given world w_i and rule r. The joint distribution over all variables can be summarized as follows:

$$
\begin{aligned}
P(r) &\propto \lambda^{|r|} \\
P(n_i) &\propto 1 \\
P(w_i|r, n_i) &= \begin{cases} \frac{1}{c(r,n_i)} & \text{if } n_i \text{ is consistent with } w_i \text{ and } r \\ 0 & \text{otherwise} \end{cases} \\
P(l_i|r, w_i) &= \begin{cases} 1 & \text{if } l_i \text{ is consistent with } w_i \text{ and } r \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

where $c(r, n_i)$ is the number of worlds that are consistent with n_i and r.

Computing the predictions of our model will be challenging in general, but for all applications in this paper we can carry out inference by enumerating the entire hypothesis space of rules and the entire set of possible worlds. Future work can attempt to develop sampling methods that efficiently compute the predictions of the model, but before beginning this enterprise it is important to explore whether our approach makes accurate predictions about human behavior. The next two sections describe experiments that highlight two distinctive aspects of our model.

Experiment 1: Learning relational concepts

Allowing for multiple worlds provides a natural way to handle relational concepts such as "biggest" or "duplicate" that depend on the world or the set of objects currently under consideration. Our first experiment explores how people learn several basic concepts, including three relational concepts. All of our participants are adults, but developmental psychologists have established that even young children are able to learn simple relational concepts. The concepts we consider are based on a developmental study of Smith (1984), and we demonstrate that our model is able to learn all of the concepts that she discusses.

Materials and Method. 14 adults participated for course credit and interacted with a custom-built computer interface. Participants were told that they would be learning about the preferences of six aliens (i.e. learning six concepts). For each alien three worlds containing three objects each were simultaneously presented on screen. The worlds used for each alien are shown in Figure 3, where each column corresponds to a world. The first five concepts are loosely inspired by the concepts studied by Smith (1984). For example, Smith's version of 3d used worlds that included objects such as a green plane, a red apple, and a yellow pig, and the underlying concept was "green plane." The objects in each world were arranged vertically on screen, and were initially ordered as shown in Figure 3, where each column corresponds to a world. Within each world, the objects could be dragged around and reordered, and participants were invited to sort them however they liked.

For all six aliens, the first two worlds had labels indicating whether or not the alien liked each object. Participants were asked to predict the alien's preferences by choosing a binary label for each object in the third world. After labeling the third world, participants provided a short written description of the current alien's preferences. The order of the six aliens (i.e. concepts) was counterbalanced across participants, and the roles of the three dimensions (ball position, color, and size) were also counterbalanced.

Model predictions. Let $w_{1:3}$ represent the three worlds and $l_{1:2}$ represent the two observed label vectors. We compute the distribution $P(l_3|w_{1:3}, l_{1:2})$ by integrating over all possible rules r. For each concept, the black bars in Figure 3 show the posterior probability that each object in the third world is an instance of the concept.

Results. For each concept the white bars in Figure 3 show the proportion of participants who chose positive labels for the objects in the third world. Responses were consistent in

