Context-Sensitive Induction

Patrick Shafto[1], Charles Kemp[1], Elizabeth Baraff[1], John D. Coley[2], & Joshua B. Tenenbaum[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Psychology, Northeastern University

Abstract

Different kinds of knowledge are relevant in different inductive contexts. Previous models of category-based induction have focused on judgments about taxonomic properties, but other kinds of models are needed for other kinds of properties. We present a new model of reasoning about causally transmitted properties. Our first experiment shows that the model predicts judgments about a disease-related property when only causal information is available. Our second experiment uses a disease-related property and a genetic property in a setting where both causal and taxonomic information are available. Our new model accounts only for judgments about the disease property, and a taxonomic model accounts only for judgments about the genetic property. This double dissociation suggests that qualitatively different models are needed to account for property induction.

Any familiar thing can be thought about in a multitude of ways. A cat is a creature that climbs trees, eats mice, has whiskers, belongs to the category of felines, and was revered by the ancient Egyptians. Knowledge of all of these kinds plays an important role in inductive inference. If we learn that cats suffer from a recently discovered disease, we might think that mice also have the disease — perhaps the cats picked up the disease from something they ate. Yet if we learn that cats carry a recently discovered gene, lions and leopards seem more likely to carry the gene than mice. Flexible inferences like these are a hallmark of human reasoning, which is notable for the selective application of different kinds of knowledge to different kinds of problems.

Psychologists have confirmed experimentally that inductive generalizations vary depending on the property involved. When told about genes or other internal anatomical properties, people generalize to taxonomically related categories (Osherson, Smith, Wilke, López, and Safir, 1990). When told about novel diseases, however, people generalize to categories related by the causal mechanism of disease transmission (Shafto and Coley, 2003). Psychologists have also suggested, at least in principle, how complex inferences like these might work. Flexible inductive inferences are supported by *intuitive theories*, or "causal relations that collectively generate or explain the phenomena in a domain" (Murphy, 1993). Many theories may apply within a single domain, and very different patterns of inference will be observed depending on which theory is triggered.

Although a theory-based approach is attractive in principle, formalizing the approach is a difficult challenge. Previous work describes a theory-based taxonomic model (Kemp and Tenenbaum, 2003), and here we use the same Bayesian framework to develop a theory-based model for induction about causally transmitted properties like diseases. The models differ in the causal knowledge used to generate probability distributions over potential hypotheses, resulting in qualitatively different patterns of generalization for different theories. These are but two of the many models that may be needed to explain the full set of inductive contexts, and extending our framework to deal with a broader range of contexts is an ongoing project.

Our work goes beyond previous formal models, which find it difficult to capture the insight that different kinds of knowledge are needed in different inductive contexts. In the similarity-coverage model, a representative and often-cited example, the domain-specific knowledge that drives generalization is represented by a similarity metric (Osherson et al., 1990). Even if we allow a context-specific notion of similarity, a similarity metric is too limited a representation to carry the richly structured knowledge that is needed in some contexts. In contrast, the knowledge that drives generalization in our Bayesian framework can be as complex and as structured as a given context demands.

We begin by introducing our new model of reasoning about causally transmitted properties. We then present experiments showing that our new model predicts human generalizations about diseases, but not about genetic properties. The theory-based taxonomic model has complementary strengths, and predicts generalizations about genetic properties, but not diseases. We finish by comparing our model to previous approaches, and describing some of the challenges to be surmounted in developing a truly comprehensive theory of context-sensitive induction.

Theory-based induction

Our theory-based framework includes two components: an engine for Bayesian inference and a theory-based prior. The inference engine implements rational statistical inference, and remains the same regardless of the inductive context. We model these theories using probabilistic processes over structured represenations of causal knowledge.

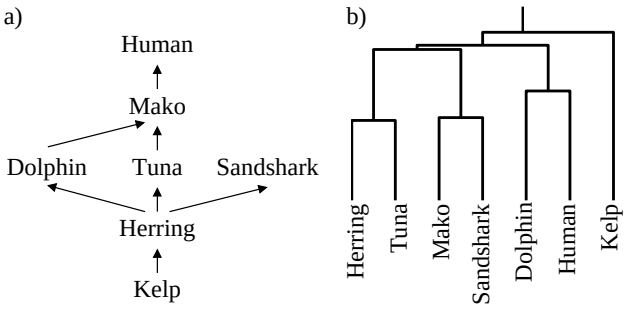

Figure 1: Two different theories of domain structure. (a) Food web structure. (b) Taxonomic structure.

