This is an earlier, submitted version of
Shultz, T. R. (2007). The Bayesian revolution approaches psychological development. *Developmental Science, 10,* 357-364.

The definitive version is available at www.blackwell-synergy.com

Running head: Bayesian Revolution Approaches Development

The Bayesian Revolution Approaches Psychological Development

Thomas R. Shultz

McGill University

Abstract

This commentary reviews five articles that apply Bayesian ideas to psychological development, some with psychology experiments, some with computational modeling, and some with both experiments and modeling. The reviewed work extends the current Bayesian revolution into tasks often studied in children, such as causal learning and word learning, and provides evidence that children's performance can be optimal in a Bayesian sense. There remains much to be done in terms of understanding how representations are created, how development occurs, how Bayesian computation might be neurally implemented, and in reconciling the new work with older evidence that even skilled adults are incompetent Bayesians.

A Bayesian Revolution

A major current revolution in cognitive science concerns the rapid ascendance of Bayesian modeling of probabilistic reasoning (Chater, Tenenbaum, & Yuille, 2006). This collection of papers on child development is an important component of this revolution, but the Bayesian revolution is much more general. Although Bayesian modeling of psychology has been done for years, it seems to have recently surpassed both symbolic and connectionist methods in frequency of conference presentations and publications. An entire special issue of *Trends in Cognitive Sciences* (July 2006) was devoted to probabilistic models of cognition. Bayesian modeling is now being successfully applied to a wide range of problems in cognitive science including sensorimotor control (Körding & Wolpert, 2006), vision (Yuille & Kersten, 2006), conditioning (Courville, Daw, & Touretzky, 2006), induction and inference (Tenenbaum, Griffiths, & Kemp, 2006), and language (Chater & Manning, 2006). In each of these areas, Bayesian models are accounting for (and in some cases predicting) subtle data patterns in a wide variety of psychological experiments.

A general conclusion emerging from this work is that people (and other animals) optimize their performance by conforming to Bayes' rule that specifies how posterior conditional probabilities (of a hypothesis being true, given a data pattern) are computed from the product of the prior probability of the hypothesis and the likelihood of those data given the truth of the hypothesis. This prior x likelihood product is divided by the sum of similar products for all relevant hypotheses. This divisor is a normalizing sum known as the marginal probability of the data; it ensures that the posteriors for all relevant hypotheses sum to 1.0.

Part of the appeal of this approach is that, unlike the emphasis on representational structure in symbolic models and soft statistical constraints in connectionist models (Shultz, 2003), Bayesian approaches emphasize both structure and statistics – essentially by computing statistics over structures, in a way that describes how knowledge is modified by new evidence. I discuss each of the five papers in order from those that emphasize experimental psychology to those that focus on modeling, before addressing some general issues about this approach.

Conditional Intervention

The article by Schulz, Gopnik, and Glymour concerns the conditional intervention principle of causation, which is said to be missing from both mechanistic and covariation accounts of causal reasoning. This principle, common to both Bayesian approaches and the logic of experimental design, is that knowing that event A causes event B implies that intervening on A can change B. In a causal Bayes' network representation, interventions are implemented as additional independent variables that fix the values of certain other variables. Knowing a causal graph structure enables inferences about the effects of such interventions. Or, if the causal graph structure is unknown as in the case of their first experiment, data from interventions enable the learning of this causal structure.

Schulz et al. showed 4- and 5-year-olds causal interventions in the interesting case of interlocking gears, where a gear can be either the cause or the effect of movement of an adjoining gear. Given evidence about which gear was switched on, children distinguished

one causal chain from another and from common-cause (where a switch turns on both gears) and conjunctive (where both a switch and a gear are needed to cause another gear to move) structures. Children made this distinction without any mechanistic cues about which gear was causing the other one to move. Although Schulz et al. note that the common-cause and conjunction problems conflict with normative adult knowledge of gears, their preschoolers did not find these problems any more difficult than conventional three-term causal chains in which a switch activates one gear that, in turn, activates an adjacent gear.