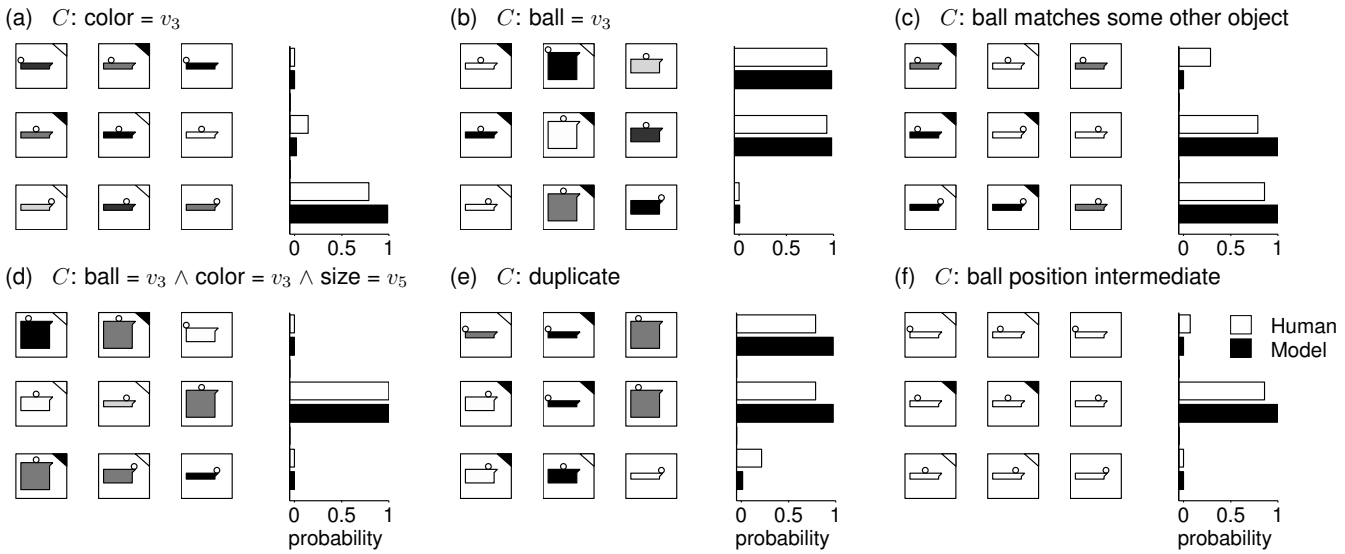

Figure 3: Stimuli and results for Experiment 1. In each panel, the three columns show three worlds that include three objects each. The objects in the first two worlds have category labels, and the bar charts show model predictions and human inferences about the labels of the objects in the third world. The black bars show probabilities generated by the model, and the white bars show the proportion of participants that chose a positive label for a given object.

all cases with the predictions of the model. If we create binary prediction vectors by thresholding the model predictions at 0.5, then 10 or more of the 14 responses for each concept were identical to the model predictions.

The written descriptions provided further evidence that participants were able to infer sensible rules and the concept descriptions in Figure 3 are based on the most common descriptions provided in the experiment. In several cases, however, participants provided rules that were more specific than the descriptions in Figure 3. For example, in condition 3a, one participant indicated that the alien liked *small* objects of a certain color.

Experiment 1 explored how participants learn relational concepts then apply them in new contexts, but did not address the topic of modal reasoning. Three worlds were shown on screen for each problem, and participants were not required to think about alternative worlds that they had not observed. Our second experiment focuses directly on modal reasoning, and uses a task where participants must go beyond the handful of worlds observed on screen and make inferences about the full space of possible worlds.

Experiment 2: Modal reasoning

Understanding the meaning of multiple concepts should allow a learner to predict which relationships between these concepts are possible and which are necessary. For example, any mother must also be a woman, and it is possible for the same person to be a mother and a grandmother. In our second experiment we asked participants to learn two concepts C_A and C_B then asked them to rate the possibility of statements involving these concepts.

Materials and Method. 15 adults participated for course credit and interacted with the same computer interface developed for Experiment 1. Participants were told that they would be learning about the preferences of five pairs of aliens. For each pair they were shown four worlds with four objects, and these worlds are shown as columns in Figure 4. The objects in the first two worlds were labeled to indicate the preferences of the first alien, and the objects in the remaining two worlds were labeled to indicate the preferences of the second alien. After observing these worlds, participants were asked to provide written descriptions of the preferences of the two aliens. Participants were then asked to rate the possibility of the six statements shown at the bottom right of Figure 4. These ratings were elicited using questions that referred to the preferences of the two aliens rather than concepts C_A and C_B: for example, ratings for statement six were based on the question "Do Mr. X and Mr. Y have identical preferences?," where X and Y were names for the two aliens in the current pair. All ratings were provided on a seven point scale with the endpoints labeled "Probably not" and "Probably." For all cases except statement 6, participants who chose a rating of 4 or above were asked to generate an example of a four-object world that satisfied the constraints in the statement. These worlds were generated using buttons that could add or remove objects from the screen, and could adjust the appearance of the objects along each of the three dimensions. The presentation order of the five pairs of concepts and the six possibility statements was counterbalanced across participants, and the role of the three dimensions was also counterbalanced.

Model predictions. Let $w_{1:4}$ represent the four worlds and $l^A_{1:2}$ and $l^B_{3:4}$ represent the label vectors for these worlds. Our approach to the task is captured by the graphical model in Figure 2b, where hypothesis $h = 1$ if rules r_A and r_B are identical and $h = 0$ if the rules are drawn independently from $P(r)$. If rules r_A and r_B are known with certainty, than each of the

