REPORT DOCUMENTATION PAGE

1. REPORT DATE June 30, 2010	2. REPORT TYPE Final Technical Report	3. DATES COVERED 12/1/06 - 11/30/09

4. TITLE AND SUBTITLE

Theory-Based Bayesian Models of Inductive Inference

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-07-1-0075

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Joshua B. Tenenbaum

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR
875 N. Randolph St
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A = Approved for public release; distribution is unlimited

AFRL-OSR-VA-TR-2012- 0034

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our research aims to develop computational models of inductive inference in higher-level human cognition, specifically the ability to generalize from sparse, noisy, ambiguous data, and to build more human-like machine learning systems. Work has focused on two areas of cognition: learning about categories and their properties, and learning about relational structures -- specifically, systems of causal and social relations. We have developed a theory-based Bayesian framework for modeling these learning tasks as statistical inference over hierarchies of structured knowledge representations. Our models have made several contributions. First, they have explained a broad range of phenomena with high quantitative accuracy, using a minimum of free parameters. Second, they have provided a rational framework for explaining how and why everyday induction works, in terms of approximations to optimal statistical inference in natural environments. Third, they have provided tools for elucidating people's implicit theories about the structure of the world – describing the form and content of the prior knowledge that guides inductive inference and explaining how it may be acquired from experience. Finally, our models have led to improved algorithms for machine learning of category structures and their property distributions, causal networks, and the structure of social relations, thus bringing artificial intelligence systems closer to the capacities of human intelligence at the same time as they support a better understanding of human intelligence.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Joshua B. Tenenbaum
					19b. TELEPHONE NUMBER 617-452-2010

20120918137

THEORY-BASED BAYESIAN MODELS OF INDUCTIVE INFERENCE

Final Technical Report

Joshua B. Tenenbaum (PI)
Department of Brain and Cognitive Sciences
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology

Project Summary

Our research aims to develop computational models of inductive inference in higher-level human cognition. Work has focused on two areas of cognition: learning about categories and their properties, and learning about relational structures -- specifically, systems of causal and social relations. The modeling challenge is to explain how people are able to make strong inductive leaps -- or generalizations to novel unseen cases -- that go far beyond the sparse, noisy, ambiguous data they observe during learning. Our research seeks to understand the computational basis of these everyday inductive leaps -- to model them quantitatively, to explain in principled rational terms how they can be as successful as they are, and to build machines with the same kinds of common-sense inductive capacities.

Traditional accounts of induction emphasize either the power of statistical learning, or the importance of strong constraints from structured domain knowledge, intuitive theories or schemas. Our work is based on the premise that both components are necessary to explain the nature, use and acquisition of human knowledge. We have developed a theory-based Bayesian framework for modeling inductive learning and reasoning as statistical inferences over hierarchies of structured knowledge representations. This theory-based Bayesian framework is primarily and traditionally posed at Marr's level of computational theory. We have also begun to explore rational algorithmic or process-level accounts of human cognition, based on sensible ways of approximating intractable Bayesian computations on sparse data.

Our theory-based Bayesian models have made several contributions. First, they have explained a broad range of phenomena with high quantitative accuracy, using a minimum of free parameters. Second, they have provided a rational framework for explaining how and why everyday induction works, in terms of approximations to optimal statistical inference in natural environments. Third, they have provided tools for elucidating people's implicit theories about the structure of the world. For the inductive inference tasks we study, people are able to make strong inferences from very limited data, and these inferences must depend on some form of prior knowledge. Our models provide ways of describing that prior knowledge and explaining how it too may be acquired from experience. Finally, our models offer a two-way bridge to the state of the art in AI. They have led to improved algorithms for machine learning of category structures and their property distributions, causal networks, and the structure of social relations, thus bringing artificial intelligence systems closer to the capacities of human intelligence at the same time as they support a better understanding of human intelligence.

Our work makes several contributions to Air Force research goals. By better characterizing human learning and inference in computational terms, our models can suggest ways to improve training and agent modeling for simulations. By developing more human-like algorithms for machine learning and reasoning, our work could lead to computer systems that can replace, supplement or extend the existing capacities of Air Force personnel. Finally, an extra payoff comes from pursuing these two goals together. By developing better machine-learning and reasoning systems that operate on the same principles as human cognition does, we should be able to make human-machine interaction and collaboration more valuable and efficient.

