# Amortized Inference in Probabilistic Reasoning

**Samuel J. Gershman**[1] **(sjgershm@mit.edu)** and **Noah D. Goodman**[2] **(ngoodman@stanford.edu)**
[1]Department of Brain and Cognitive Sciences, MIT
[2]Department of Psychology, Stanford University

**Abstract**

Recent studies of probabilistic reasoning have postulated general-purpose inference algorithms that can be used to answer arbitrary queries. These algorithms are memoryless, in the sense that each query is processed independently, without reuse of earlier computation. We argue that the brain operates in the setting of amortized inference, where numerous related queries must be answered (e.g., recognizing a scene from multiple viewpoints); in this setting, memoryless algorithms can be computationally wasteful. We propose a simple form of flexible reuse, according to which shared inferences are cached and composed together to answer new queries. We present experimental evidence that humans exploit this form of reuse: the answer to a complex query can be systematically predicted from a person's response to a simpler query if the simpler query was presented first and entails a sub-inference (i.e., a sub-component of the more complex query). People are also faster at answering a complex query when it is preceded by a sub-inference. Our results suggest that the astonishing efficiency of human probabilistic reasoning may be supported by interactions between inference and memory.

**Keywords:** induction, Bayesian inference, memory

"*Cognition is recognition*." – Hofstadter (1995)

## Introduction

One view of probabilistic reasoning holds that our brains are equipped with general-purpose inference algorithms that can be used to answer arbitrary queries (Griffiths et al., 2012; Pouget et al., 2013). An under-appreciated property of such algorithms borrowed from computer science is that they are *memoryless*: each query is (at least in principle) processed independently of others. While this property guarantees that inferences will not interfere with one another, it can also lead to gross computational inefficiency, since inferences are never reused; memorylessness implies that answering the same query twice requires the same amount of computation as answer two unique queries.[1]

Whatever inference algorithms the brain uses, they are unlikely to be memoryless. Consider, for example, the image in Figure 1 (Gregory, 1970). Upon viewing it for the first time, most observers find it extremely difficult to identify what the image depicts.[2] However, once the image has been deciphered, all subsequent views are instantly recognized. Clearly, the visual system is not running a computationally expensive inference algorithm upon each viewing; the inference is simply reused.

In reality, it is rare to be faced with the exact same query multiple times. Much more pervasive is the appearance of similar or related queries. For example, as you view an image, your head and eyes are continuously moving, generating an infinitude of slightly different queries. For these queries, it may be inaccurate to reuse a stored inference without modification. This raises the problem of *amortized inference*: how to flexibly reuse inferences so as to answer a variety of related queries. Recently, Stuhlmüller et al. (2013) addressed this problem by using stored samples to estimate local conditional distributions, and then approximating answers to more complex queries by composing the local distributions. The work described in this paper seeks experimental evidence for a similar kind of flexible reuse in human reasoning.

We presented subjects with a simple Bayesian network and asked them to answer a series of queries about it. One of these queries (the "target") could be answered by reusing the answer to another query (the "sub-query"). We hypothesized that the effects of reuse would be evident compared to an inference with the same structure but no re-usable sub-query. Further, we hypothesized that this effect would be present only if the target was presented after the sub-query. Accordingly, we manipulated (between subjects) whether the target came before or after the sub-query. This design allowed us to look for two key signatures of reuse: correlations between related inferences (Experiment 1) and faster responses for inferences that exploit reuse (Experiment 2).

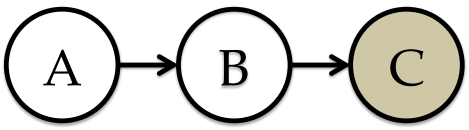
Figure 1: What does this image depict?

## Amortized inference in Bayesian networks

In this paper, we will restrict our attention to amortized inference for Bayesian networks. Let $p(x)$ denote a probability distribution on variables $x = \{x_1, \ldots, x_M\}$. A Bayesian network $G$ is a directed acyclic graph with nodes corresponding

[1]To be fair, inference algorithms for dynamical systems, like Kalman filtering, involve reuse in a certain sense. However, these algorithms are not designed to reuse inferences when applied to several independent time series (even if the time series are identical).

[2]Answer: a dalmatian.