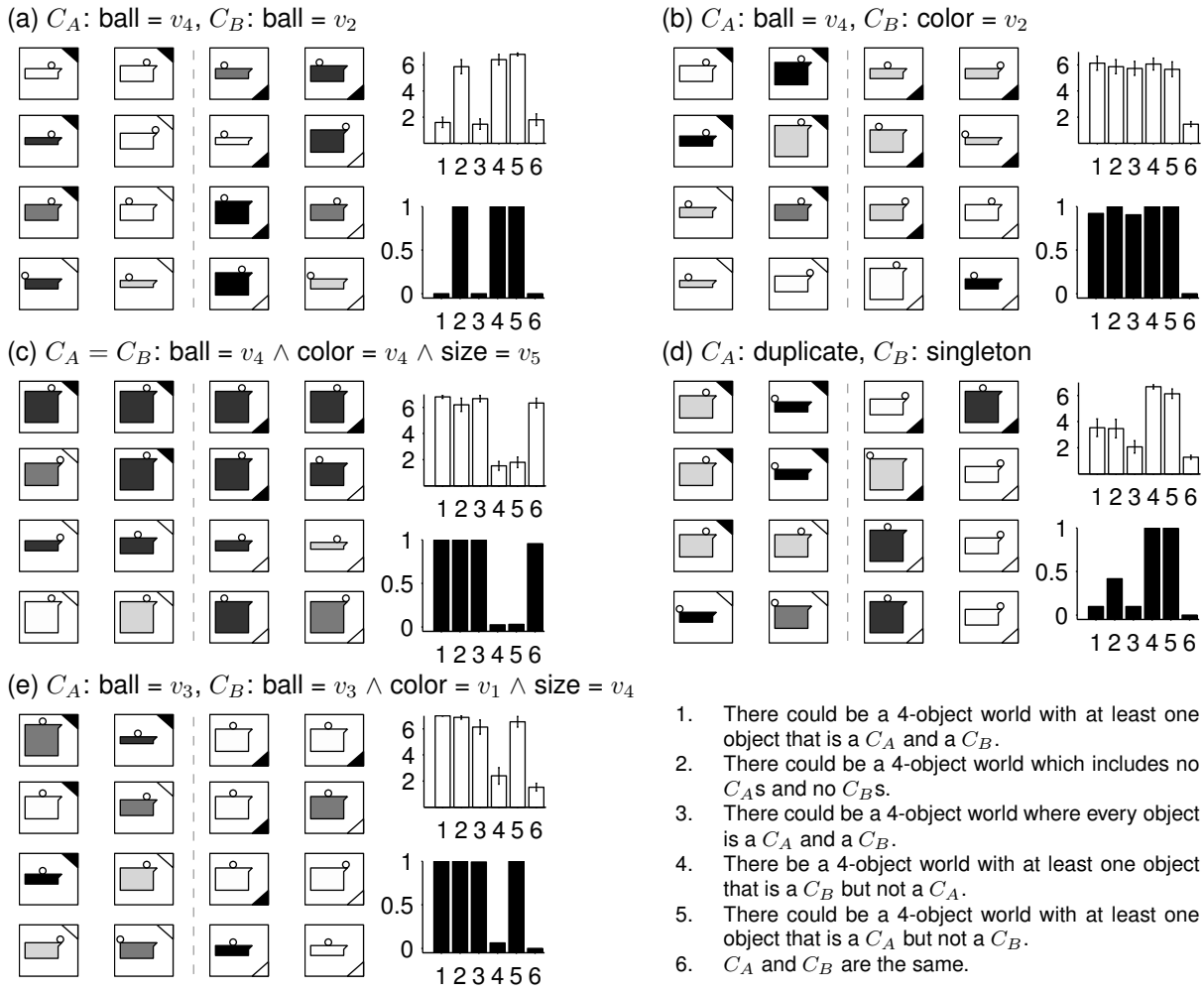

Figure 4: Stimuli and results for Experiment 2. In each panel the four columns show four worlds that include four objects each. The objects in the first two worlds are labeled with respect to concept C_A (upper right triangles), and the remaining two worlds are labeled with respect to concept C_B (lower right triangles). The bar charts show model predictions and human inferences about the six questions shown at the bottom right of the figure. Model predictions (black bars) are probabilities, and human inferences (white bars) are shown as mean judgments on a 7 point scale. Error bars show the standard error of the mean.

six statements in Figure 4 is either true or false—for example, statement 1 is true if there is at least one world with an object assigned a positive label by both r_A and r_B. We compute the posterior probability of each statement using a uniform prior $P(h)$ and integrating out rules r_A and r_B and hypothesis h. If desired, each of the six statements could be expressed in a logical language with operators that express necessity and possibility, and our approach could be combined with an explicit formal semantics for modal logic (Kripke, 1963). This degree of formalization does not seem useful for our current purposes, but may be useful in other contexts.

For each pair of categories, the black bars in Figure 3 show the posterior probabilities of the six possibility statements. Note that each pair of categories leads to a qualitatively different pattern of predictions.