It might prove interesting to study these tasks developmentally to determine how evidence on conditional intervention would interact with knowledge of how gears actually work. Using gears with looser fittings could enable diagnosis of such mechanistic knowledge by permitting notice of which gear starts moving before the other. Based on other evidence of the precedence of understanding of causal mechanisms (Shultz, 1982), such knowledge might alter participants' use of conditional intervention evidence. In particular, knowing how gears actually work could raise suspicions about evidence inconsistent with this knowledge.

In a second experiment, Schulz et al. found that knowledge of causal structure enabled prediction of the effects of interventions. In their third and final experiment, there is some evidence that children could construct their own interventions in play to discover relevant causal structure. Considered together, these experiments contribute to an already extensive line of research suggesting children's use of Bayesian principles in causal inference.

Backward Blocking

In their study, Sobel and Kirkham tested 5-month-old infants in a version of the backward-blocking paradigm. Previously, Sobel and colleagues found support for backward blocking in preschoolers (Sobel, Tenenbaum, & Gopnik, 2004). *Backward blocking* reverses the normal order of events used in the somewhat more familiar *blocking* paradigm, studied in both animal classical-conditioning and human causal-inference literatures. In blocking, two stimuli (causes) are presented together with a reward (effect), but only after an association has already been formed between one of the stimuli on its own and the reward. The previous association between the first-experienced stimulus and the reward blocks an association from being formed between the second stimulus and the reward. As Sobel and Kirkham point out, many conditioning theories, as well as a Bayesian analysis, predict the blocking phenomenon. What many of these conditioning theories have trouble with is backward blocking, in which the order of training experiences is reversed: training with the two stimuli and the reward now precedes training with only one stimulus and the reward. Again, only the stimulus that appears alone with the reward is able to evoke a conditioned response. Backward blocking is particularly interesting within the domains of causal learning and classical conditioning precisely because it is naturally predicted by Bayesian methods and is comparatively awkward for conditioning models.

Sobel and Kirkham framed their experiment even more generally as testing infants' understanding of the Markov assumption, which states that each variable in a causal Bayes' network is independent of all other variables except its effects, conditional on its

direct causes. In the case of backwards blocking and ordinary forward blocking, participants must distinguish an effective (i.e., unblocked) cause from an ineffective (i.e., blocked) cause.

In a subsequent study, Sobel and Kirkham (in press) found evidence for backward blocking in 8-month-olds using an anticipatory eye gaze measure. Those infants were familiarized with two events (A and B) that predicted a third event (C), and then observed that event A alone predicted either event C (backwards-blocking condition) or event D (control condition). When later shown event B, they looked longer at D than at C in the backwards-blocking condition , but longer at C than at D in the control condition, thus conforming to some of the backward-blocking predictions. I would have thought that an idealized prediction would entail no looking difference in the backwards-blocking condition (because the association to the ineffective stimulus is merely blocked, not driven to negative), but the obtained interaction at least suggested discrimination between effective and ineffective stimuli.

The current Sobel and Kirkham experiment extends this same paradigm to even younger, 5-month-old infants. These younger infants looked longer at C than D in the backwards-blocking condition but showed no difference in the control condition. What does all this mean? Sobel and Kirkham tentatively conclude that sensitivity to the Markov assumption develops between 5 and 8 months of age, but I'm not so sure. The 5-month-olds conform to none of the idealized backward-blocking predictions but do show some other systematic looking preferences; 8-month-olds conform to only half of the idealized backward-blocking predictions and show an additional, unpredicted looking preference (for event D) in the backward-blocking condition. Thus, it is not clear that either age group exhibits true backward blocking. Compounding these interpretation difficulties are that many test trials (40%) were excluded from analysis due to poor attention to familiarization events, and the well-known uncertainties about what looking preferences really mean (Cohen, 2004; Haith, 1998). In this context, Sobel and Kirkham's idea to study the infant's own causal interventions, perhaps through sucking rates, seems promising. Because there is evidence that the amount of strength reduction in backward blocking is weaker than the amount of reduction in forward blocking, this may pose a problem for Bayesian and other models that are insensitive to trial orders (Kruschke, in press).