Query 1: $P(B|C) = P(C|B)P(B)/P(C)$

Query 2: $P(A|C) = \sum_B P(A|B)P(B|C)$

Figure 2: A Bayesian network in which variable A is observed. Query 1 is a sub-query of Query 2. The reused conditional distribution is shown in blue.

to variables and edges corresponding to probabilistic dependencies. The graph expresses a factorization of $p(x)$ into a product of conditional distributions:

$$p(x) = \prod_{m=1}^{M} p(x_m|\text{pa}_G(x_m)), \quad (1)$$

where $\text{pa}_G(x_m)$ is the set of parents of node $m$ in graph $G$. A query $q$ is a tuple $\{Q, \mathcal{E}, y\}$, where $Q \subseteq \{1,\ldots,M\}$ denotes a set of unobserved (latent) variables and $\mathcal{E} \subseteq \{1,\ldots,M\} \setminus Q$ denotes a set of observed variables with values $y$. An inference algorithm is any function that takes as input $q$ and returns the conditional distribution $p(x_Q|x_\mathcal{E} = y)$. Many algorithms are available for this task, such as belief propagation, Markov chain Monte Carlo, and importance sampling (Koller & Friedman, 2009).

Almost all widely used inference algorithms are memoryless: their operation does not depend on a memory trace of past inferences. In contrast, we will consider amortized inference algorithms that reuse past inferences. One simple form of flexible reuse is caching (or, in computer science lingo, "memoizing"; Michie, 1968) sub-computations that are invoked by multiple queries. A simple example is shown in Figure 2, where the conditional distribution computed for Query 1 can be reused to answer Query 2. We refer to Query 1 as a "sub-query" of Query 2, and its corresponding inference as a "sub-inference."

Memoized reuse has a number of implications for human reasoning, which we test in the experiments reported below. Let us imagine a simple query that is presented earlier than, and entails sub-computations of, a more complex query. The most immediate implication of caching (tested in Experiment 2) is that answers to the more complex queries should be faster compared to similar queries where reuse is unavailable, since retrieval of an inference is presumably faster than computing the inference from scratch. A second implication (tested in Experiments 1 and 2) is that variation in answers for the complex query should be systematically predictable from the corresponding answers to the simpler query. In other words, individual differences in the answer to a simple query should propagate to more complex queries, under conditions


Figure 3: Bayesian network presented to subjects.

where the computations for the simple query can be reused.

We should note an important subtlety to this argument: at least in principle, reuse could still occur if the complex query precedes the sub-query, since the same sub-computations may be invoked regardless of order. However, a complex query can be answered in a number of different ways, which may or may not invoke the same sub-computations as the sub-query. We conjecture that query order biases the complex query to be answered in different ways based on what sub-computations are available for reuse. This conjecture is consistent with our finding below of order effects in correlations and prediction errors (Experiment 1), but more research is needed to understand this issue completely.

## Experiment 1

We presented subjects with a sequence of queries about a Bayesian network representing a hierarchy of military officers (Figure 3). Subjects were told that the enemy was attempting to bribe officers, and that their job was to infer whether a particular officer would defect based on information about the defection of other officers. We constructed the queries such that one query (the "target") was a sub-query of another. Thus, we provided subjects with the opportunity to reuse their sub-inference. For comparison, we had a separate group of subjects answer the same queries, but in this case the sub-query was presented *after* the target query. We hypothesized that when the sub-query could be reused, individual responses for the sub-query could be used to predict responses to the larger query—that is, the two queries would be non-independent.

### Methods

**Subjects**. 146 subjects (73 in each condition) were recruited through Amazon Mechanical Turk and paid $0.50.

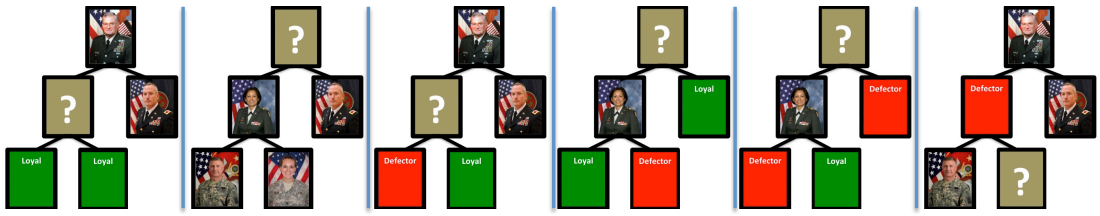



Figure 4: Experiment 1 design. Each query (labeled A-F) is shown, along with the serial positions for each condition (S1 and S2). Subjects in both experiments were assigned to either S1 or S2. On each trial, subjects were asked to make a guess about whether the queried officer (indicated by a question mark) would defect or stay loyal. Observed variables are shown by colored squares (green = loyal, red = defect).