Results. The white bars in Figure 4 show average human ratings for the six possibility statements. Responses for the five different pairs of categories are qualitatively different, and in all cases there is a relatively close correspondence between human ratings and model predictions.

Consider first the difference between pairs 4a and 4b. Both pairs include concepts that correspond to single values along a dimension: for example, concept C_A in both cases picks out objects with value v_4 along the ball position dimension. In pair 4a, concept C_B corresponds to a different value along the ball position dimension, which means that it is impossible for an object to simultaneously be a C_A and a C_B. The first bar in Figure 4a suggests that participants were able to make this inference. In pair 4b, however, concept C_B corresponds to a value along a different dimension, and participants were confident that there could be an object that was a C_A and a C_B. Note that participants never observed an object with labels for both concepts, which means that they had to go beyond their direct experience when making inferences about the compatibility of the two concepts.

Pairs 4c and 4e form a second natural comparison set. In

both cases, concept C_B includes only objects with a specific value along each dimension. In 4c, concept C_A specifies the same values along each dimension, and participants were confident that concepts C_A and C_B were identical (bar 6). In 4e, concept C_A specifies values along only one dimension, and the C_A objects are a superset of the C_B objects. Participants inferred that C_A and C_B are different concepts (bar 6), that every C_B is a C_A (bar 4), but that some C_A objects are not C_B objects (bar 5).

The main discrepancy between model predictions and human responses occurs for pair 4d and the first possibility judgment. The model infers that C_A includes duplicates and C_B includes singletons, and concludes that no object can be both a C_A and a C_B. Eight out of 15 participants made a similar inference, and chose ratings of 2 or below on a seven point scale, but five participants chose ratings of 6 or above, producing a mean rating of around 3.5. Many of these five participants gave complex disjunctive definitions when describing concept C_A, suggesting that they may have focused on the individual characteristics of the positive examples without reflecting on the relationships of these positive examples to the other objects in the world.

Although we know of no previous studies that combine modal reasoning and concept learning, previous work on modal reasoning has explored how people arrive at conclusions given premises supplied by the experimenter. For example, given that all artists are beekeepers and that Lisa is a beekeeper, it is possible that Lisa is an artist (Evans et al., 1999). The mental models approach can account for inferences of this kind, but note that our task is rather more challenging. We explored cases where the "premises" for modal reasoning (i.e. the meanings of the concepts) are not supplied but must instead be learned from a small number of examples. In order to handle the inductive aspect of our task, a computational approach must incorporate a human-like inductive bias, and the mental models approach is not well-equipped to satisfy this criterion. Our results, however, suggest that human inferences can be accurately predicted by combining a possible worlds framework with a description length prior over logical rules.

Conclusion

We developed a model of concept learning that relies on the notion of possible worlds and evaluated it in two experiments. Our first experiment suggests that our approach helps to explain how humans learn relational concepts such as "bigger" or "duplicate." Our second experiment demonstrates that humans readily make modal inferences about concepts, and illustrates that a possible worlds approach can account for this ability.

Although modal reasoning is an especially natural application for a possible worlds approach, the same approach should help to illuminate other aspects of human learning and reasoning. Philosophers and linguists have used the possible worlds framework to clarify the meaning of counterfactual statements (Lewis, 1973), and to characterize the content of claims about belief, desire, and knowledge (Hintikka, 1962). The psychological implications of these projects have received relatively little attention, but the possible worlds approach is a promising way to study the many different ways in which human concepts are put to use.

Acknowledgments This work was supported in part by NSF grant CDI-0835797.

References

Bell, V. A., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science*, *22*(1), 25–51.

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), 1–43.

Evans, J. S. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: a test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(6), 1495–1513.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630-633.

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

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Hintikka, J. (1962). *Knowledge and belief*. Cornell University Press.

Johnson-Laird, P. N. (1982). Formal semantics and the psychology of meaning. In S. Peters & E. Saarinen (Eds.), *Processes, beliefs and questions* (pp. 1–68). D. Reidel.

Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In *Advances in Neural Information Processing Systems 22* (pp. 934–942).

Kripke, S. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica*, *16*, 83–94.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

Nichols, S. (2006). Imaginative blocks and impossibility: an essay in modal psychology. In S. Nichols (Ed.), *The architecture of the imagination: new essays on pretence, possibility and fiction* (pp. 237–255). Oxford: Oxford University Press.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Osherson, D. (1977). *Logical abilities in children*. L. Erlbaum Associates.

Smith, L. B. (1984). Young children's understanding of attributes and dimensions: a comparison of conceptual and linguistic measures. *Child Development*, *55*, 363–380.

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