Example Sampling in Word Learning

The Xu and Tenenbaum article presents a fascinating empirical study and model that usefully extends their previous work on word learning. In deciding how to generalize a novel word beyond some given examples, both 4-year-olds and adults were consistent with a Bayesian analysis of the way in which examples were sampled. When three examples of a novel word were generated by a knowledgeable teacher, participants were justified in assuming that these examples represented a random sample from the word's extension, and they consequently restricted their generalization to a specific, subordinate meaning. In contrast, when a similar set of three examples was given but only one of them was a genuinely random instance of the word meaning (the other two having been selected by the learner), they generalized more broadly, to the basic level, just as in previous studies that had provided only one example. Results were strong and accounted for by an explicit Bayesian model. Presumably, the authors argue, neither rational nor

associative theories could predict or explain these results because sampling of examples is not part of those theories.

These results appear to depend on the participants trying to win a prize by picking out two correct examples of the new word, an apparent violation of random sampling because they would choose examples conservatively in order to be correct. An even stronger test of the Bayesian model might involve varying the pragmatic context beyond prize winning to generate a wider range of testable predictions.

The slightly stronger results for adults in the teacher-driven condition only hint at a possible developmental effect, but they do raise the question of whether children are so concerned with sampling issues from the very beginning of word learning. Because of their strength, subtlety, and counter-intuitiveness, these findings are important and provocative for other theoretical approaches.

Learning Overhypotheses

In their modeling paper, Kemp, Perfors, and Tenenbaum argue that inductive learning requires overhypotheses, which they define as abstract knowledge that sets up a hypothesis space at a less abstract level, thus constraining the learner's hypotheses at that lower level. Because their paper presents so many novel and complex ideas, it is perhaps the most difficult of the batch to summarize concisely and evaluate. Asserting that some overhypotheses are innate, the paper focuses on the idea that hierarchical Bayesian models can explain how the rest can be learned. This is illustrated by models that learn overhypotheses about the shape bias in word learning and about bias variations between two different ontological types – objects and substances.

Smith and colleagues had found that shape tends to be homogeneous within object categories (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Given only one exemplar of a novel object category, children extended that category label to similarly shaped objects, rather than to objects with similar texture or color to the exemplar. Even with novel categories, shape was a reliable indicator of category membership. Kemp et al.'s hierarchical models cover the basics of these experiments and generate several interesting predictions, e.g., the optimal number of examples per category is two (assuming a fixed number of total examples), and learning is sometimes faster at higher than lower levels of abstraction, thus explaining why abstract knowledge may appear to be innate even when it is not. This could happen in situations where a child encounters many sparse or noisy observations such that any individual observation may be difficult to interpret, but taken together the observations might support a general conclusion. Mirroring some aspects of other psychological data, a related model learns to choose a shape match for a solid exemplar and a material match for a non-solid exemplar.

There are connectionist models of learning shape bias (Colunga & Smith, 2005; Samuelson, 2002), but Kemp et al. argue that connectionist models operate at an implementation level, whereas Bayesian models are pitched at a computational level (in Marr's terms). Kemp et al. also claim that abstract structures cannot be discerned in connectionist models. This is admittedly more difficult than when abstract structures are pre-specified as in many Bayesian models, but it can be done even with neural networks (Shultz, 2003). Some constraints are built in by the neural modeler, while others can be identified by statistical analysis of a network's knowledge representations. Of course, the

representations so obtained may differ from those expected by researchers, but that is part of the appeal of connectionist modeling – a neural network may have its own way of doing things. In the Bayesian models that I have seen, Bayes' rule does not generate abstract structures – rather it computes statistics over structures designed by the modelers. In contrast, connectionist approaches sometimes show how structures are created, as when a linear structure is created by a network learning a collection of weights of systematically increasing strength (Shultz & Vogel, 2004).