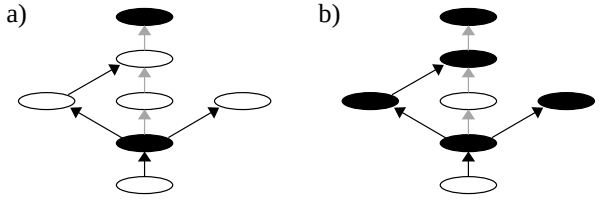

Figure 2: One possible disease simulation. (a) Initial step showing species hit by the background rate (black ovals) and active routes of transmission (black arrows). (b) Total set of species with disease via background and transmission.

The Bayesian inference engine

Suppose that we observe a set of examples D that share a novel property, and wish to compute the strength of the conclusion that an unobserved instance, y, also has the property. Let H be the hypothesis space — the set of all possible extensions of the property. The probability that y has the property can be computed by summing over all hypotheses:

$$p(y|D) = \sum_{h \in H} p(y|h)p(h|D) = \sum_{h} \frac{p(y|h)p(D|h)p(h)}{p(D)}$$

where the last step follows from Bayes' rule.

Now $p(y|h)$ equals one if $y \in h$ and zero otherwise. We assume that $p(D|h)$, the likelihood of observing the data given h, is 1 if the data are consistent with the hypothesis and 0 otherwise. Then

$$p(y|D) = \sum_{h: y \in h, D \subset h} \frac{p(h)}{p(D)}$$

The denominator can be expanded by summing over all hypotheses: $p(D) = \sum_h p(D|h)p(h) = \sum_{h: D \subset h} p(h)$. Thus

$$p(y|D) = \frac{\sum_{h: y \in h, D \subset h} p(h)}{\sum_{h: D \subset h} p(h)} \quad (1)$$

The generalization probability $p(y|D)$ is therefore the proportion of hypotheses consistent with D that also include y, where each hypothesis is weighted by its prior probability $p(h)$. If the conclusion y and observed examples D are included in many hypotheses with high prior probability, then the probability of generalization will be high.

Theory-based priors

The prior probabilities $p(h)$ in Equation 1 are determined by intuitive theories of the domain under consideration. We model these theories using a combination of structured representations of causal knowledge and parameters for generating a probability distribution over hypotheses.

A generative model of causally transmitted properties. Our model will apply to any property that depends on asymmetric causal relationships among objects in a domain. For concreteness, however, assume that the domain is a set of species, and the novel property is a disease spread by predator-prey relations between species. We generate a prior distribution using a theory with two components. First, the theory states the predator-prey relations that hold over the domain. The set of relations can be represented as a food web (see Figure 1). Second, the theory assumes that properties are generated by a probabilistic process over the web. The process has two parameters, a base rate and a transmission rate. The base rate is the probability that an animal inherits the property from a cause external to the web. The transmission rate is the probability that the property is transmitted from a species along an arrow to a causally related species. Assuming that the base rate affects each species independently and the transmission rate affects each arrow independently, we obtain a prior distribution over all extensions of the property.

The prior can be generated by repeatedly simulating the arrival and transmission of disease in the system. A single simulation chooses a set of animals that are affected by the base rate, and a set of causal links that are active (Figure 2a). These choices imply that a certain set of animals will catch the disease, and that hypothesis is the output of the simulation (Figure 2b). If we repeat the simulation many times, the prior probability of any hypothesis is the number of times it is chosen as output. Reflecting on these simulations should establish that the prior captures two basic intuitions. First, species that are linked in the web are more likely to share the property than species that are not directly linked. Second, property overlap is asymmetric: a prey animal is more likely to share the property with its predator than vice versa.

Several qualitative predictions emerge from the model, and will be tested in the experiments that follow. *Asymmetry* is the prediction that generalizations from prey to predator are stronger than generalizations from predator to prey. Asymmetry is a consequence of the assumption that causal transmission is directed. *Dissipation* is the prediction that the strength of generalization decreases with increasing distance in the web, and should be seen

because the mechanism of causal transmission is fallible. Finally, *fanning* is the prediction that generalization from predator to prey depends on the number of other species that the predator eats. If the predator has the property but eats only one species, it is fairly likely that the predator caught the disease from that species. If the predator eats many species, the chance that it inherited the property from any given prey species is small, since there are many alternative sources of the disease.