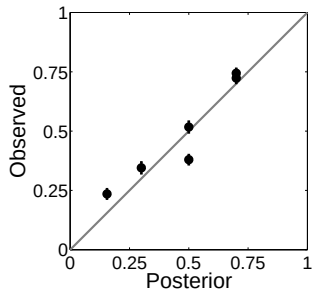



Figure 5: Average observed probabilities in Experiment 1 plotted against the true posterior probabilities. Error bars show standard error of the mean.

**Procedure**. At the beginning of the task subjects were shown a hierarchy of military officers (Figure 3), and told that the enemy was attempting to bribe the officers; the probability that an officer defects is 0.7 if his/her superior defects, and 0.3 if his/her superior remains loyal. In addition, the general at the top of the hierarchy defects with probability 0.5. On each trial, subjects were presented with one of 6 queries, shown in Figure 4, and asked to make a binary guess about whether the queried officer defected or not, followed by a confidence rating using a slider bar. The set of queries were shown in two orders: subjects in condition S1 saw the order A, B, C, D, E, F, and subjects in condition S2 saw the order A, B, F, D, E, C. The only difference between these conditions is that the serial positions of C and F are swapped. The main queries of interest were C, D and E (as we explain below); the other queries were included to help ascertain how well calibrated subjects' responses were with the true posterior probabilities.

## Results

Binary judgements with their confidence ratings were converted to probabilities by linearly rescaling the confidence, such that minimum confidence is mapped to probability 0.5, using the choice to determine if the probability was above or below 0.5. Because this mapping might not correspond to subjective probabilities, we first sought to confirm that this mapping is well-calibrated with the true posterior probabilities. Figure 5 shows the "observed" probabilities (i.e., converted confidence ratings) plotted against the posterior probabilities predicted by from the Bayesian network via Bayes' rule, collapsing across conditions. While some systematic deviations are evident, the two sets of probabilities are strongly correlated ($r = 0.92, p < 0.01$). Thus, we can reasonably use these observed probabilities as proxies for subjects' inferences.

If subjects in condition S1 are reusing their inference for Query C to answer Query E, then we should be able to systematically predict their answers to Query E from their answers to Query C. We tested this hypothesis by plugging each subject's answer to Query C into the computation of Query E using probability theory. In detail, take the natural (forward) decomposition of the Bayes net in Figure 3 to be

$$P(X,Y,M,Z,L) = P(X|M)P(Y|M)P(M|L)P(Z|L)P(L), \quad (2)$$

where $X$ and $Y$ are the two leaf nodes, $M$ and $Z$ are the two middle nodes, and $L$ is the root node. Then Query C is $P(M|X,Y)$ and Query F is $P(L|X,Y,Z)$. The two are related by:

$$\begin{aligned} P(L|X,Y,Z) &= \sum_M P(L,M|X,Y,Z) \\ &= \sum_M \frac{P(L|M,Z)\color{blue}{P(M|X,Y)}P(Z|M)}{P(Z|X,Y)}, \end{aligned} \quad (3)$$

where the reused computation is shown in blue. We compared this predicted probability to the reported probability for Query E. As shown in Figure 6A, the observed and predicted probabilities were significantly correlated in condition S1 ($r = 0.28, p < 0.05$), but not in condition S2 ($r = -0.11, p = 0.36$). Further corroborating our hypothesis, we found that

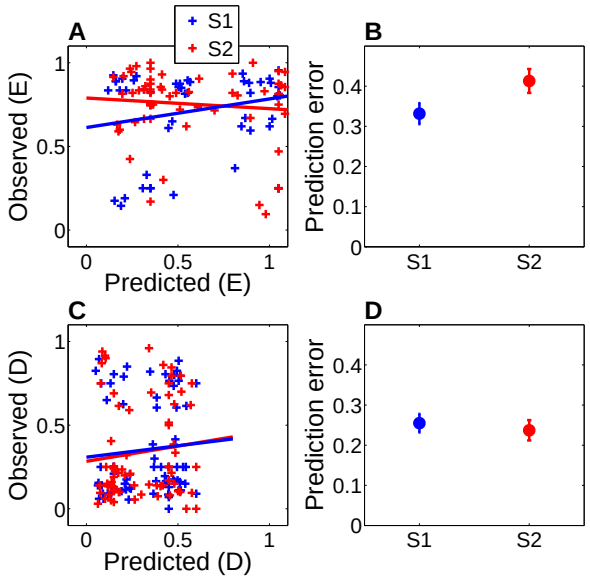



Figure 6: Results of Experiment 1. (*A*) Relationship between observed and predicted probabilities for Query E, where the predictions are computed by plugging each subject's answer to Query C into Bayes' rule. Least-squares lines are superimposed on individual data points. (*B*) Average prediction error, defined as the absolute difference between the predicted and reported probabilities for Query E. Error bars show standard error of the mean. (*C* and *D*) Same format as *A* and *B*, but for Query D instead of Query E.

the absolute difference between the reported and predicted probabilities (the prediction error; Figure 6B) was significantly higher in S2 compared to S1 $[t(144) = 2.06, p < 0.05]$.