1 General technical approach

Theory-based Bayesian models of induction focus on three critical questions: what is the content of people's intuitive theories of the world, how are they used to support rapid learning, and how can they themselves be learned? The learner evaluates hypotheses *h* about some aspect of the world -- the meaning of a word, the extension of a property or category, or the presence of a hidden cause – given observed data *d* and subject to the constraints of a background theory *T*. Hypotheses are scored by computing posterior probabilities via Bayes' rule:

$$P(h \mid d, T) = \frac{P(d \mid h, T) P(h \mid T)}{\sum_{h' \in H_T} P(d \mid h', T) P(h' \mid T)} \quad . \qquad (1)$$

The likelihood $P(d|h,T)$ measures how well each hypothesis predicts the data, while the prior probability $P(h|T)$ expresses the plausibility of the hypothesis given the learner's background knowledge. Posterior probabilities $P(h|x,T)$ are proportional to the product of these two terms, representing the learner's degree of belief in each hypothesis given both the constraints of the background theory *T* and the observed data *d*. Adopting this Bayesian framework is just the starting point for our cognitive models. The challenge comes in specifying hypothesis spaces and probability distributions that support Bayesian inference for a given task and domain. In theory-based Bayesian models, the domain theory plays this critical role.

More formally, the domain theory *T* generates a space H_T of candidate hypotheses, such as all possible meanings for a word, along with the priors $P(h|T)$ and likelihoods $P(d|h,T)$. Prior probabilities and likelihoods are thus not simply statistical records of the learner's previous observations. Rather, they are products of abstract systems of knowledge that go substantially beyond the learner's direct experience of the world, and can take qualitatively different forms in different domains.

It is often crucial to distinguish multiple levels of knowledge in a theory. The base level of a theory is a structured probabilistic model that defines a probability distribution over possible observables -- entities, properties, variables, events. This model is typically built on some kind of graph structure capturing relations between observables, such as a taxonomic hierarchy or a causal network, together with a set of numerical parameters. The graph structure determines qualitative aspects of the probabilistic model; the numerical parameters determine more fine-grained quantitative details. At a higher level of knowledge are abstract principles that generate the class of structured models a learner may consider, such as the specification that a given domain is organized taxonomically or causally. Inference at all levels of this theory hierarchy – using theories to infer unobserved aspects of the data, learning structured models given the abstract domain principles of a theory, and learning the abstract domain principles themselves -- can be carried out in a unified and tractable way with hierarchical Bayesian models.

The following sections describe more specific theory-based Bayesian models for the two areas of focus in this project, category learning and learning causal and social relations. Numbered references refer to publications cited at the end of this report.

2 Learning categories

Much of human knowledge is organized into categories, which summarize the properties of objects and the ways in which we act towards them. For example, in a military setting a person might need to discriminate many different categories of vehicles, knowing the capabilities of each, and know how to identify whether those vehicles are likely to be friends or foes. How are categories structured, and how are they learned? We have developed theory-based Bayesian models to answer these questions using several techniques that build on – and contribute to – state-of-the-art research in machine learning, statistics, and artificial intelligence. Nonparametric Bayesian methods allow us to build models of categorization whose complexity does not have to be fixed in advance. The number of categories, along with the statistical features of each category, can be discovered automatically from observations. Inference can be performed using Monte Carlo methods that grow the effective number of categories just as the data require, naturally embodying a version of Occam's razor that balances representational simplicity and fit to the data in inferring the appropriate number of categories. Hierarchical Bayesian methods allow us to express higher-level abstract constraints on the learner's more concrete hypotheses in forms that can themselves be learned from data. Probabilistic models defined over rule-based representations can capture concepts whose structure is defined symbolically. Priors on rule-based concepts can be defined using probabilistic grammars – again embodying a natural version of Occam's razor in a way that reflects the intuitions of "minimum description length" (MDL) learning, but embedded in a probabilistic framework that gives a principled basis for people or machines should express uncertainty about category membership.