Although our experiments only consider the case of disease transmission in a food web, many other inductive contexts fit the pattern of asymmetric causal transmission that the model is designed to capture. Within the domain that we focus on, the causal model should also apply to inferences about the transmission of toxins. Outside of this domain, the model could be used, for example, to model inference about the transmission of lice between children at a day care, the spread of secrets through a group of colleagues, or the progression of fads through a society.

A generative model of taxonomic properties. A second method for generating a prior distribution is described by Kemp and Tenenbaum (2003). The model is based on two key ideas: species fall at the leaves of a known taxonomic tree (see Figure 1), and the novel properties are generated by a mutation process over the tree. Imagine a property that arises at the root of the tree, and spreads out towards the leaves. The property starts out with some value (on or off), and at each point in the tree there is a small probability that the property will mutate, or switch its value. The mutation process has a single parameter — the average rate at which mutations occur. As for the previous generative model, this stochastic process induces a prior distribution over all possible extensions of a novel property. The closer two species lie in the tree, the more likely they are to share the property.

Observe that this taxonomic model is closely related to the model for causally transmitted properties. Both models are based on structured representations (food webs or trees), and incorporate stochastic processes over those representations. Both were also built by thinking about how properties are distributed in the world. It is therefore not surprising that both models match models used by scientists in other fields — models like the causal model are used by epidemiologists, and models like the taxonomic model are used by evolutionary biologists. In developing these models, then, we have given rational analyses of inductive inference in two contexts (Anderson, 1990). Extending our approach to other contexts is a matter of formally specifying how the properties covered by those contexts are distributed in the world.

Our primary goal so far has been to characterize the knowledge that plays a role in generalization (theory-based priors), and the input-output mapping that allows this knowledge to be converted into judgments of inductive strength (Bayesian inference). Our work is located at the most abstract of Marr's levels — the level of computational theory (Marr, 1982) — and we make no suggestion about the process by which people make generalizations, or the representations they may use. Inference in both our models can be implemented using belief propagation over Bayes nets, and we are encouraged to know that efficient implementations exist. We are not committed to the claim, however, that inference in these Bayes nets resembles cognitive processing.

Experiments

This section compares model predictions to human generalizations in two experiments. The first experiment focuses on the performance of the new causal model, comparing model predictions with human generalizations when only causal information is available. The second experiment considers generalizations of diseases and genetic properties, and compares the performance of the causal and the taxonomic models.

Experiment 1: Testing the Causal Model

Participants. Nineteen people participated in this experiment.

Materials. Participants were given a set of 7 cards describing feeding relations between the animals. The cards presented a blank label for the animal ("Animal X") and the creature's immediate predators and prey. This information is presented schematically in Figure 1. Note that the full graph was never presented and no taxonomic information was available to participants in this experiment.

Procedure. The experiment contained two phases: training and a generalization task. In the training phase, participants were given a set of 7 cards corresponding to the animals in a set. Before proceeding, participants were required to score 85% on a true/false test to demonstrate that they were familiar with the information on the cards. In the generalization phase, participants were presented with a series of 42 questions (all possible pairs) of the form, "Animal X has a disease. How likely is it that animal Y has the same disease as animal X?" Participants rated the likelihood for each question on a scale of 1-7, where 1 indicated "very likely" and 7 indicated "very unlikely." Questions appeared in random order.

Results. During the experiment, some participants noted that they had used the rating scale backwards. These subjects' ratings were inverted as they requested, and we inspected the data for other participants who may have used the scale backwards. First we identified all questions for which the average rating was < 2 or > 6. If a participant provided opposite ratings (> 6 on a question for which the average was < 2 or vice versa) for more than 2/3 of these questions, we eliminated them from further analysis. No participants were eliminated from experiment 1.

We fit the model to the data by searching for the best values of the base and transmission rates. The model shows robust performance across a range of parameter values, and all the results in this paper used a base rate of 0.1 and a transmission rate of 0.5. Figure 3 shows that the model gives a good account of the human judgments.