Kemp et al. note that a common reservation about Bayesian models is that their success depends on the modeler's ability to choose the correct prior probabilities. Interestingly, hierarchical models solve this problem in that abstract knowledge need not be specified in advance, but can be learned from data. Another important contribution is that hierarchical models can integrate bottom-up with top-down approaches. Kemp et al. propose to keep adding levels of abstraction until knowledge is simple enough or general enough that it can be plausibly assumed to be innate. This is an interesting proposal that may well hold in some domains, but in some other domains like physics, highly abstract knowledge (e.g., string theory) seems unlikely to be either simple or innate (R. Shiffrin, personal communication).

Kemp et al. acknowledge that the idea of a set of candidate hypotheses being known in advance seems inconsistent with the intuition that the repertoire of a learner can grow over time, as in constructivist approaches. In the static Bayesian models of their paper, the hypotheses are designed by the modelers as probability distributions. Perhaps future Bayesian models can be allowed to grow and thereby expand their hypothesis space at various levels of abstraction, as some connectionist models do (Shultz, 2003). Later, I discuss how this issue is linked to that of creating representations. In any case, this is a fascinating paper that makes contact with several important developmental issues and makes a number of predictions, some of which are supported, some unsupported, and others well worth testing.

Connectionist Causal Inference

The odd-men-out in this group of authors are McClelland and Thompson, who do not offer a Bayesian product but rather a connectionist model of causal inference experiments that were claimed to be out reach for associative models. The psychology experiments in question are those with a *blicket* detector contrasting one vs. two causes (Gopnik, Sobel, Schulz, & Glymour, 2001) and backward blocking vs. screening off (Sobel et al., 2004). Although it is well known that most of these phenomena can be covered by a variety of associative models (Dayan & Abbott, 2001; Read & Montoya, 1999), Bayesian researchers have claimed that the unique ability of Bayes to cover backward blocking implicates the role of Bayesian inference and even a dedicated causal module. By covering the psychological data with a connectionist network, McClelland and Thompson counter both of these claims.

Connectionist networks are based on associative principles, but of course are vastly more powerful than classical associations by virtue of employing large, multilayer topologies containing units with nonlinear activation functions. The authors employ specially designed feedforward networks simulating both slow cortical learning and rapid hippocampal learning. In a pre-experiment cortical-learning phase, the network learns

which object features predict activation of a blicket detector. With those cortical weights frozen, the network then enters a blicket-experiment phase in which particular patterns of covariation between candidate causes and effects are presented.

In a simulation of other psychological work (Schulz & Gopnik, 2004), a similar model employs prior beliefs about context-specific cause-effect relations in the form of physical or psychological causation but overrides them in specific cases when presented with contradictory evidence that events in one domain covary with effects in another domain. In both simulations, McClelland and Thompson use existing knowledge to bias the newer learning that occurs in the typical causal inference experiment.

This work is more at the implementation level than are Bayesian approaches and is inspired by neuroscience evidence of slow cortical and fast hippocampal learning (McClelland, McNaughton, & O'Reilly, 1995). Although these networks are not specially designed for causal learning and inference, they are dedicated to using existing knowledge in new learning, a powerful and understudied idea. Other knowledge-based connectionist systems, that are less specialized and based on constructivist principles, recruit existing knowledge and adapt it to learn new tasks (Shultz & Rivest, 2001).

Future work along this promising line might more strongly justify the nature of the pre-experiment training, run multiple networks to gage their variability, and analyze network knowledge representations to provide additional insight into their solutions. Even if connectionist models can simulate such phenomena, it is noteworthy that these phenomena were first identified and predicted with Bayesian approaches. This underscores that Bayesian ideas were particularly useful in initiating and directing this research. Nonetheless, it will eventually be important to understand how brains implement Bayesian learning and inference.