We have used these methods separately and in combination to model a number of central phenomena in human category learning, and to build more human-like machine learning systems. With Charles Kemp, Amy Perfors, and Mike Frank, we have built hierarchical nonparametric models for how children learn to learn categories and word meanings (1, 25), and syntactic constructions (8, 23). These models not only learn new concepts from few examples, but also learn what abstract properties "good" concepts have in common which allows them to learn how to learn new concepts more quickly. With Mike Frank and Noah Goodman, we have built hierarchical Bayesian models for learning word meanings that jointly infer lexical concepts and speakers' referential intentions in particular communicative acts, which substantially improves accuracy of learned meanings (5, 17).

We have built hierarchical nonparametric models for discovering multiple cross-cutting ways to categorize a given domain (13) – for instance, learning that animals are best grouped taxonomically (into mammals, fish, birds, reptiles, etc.) in order to explain their anatomical and physiological features, but can also be grouped according to ecological niches (e.g., land, air, sea predators or prey) in order to explain their behavioral features.

With Noah Goodman, Tom Griffiths and Jacob Feldman, we have built grammar-based models for learning rule-based concepts (2, 30). We have begun to explore how these grammar-based approaches can be applied to learning compositional aspects of semantics, with Steve Piantadosi (20), and to learning structured object concepts, with Virginia Savova and Frank Jakel (22, 26).

With Charles Kemp, we have built hierarchical Bayesian models on top of grammar-based representations for learning the abstract structural form of a domain (3). Different grammars can capture the tree-structure of taxonomic categories in biology, the linear structure of political identities in voting systems, or the low-dimensional spatial structure of perceptual categories for faces or colors. With Kemp and also Pat Shafto we have shown how probabilistic models with appropriate abstract structural forms can capture people's property induction judgments in a range of domains (4, 29).

Finally, building on the work above and also inspired by work of Tom Griffiths and colleagues on nonparametric models of human concept learning, we have built nonparametric models for machine concept learning that can learn more complex concepts more accurately than previous Bayesian classification systems (15).

3 Learning causal and social relations

Categories are fundamental to how we organize our knowledge of the world, but they do not exhaust our knowledge. People also deploy much richer systems of knowledge – what cognitive scientists have called *intuitive theories*. To explain how people learn and use intuitive theories, we have taken the technical methods described above for modeling category learning and extended them to work with *relational representations*: representations using predicate logic that describe how entities of one or more types relate to each other. Our focus has been on systems of causal relations and social relations, and on the abstract knowledge that constrains how these systems can be induced from sparse data.

With Tom Griffiths we have shown how to define Bayesian models over predicate logic representations to capture intuitive theories of simple causal systems, and the constraints these place on learning specific causal relations (6, 27, 28). With Charles Kemp, Noah Goodman, Yarden Katz and Tomer Ullman we have shown how to learn abstract knowledge of causality – including abstract theories of specific causal domains as well as more general knowledge of how causality works across domains (9, 10, 21) – by combining both hierarchical Bayes and nonparametric Bayes methods with logical representations. With Liz Bonawitz we have tested a simple version of this approach in experiments with pre-school age children, showing that they can learn appropriate laws and abstract categories for the theory of magnet poles in ways that mirror the historical development of these concepts in science (12).

With Noah Goodman we have modeled how people can ground abstract causal knowledge in perception, learning how to carve up the continuous spatiotemporal flux of perceptual experience into discrete events at the same time as they learn how these events are related causally (16). Specifically, we have shown empirically that human learning of perceptually grounded causal models is best explained, as in our model, in terms of a joint Bayesian inference about both the variable structure and the causal structure relating those variables. With David Wingate and Dan Roy, we have shown how the same ideas can serve as the basis for more sophisticated nonparametric latent-event models of dynamical systems (24).

We have applied analogous methods for hierarchical and nonparametric Bayesian learning over predicate logic representations to model how people learn a wide range of social relations, and most interestingly, how they can learn the abstract form of a relational system that supports generalization about the social relations of new agents from minimal observations (7, 18, 19). For instance, we have modeled how people can learn that a particular social system follows a tree structure, or a ring structure, or a clique structure, and use that knowledge to infer which social relations a new actor will engage in after seeing just a single interaction between that actor and one other person in the network.