Qualitative results were also obtained for the three

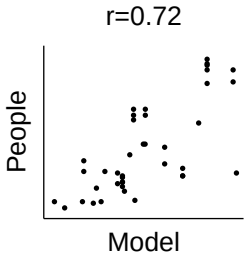

Figure 3: Human judgments of argument strength vs. model predictions.

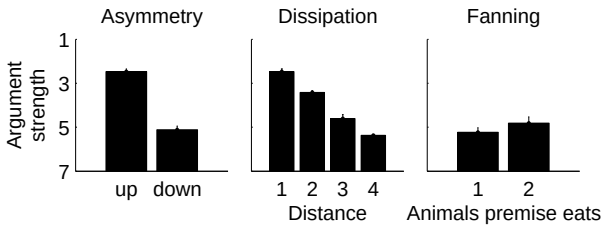

Figure 4: Qualitative effects: Experiment 1.

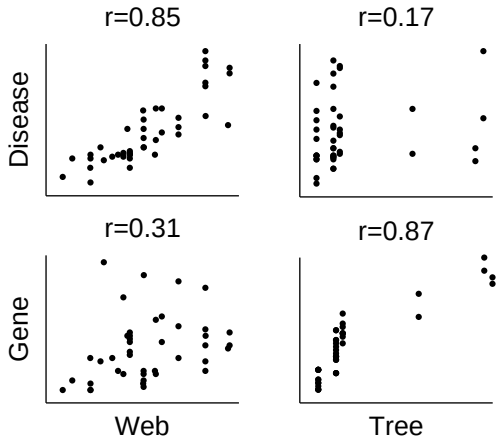

Figure 5: Human judgments of argument strength vs. model predictions for two models and two conditions.

causal phenomena: causal asymmetry, dissipation, and fanning. To test for causal asymmetry, the average ratings for generalizations up the chain (from prey to predator) were compared to generalizations down the chain (from predator to prey) using a two-tailed t-test. The results indicate a significant difference (see Figure 4), $t(12) = 18.00$, $p < 0.001$, with generalizations up the food chain stronger than generalizations down the food chain, as predicted by the model. To test for the dissipation effect, generalizations up the chain were collapsed into four categories based on the distance from the premise to the conclusion. Dissipation predicts that generalization strength should decrease with increasing distance. Because there was only one argument with a distance of 4, statistical significance could not be evaluated for this case. The results indicate significant differences $1 < 2$, $t(11) = 10.47$, $p < 0.001$, and $2 < 3$, $t(6) = 7.10$, $p < 0.001$. Finally, to test the fanning effect we collapsed generalizations from predators to prey into two categories based on the total number of prey (1 or 2). The fanning effect predicts that generalizations from animals with multiple prey should be weaker. Results indicate no evidence for the fanning effect, $t(5) = 1.09$, $p > 0.20$, and we consider possible explanations in the next section.

Experiment 2: Different Domain Theories

This experiment tests whether people use different kinds of prior knowledge in a single domain. Participants were taught about food web and taxonomic relations between a set of familiar creatures (see Figure 1). Participants were then required to generalize novel genetic or disease properties.

Participants. There were 20 participants in the disease condition and 9 in the gene condition.

Materials. Participants were again given 7 cards representing the species. For this experiment, however, each card contained the name of a species (e.g. "Herring"), its immediate neighbors in the food web, and all other species in the same taxonomic category. Note that neither the full food web nor the taxonomic tree was ever presented to participants.

Procedures. In the training phase, participants were tested on both the taxonomic and the food web relations among creatures. In the generalization task, participants were assigned to a novel disease or novel gene condition. Participants in the novel disease condition were presented with questions of the form, "Tuna have a disease. How likely is it that dolphin have the same disease as tuna?" For the novel gene condition, the predicate "have a disease" was changed to "have a gene called XR-23."

Results. The data were inspected using the criteria described in the first experiment, and one participant was eliminated.

The causal and taxonomic models were both fit to the data from each condition (Figure 5). The food web model used the same parameters as in the previous experiment. We fit the taxonomic model by searching for the best value for the mutation rate. The results show a double-dissociation between model predictions and human generalizations. The predictions of the causal model fit human judgments about diseases, but not judgments about genes. Conversely, the predictions of the taxonomic model fit human generalizations of genes, but not generalizations of diseases. The weak correlations of the web model with the gene data and tree model with the disease data reflect the fact that taxonomic relations were not orthogonal to food web relations (mammals appear high in the food web while plants and fish appear relatively low).