How Can Incompetent Bayesians be Bayesian?

The resurgence of interest in Bayesian methods is somewhat surprising given the Nobel-Prize-winning work showing that people are rather poor Bayesians, subject to such biases as the base-rate fallacy and the representativeness heuristic (Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996; Tversky & Kahneman, 1974, 1981). There is also evidence that people confuse the direction of conditional probabilities, e.g., the probability of a symptom given a disease vs. the probability of a disease given a symptom (Eddy, 1982; Gluck & Bower, 1988). Even experienced medical professionals deviate from Bayes in these ways, creating medical inefficiencies and sometimes disastrous outcomes.

Clearly, the new evidence that people are Bayesian optimizers needs to be reconciled with this older work suggesting that people routinely ignore prior probabilities and confuse the direction of conditional probabilities. One hypothesis currently being floated to explain this discrepancy is that people are implicit, but not explicit, Bayesians (Chater et al., 2006). This seems a bit implausible given that the implicit vs. explicit distinction can be rather vague and may not make much of a psychological difference in any case. It also seems to conflict with results showing that prior probabilities are more likely to be used if made more explicit (Bar-Hillel & Fischoff, 1981; Fischoff, Slovic, & Lichtenstein, 1979; Gigerenzer, Hell, & Blank, 1988).

Interestingly, however, the implicit-explicit hypothesis does seem to be testable. One could, for example, design psychology experiments that a) take a modern experiment showing implicit conformity to Bayes' rule and add a condition in which the numeric features are explicit, or b) take an older experiment showing explicit deviations from Bayesian rationality and add a condition in which the numeric features are implicit. According to the implicit-explicit hypothesis, such new conditions should reverse the usual trends. One experiment happened to follow strategy *b* by requiring that participants learn a probabilistic diagnosis problem from examples, as opposed to being given summary probabilities; there was still evidence of neglecting priors and confusing the direction of conditionals (Gluck & Bower, 1988).

Another, related hypothesis is that what disrupts explicit Bayesian reasoning is the very use of probabilities in problem descriptions. There is evidence that people did considerably better on otherwise explicit uncertainty problems if the numerical information was presented in terms of frequencies, rather than probabilities, although participants still didn't come very close to Bayesian norms (Chase, Hertwig, & Gigerenzer, 1998; Gigerenzer & Hoffrage, 1995; Gigerenzer & Todd, 1999).

Still other explanations for these discrepancies may be worth considering. One is that the task is often quite different in experiments that show Bayesian failures than in experiments that show Bayesian successes (Y. Takane, personal communication). In typical *failure* experiments, participants largely unfamiliar with Bayes' rule are given priors and likelihoods and asked for posteriors, a difficult computation to perform mentally. In typical *success* experiments, participants perform some other task, such as causal learning or word learning. Then a skilled modeler typically tries to match observed posterior distributions by constructing plausible priors and likelihoods that fit together in Bayes' rule. If asked to predict the better Bayesian, I would bet on the modeler. The nature of the actual computations performed by participants in the success experiments is still unknown.

Model Comparisons

There is little doubt that Bayesian approaches are making rapid progress in the understanding of a wide range of psychological phenomena. There is, however, resistance to the claim that Bayesian models uniquely account for the phenomena that they predict and explain. Theorists are naturally inclined to make such uniqueness claims because, if true, they would enhance the value of a model. But Newell (1990) cautioned that models can at best provide sufficient but never necessary explanations, because a better model can be expected to appear at some future time. Still, at any particular moment, there can

be a strong urge to demonstrate that one's model is the only extant one to explain something. And theoretical competitors often eagerly respond to such challenges. The principal responses to Bayesian challenges are currently coming from connectionists, as well represented here by the McClelland and Thompson article.