Finally, with Dan Roy, we have developed new nonparametric Bayesian machine learning methods for relational data; our methods infer a hierarchical tree structure of latent groups that best explains a whole set of social relations based on possibly different tree-consistent partitions of the objects for each relation (14). Roy and colleagues (chiefly Yee Whye Teh) have extended these approaches to a novel class of nonparametric models known as the Mondrian Process.

PROJECT PUBLICATIONS

Journal Articles

1. Kemp, C., Perfors, A. and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science 10*(3), 307-321.
2. Goodman, N., Tenenbaum, J. B., Griffiths, T. L., and Feldman, J. (2008). A rational analysis of rule-based concept learning. *Cognitive Science 32*:1, 108-154.
3. Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences.* 105(31), 10687-10692.
4. Kemp, C. and Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review 116*(1), 20-58.
5. Frank, M., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20, 578–585.
6. Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review 116,* 661-716.
7. Kemp, C., Tenenbaum, J. B., Griffiths, T. L., and Niyogi, S. (2010). A probabilistic model of theory formation. *Cognition 114*(2), 165-196
8. Perfors, A., Tenenbaum, J. B., and Wonnacott, E. (2010) Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language 37*, 607-642.
9. Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (in press). Learning to learn causal relations. *Cognitive Science.*
10. Goodman, N. D., Ullman, T., and Tenenbaum, J. B. (in press). Learning a theory of causality. *Psychological Review.*
11. Frank, M. C. and Tenenbaum, J. B. (under review). Three ideal observer models for rule learning in simple languages. *Cognition.*
12. Bonawitz, E. B. and Tenenbaum, J. B. (under review). Sticking to the Evidence? A computational and behavioral case study of micro-theory change in the domain of magnetism. *Cognition.*
13. Shafto, P., Kemp, C., Mansinghka, V. K., and Tenenbaum, J. B. (revision under review). Learning cross-cutting systems of categories. *Cognition.*

Papers in Refereed Conference Proceedings

14. Roy, D., Kemp, C., Mansinghka, V. K., and Tenenbaum, J. B. (2007). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems 19.*
15. Mansinghka, V., Roy, D., Rifkin, R., and Tenenbaum, J. B. (2007). AClass: A simple online parallelizable algorithm for probabilistic classification. *AISTATS 2007.*
16. Goodman, N. D.., Tenenbaum, J. B., and Mansinghka, V. K. (2007). Learning grounded causal models. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.*
17. Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2008). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems 20.*
18. Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2008). Learning and using relational theories. *Advances in Neural Information Processing Systems 20.*
19. Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*
20. Piantadosi, S. T., Goodman, N. D., Ellis, B. A., and Tenenbaum, J. B. (2008). A Bayesian Model of the acquisition of compositional semantics. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*
21. Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., and Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*
22. Savova, V., and Tenenbaum, J. B. (2008). A grammar-based approach to visual category learning. *Proceedings*

of the Thirtieth Annual Conference of the Cognitive Science Society.

23. Frank, M., Ichinco, D., and Tenenbaum, J. B. (2008). Principles of generalization for learning sequential structure in language. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*
24. Wingate, D., Goodman, N. D., Roy, D., and Tenenbaum, J. B. (2009). The infinite latent events model. *Uncertainty in Artificial Intelligence (UAI) 2009.*
25. Perfors, A. F., and Tenenbaum, J. B. (2009). Learning to learn categories. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*
26. Savova, V., Jakel, F., and Tenenbaum, J. B. (2009). Grammar-based object representations in a scene parsing task. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*

Book Chapters

27. Tenenbaum, J.B., Griffiths, T. L., and Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik and L. Schulz (eds.), *Causal Learning*. Oxford University Press.
28. Griffiths, T. L. and Tenenbaum, J.B. (2007). Two proposals for causal grammar. In A. Gopnik and L. Schulz (eds.), *Causal Learning*. Oxford University Press.
29. Tenenbaum, J. B., Kemp, C., Shafto, P. (2007). Theory-based Bayesian models for inductive reasoning. In A. Feeney and E. Heit (eds.), *Induction*. Cambridge University Press.
30. Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., & Feldman, J. (2008). Compositionality in rational analysis: Grammar-based induction for concept learning. M. Oaksford and N. Chater (Eds.). *The probabilistic mind: Prospects for rational models of cognition.* Oxford: Oxford University Press.