Qualitative comparisons were also conducted. We ex-

pected only the participants in the disease condition to demonstrate causal asymmetry, dissipation, and fanning. As predicted we found participants in the disease condition demonstrated significantly stronger generalizations from prey to predator than from predator to prey, $t(12) = 8.69$, $p < 0.001$. Results for the gene condition indicate no differences, $t(12) = 0.14$, $p > 0.20$. There was a significant dissipation effect in the disease condition, $1 < 2$, $t(11) = 8.03$, $p < 0.001$, and $2 < 3$, $t(6) = 2.71$, $p < 0.05$. The gene effect showed no corresponding effect, both $p > 0.40$. Finally, the fanning effect was not observed in either the disease or the gene conditions, both $p > 0.40$.

Across both experiments, no evidence was found for the fanning effect, one of the qualitative effects derived from the causal model. One possible explanation is methodological: subjects were asked to judge a single argument at a time, but asking them to sort the entire set in order of inductive strength might have produced more finely tuned judgments. A second, more interesting possibility is that people find reasoning backwards through a causal chain more difficult than reasoning forward. These possibilities will be pursued in future work. Despite our negative results for the fanning effect, the causal model derives strong support from the correlations achieved with human judgments.

These results of these studies demonstrate a double-dissociation between model predictions and human generalizations. When reasoning about disease, people's generalizations correlate highly with the causal model but weakly with the taxonomic model. In contrast, when reasoning about genes, people's generalizations correlate highly with the taxonomic model but weakly with the causal model. Qualitative comparisons also support the conclusion that multiple kinds of knowledge are necessary to account for reasoning in different inductive contexts.

Discussion

The importance of the inductive context has been downplayed by many previous accounts of category-based induction. Researchers have often used the domain of biology and chosen so-called 'blank predicates,' which usually trigger the default taxonomic context. For example, a subject told that 'cats have property P' is likely to assume that P is related to internal genetic and physiological factors. The emphasis on biology and on the default taxonomic context in particular may mean that some existing models for category-based induction are more limited than is commonly realized. The coverage model, for example, includes a hierarchical taxonomy of categories, a structure that is ideal for reasoning about the default biological context but of limited value in other contexts.

Despite the general emphasis on taxonomic reasoning, several researchers have argued for the importance of the inductive context. Our experiments are perhaps most closely related to those of Heit and Rubinstein (1994), who show that people draw different conclusions when asked about physiological properties than when asked about behavioral properties. Researchers have suggested that formal approaches like the coverage model and the feature-based model can account for some context effects (Heit and Rubinstein, 1994; Sloman, 1993). We argue that both approaches are insufficient to account for the full range of contexts because their representations of prior knowledge are too impoverished to capture the abstract theoretical knowledge that constrains and guides induction.

Here we give non-standard descriptions of the coverage and the feature-based models that show how they relate to our theory-based models. All three models include context-specific representations. To recap, our Bayesian approach uses a prior generated by a context-specific theory. The coverage model includes a similarity measure and a taxonomy of categories, both of which may depend on the inductive context. The feature-based model includes a set of context-specific features.

The similarity-coverage model can account for context effects such as generalizations of anatomical and behavioral properties by using different measures of similarity for different contexts. However, this approach can not account for our disease-induction data. While shared behaviors and shared anatomy can be reasonably thought of as contributing to similarity under some natural taxonomies, predator-prey relationships do not confer similarity in any conventional sense, nor are they naturally organized in a taxonomy. It may be possible to account for our data by leaving similarity behind and moving to a context-specific measure of association. Similarity alone, however, will not be able to provide a full account of context-specific reasoning.

Unlike the similarity-coverage model, the feature-based model may be able to account for our results given a sufficiently large set of features. To see why, note that the information contained in either of our theory-based priors could be represented using a large set of features sampled from that prior. Our objection to the model is that it cannot account for inductive inference in cases where people have seen few directly relevant features or perhaps none at all. Cases like these draw directly on intuitive theories, and Experiment 1 is one of them. Any features used by the subjects in this experiment must have been constructed solely from the information provided about predator-prey relations in the domain. A simple feature representation where every predator-prey pair shares a unique feature seems unlikely to work — in particular, it is hard to see how the asymmetry effect would emerge naturally from this representation.