In theoretical terms, the prospects of connectionist implementation of computational-level Bayesian ideas seem good. Due to a mathematical equivalence between Bayes' rule and commonly used neural activation functions, each neural unit can be seen as computing a posterior probability value, under certain assumptions (Jordan, 1995; McClelland, 1998). Moreover at higher levels, competitive-learning networks select the most probable of several mutually exclusive hypotheses, and constraint-satisfaction networks compute the most probable configuration of hypotheses by increasing goodness or equivalently by decreasing energy (McClelland, 1998). Three-layer feedforward networks trained with backpropagation of error have been shown to approximate Bayesian decision making that minimizes the probability of a decision error. If the decision is relatively easy, representations on the hidden layer reflect the Bayesian posterior probability distribution (Asoh & Otsu, 1989). So-called Boltzmann machine networks, although somewhat neglected because of their slowness, learn a range of outputs for a given input such that the probability of each output matches the probability of that output in the environment (Ackley, Hinton, & Sejnowski, 1985). Boltzmann machines are stochastic networks with bi-directional weights and binary units that turn on with a probability equal to a logistic function of the states of their inputs and the connecting weights. A faster-learning, feedforward version of stochastic networks has been explored (Neal, 1992), and the McClelland and Thompson work shows evidence of probability matching by a deterministic feedforward network. There was also a proposal to interpret network input and output values as probabilities and adjust network weights to maximize log likelihood rather than minimizing error (Baum & Wilczek, 1988).

If connectionist systems do function at a lower, implementation level compared to Bayesian systems, it is likely that the two theoretical approaches can usefully coexist and enhance each others' explanatory accounts. Bayesian analyses could explore the computational requirements of a task and identify optimal solutions, whereas connectionist models could show how such computations might be accomplished in neural tissue. But if the longstanding debate on symbolic vs. connectionist approaches is taken as a guideline, we can also expect that Bayesian and connectionist models could, in some cases, generate different predictions for particular experiments. Among the theoretical differences that could create differential predictions are that non-optimal solutions can emerge from neural networks, and that Bayesian networks have a particular facility for making inferences in both directions (forward and backward) that unidirectional neural networks lack.

As this literature advances, we might expect renewed efforts to design neural networks that can make bidirectional inferences based on what they learn. For example, Hinton and colleagues (Hinton, Osindero, & Teh, 2006; Hinton & Salakhutdinov, 2006) introduced generative networks called restricted Boltzmann machines that, when chained together, enable bidirectional inferences via both top-down and bottom-up weights.

Thus relations between Bayesian and connectionist systems are varied, deep, and deserving of further exploration. Neural networks seem to be good candidates for models

that could identify the circumstances under which performance can be Bayesian optimal and how brains might achieve this. Notwithstanding these implementation concerns, the demonstrated heuristic advantage of also working at the higher, computational level with Bayesian ideas should not be underestimated.

What about Development?

It is clear from this collection of papers that Bayesian approaches are addressing issues and tasks found in studies with children, including learning about causes and effects and about words and concepts. What is less clear is whether this work is currently addressing developmental issues per se. Most of the simulations and covered experiments identify nearly optimal behavior at every age tested, and the one attempt to examine development (Sobel & Kirkham) yielded somewhat ambiguous psychological data.

Perhaps an interesting question is whether children's performance becomes more optimal with development. It is difficult to imagine that they (or any other organisms) are always optimal in every domain. Unless, of course, Bayesian modelers choose to adopt a nativist stance that everything is optimal from the start. There may be a natural tension between developmental change and many current approaches to Bayesian modeling that design priors and likelihoods to fit observed posteriors. If a fit can be designed, then the hypothesis that children are little Bayesians is seen as confirmed. But it might be interesting to view development as a more serious modeling challenge.

Assuming for the moment that Bayes' rule itself does not develop, there would seem to be two Bayesian mechanisms affording developmental possibilities: the learning of priors and likelihoods, and the creation and selection of Bayesian models. Within a particular Bayesian model, performance could improve with better estimates of priors and likelihoods, and such improvements might explain some quantitative aspects of development. But if a Bayesian model is inappropriate, then no amount of parameter fitting would produce optimal performance. In such cases, qualitatively different models would need to be created and the best one selected. Although model selection might result from Bayesian inference, fully automatic creation of models, whether Bayes networks or hypotheses for new levels in a hierarchical model, seems beyond the ability of Bayes' rule itself, which is why hypotheses and structures are designed by the researcher in current Bayesian models. Making these complimentary processes of parameter fitting and model creation more fully automatic and relating them to developmental change could make an important contribution.

Interestingly, probability estimation and model building seem analogous to quantitative and qualitative developmental mechanisms, respectively, in constructive neural networks. Constructive networks start with minimal structure and recruit single hidden units or previously-learned networks as needed to reduce network error (Shultz, 2003). Change occurs through quantitative adjustment of connection weights within a particular network structure or, when that fails to solve the current problem, through qualitative recruitment of additional computational devices. This recurring cycle of adjustments and recruitments creates novel representational structures that the network could not previously express. Perhaps there are computational lessons in these mechanisms that could be of use to Bayesian modelers of psychological development.

Conclusion

The rule of Rev. Bayes is already being used to tell a very interesting story about cognition, perception, language, and action. But it is not the whole story, particularly when it comes to complex representations, development, and brain implementation. What will be especially interesting is to see how Bayesian approaches can be integrated with advances in knowledge representation, artificial neural networks, and neuroscience.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147-169.

Asoh, H., & Otsu, N. (1989). Nonlinear data analysis and multilayer perceptrons. In *Proceedings of the International Joint Conference on Neural Networks* (Vol. 2, pp. 411-415). San Diego.

Bar-Hillel, M., & Fischoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology, 41,* 671-680.

Baum, E. B., & Wilczek, F. (1988). Supervised learning of probability distributions by neural networks. In D. Andersen (Ed.), *Neural information processing systems* (pp. 52-61). Melville, New York: American Institute of Physics.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences, 2,* 206-214.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10,* 335-344.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10,* 287-291.

Cohen, L. B. (2004). Uses and misuses of habituation and related preference paradigms. *Infant and Child Development, 13,* 349-352.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review, 112,* 347-382.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294-300.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge: Cambridge University Press.

Fischoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational behavior and human decision processes, 23,* 339-359.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: General, 14,* 513-525.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gluck, M. A., & Bower, G. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Gopnik, A., Sobel, D., Schulz, L., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37,* 620-629.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development, 21,* 167-179.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18,* 1527-1554.

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

Jordan, M. I. (1995). *Why the logistic function? A tutorial discussion on probabilities and neural networks* (Computational Cognitive Science Technical Report No. 9503): MIT.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582-591.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10,* 319-326.

Kruschke, J. K. (in press). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*.

McClelland, J. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21-53). Oxford: Oxford University Press.

McClelland, J., McNaughton, J., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419-457.

Neal, R. (1992). Connectionist learning of belief networks. *Artificial Intelligence, 56,* 71-113.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal learning and causal reasoning. *Journal of Personality and Social Psychology, 76,* 728-742.

Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20 month olds. *Developmental Psychology, 38*, 1016-1037.

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*, 162-176.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1, Serial No. 194).

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Shultz, T. R., & Rivest, F. (2001). Knowledge-based cascade-correlation: Using knowledge to speed learning. *Connection Science, 13*, 1-30.

Shultz, T. R., & Vogel, A. (2004). A connectionist model of the development of transitivity. In *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society* (pp. 1243-1248). Mahwah, NJ: Erlbaum.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*, 13-19.

Sobel, D. M., & Kirkham, N. Z. (in press). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303-333.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*, 309-318.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*, 301-308.

Acknowledgements

Thanks to the Natural Sciences and Engineering Research Council of Canada for continued financial support and to Yoshio Takane for helpful comments on a previous draft.