A comparison with the feature-based model is illuminating because it underlines the importance of theories. It is not enough to present a mechanism for inference over features, and argue that it can account for human judgments given the right set of features. We also need to explain how subjects might have acquired those features. Similarly, it not enough to present a framework for Bayesian inference and argue that the framework can account for human judgments given the right prior. We also need to explain how the prior is generated. In our framework, theories provide that explanation, and it is

difficult to see how to account for real-world inductive inference without them.

Consider, for example, the case of a scientist who lives on an island where the local food web is represented by Figure 1a. Suppose that the scientist has recorded the distribution of one thousand different diseases, and (possibly unknown to him), the distribution closely matches the distribution predicted by our generative model. When asked the questions in Experiment 2, the scientist's responses match our model perfectly, but we cannot conclude that he has our theory – perhaps he is using the feature-based model over the data he has collected. Suppose, however, we ask the scientist a counterfactual question: we ask him about an ecosystem that is identical except that now people eat kelp, and makos do not eat tuna. A scientist with our theory of causally-transmitted properties will have no trouble, but a scientist without the theory will be lost. Neither of our experiments addresses this question, but we believe that people will respond flexibly when asked to reason about counterfactuals, or otherwise given information that alters an underlying theory. Prior distributions (or feature sets) alone are unable to explain these effects, and we conclude that the representational power of theories is indispensable.

In this work, we do not claim to be modeling all or most of the content and structure of people's intuitive theories. Rather, we are modeling just those aspects of a theory which appear necessary to support inductive reasoning about properties in these contexts. We are agnostic about whether people's intuitive theories contain much richer causal structures than those we have modeled here (Carey, 1985), or whether they are closer to light or skeletal frameworks with just a few basic principles (Wellman and Gelman, 1992).

Although we believe our theory-based approach is more likely than previous formal models to yield a satisfying account of context-sensitive induction, the work reported here is only a first step towards that goal. Perhaps the biggest gap in our model is that we have not specified how to decide which theory is appropriate for a given argument. Making this decision automatically will require a semantic module that knows, for example, that words like 'hormone' and 'gene' are related to the taxonomic theory, and words like 'disease' and 'toxin' are related to the theory of asymmetric causal transmission. Integrating this semantic knowledge with our existing models of inductive reasoning is an ongoing project.

We have discussed theories that account for inductive reasoning in two contexts, but it is natural and necessary to add more. Once we allow multiple theories, it is important to consider how many are needed and how they might be learned. Our Bayesian approach makes it possible to address these problems: Kemp, Perfors, and Tenenbaum (2004) present an approach to learning a single domain theory, which could be extended to learn multiple theories in a single domain.

Conclusion

Every real-world inference is embedded in some context, and understanding how these different contexts work is critical to understanding real-world induction. We have suggested that different contexts trigger different intuitive theories, and that modeling the theories involved explains patterns of induction across different contexts. We described a novel theory of causally-transmitted properties, and showed that it accounts for inductive judgments about diseases but not genetic properties. Our new theory complements a previously described taxonomic theory, which accounts for judgments about genetic properties but not diseases. Two theories will not be enough, and characterizing the space of theories that people use and the process by which they are acquired is a challenging long-term project. We believe that our Bayesian approach suggests a promising way to attack these fundamental problems.

Acknowledgements. We thank NTT Communication Science Laboratories and DARPA for research support. JBT was supported by the Paul E. Newton Career Development Chair.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

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

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

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2004). Theory-based induction. In *Proceedings of the 26th annual conference of the Cognitive Science Society*.

Kemp, C. and Tenenbaum, J. B. (2003). Learning domain structures. In *Proceedings of the 25th annual conference of the Cognitive Science Society*.

Marr, D. (1982). *Vision*. W. H. Freeman.

Murphy, G. L. (1993). Theories and concept formation. In Mechelen, I. V., Hampton, J., Michalski, R., and Theuns, P., editors, *Categories and concepts: Theoretical views and inductive data analysis*. Academic Press.

Osherson, D., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.

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

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25:213–280.

Wellman, H. and Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43:337–375.