As a control, we repeated the same analysis using Query D instead of Query E. Subjects are never shown a sub-query of Query D, and hence we do not expect any reuse. Nonetheless, we can still plug the answer to Query C into the inference for Query D, since the conditional distribution on the left branch of the Bayesian network is the same in both queries (due to symmetries in the probabilistic model). The results of this control analysis are shown in panels C and D of Figure 6. There was no correlation between the observed and predicted probabilities for either condition (both $p > 0.05$), and the prediction errors did not differ significantly ($p = 0.6$). Thus the order manipulation specifically affects the relationship between Query C and Query E.

Our data do not constrain hypotheses about particular inference algorithms used for the individual queries except in requiring them to decompose into sub-queries (a condition on re-use) and to have an element of stochasticity or individual variation (a necessity for inducing the reported correlation). To show that our results are indeed expected from such algorithms, we provide here one concrete example of how a sampling algorithm (see Griffiths et al., 2012; Vul et al., in press) can give rise to the correlations observed in this experiment. Imagine an approximate inference system that answers queries by generating samples from the desired conditional distribution. In the setting of the defection task, where all the variables are binary, the average of these samples is a Monte Carlo estimate of the conditional probability. Reuse in this system is implemented by caching and retrieving these Monte Carlo estimates. Figure 7 shows the results of simulating this system (with a small amount of noise corrupting the Monte Carlo estimates) on the same queries given to subjects, and performing the same correlation analyses on the simulated inferences. The Y-axis shows the correlation between the predicted and observed inferences, demonstrating that for Query E reuse induces a strong correlation, whereas for Query D the lack of reuse leads to a correlation of 0. Note that even the very small fluctuations in Monte Carlo estimates based on dozens of samples are enough to induce a strong correlation (though these small fluctuations could be quickly swamped by independent sources of response noise).

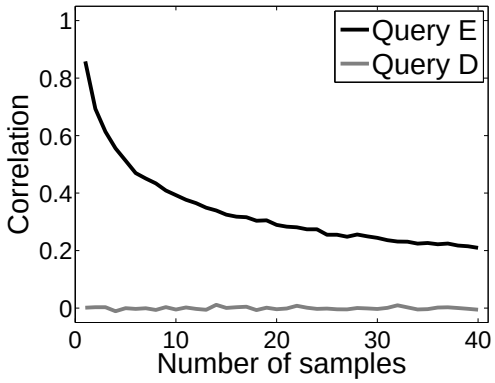



Figure 7: Simulation of reuse in a sample-based approximate inference system. See text for details.

In summary, we found that answers to complex queries are predictable from answers to sub-queries, but this predictability only occurs under specific circumstances that increase the likelihood of reuse (i.e., when a complex query is preceded by a sub-query).

## Experiment 2

In Experiment 2, we tested the prediction that reuse should lead to faster inferences. We used the same query orders as in Experiment 1, but we no longer asked subjects to make a confidence rating; instead, subjects made a speeded binary choice using the keyboard.

### Methods

**Subjects**. 134 subjects were recruited through Amazon Mechanical Turk and paid $0.50. We only analyzed data from subjects who chose the more likely hypothesis (defect) for Query E, resulting in 50 subjects in condition S1 and 53 subjects in condition S2.

**Procedure**. The procedure was identical to the procedure

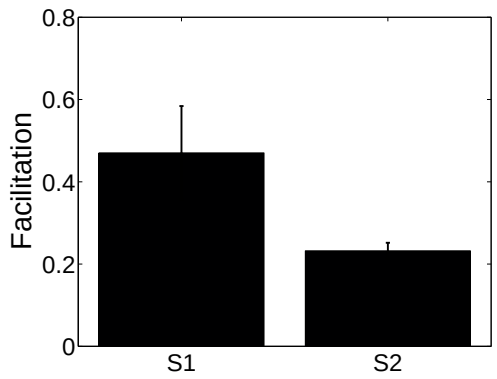



Figure 8: Median facilitation effect in Experiment 2.

used in Experiment 2, except that subjects were asked to make speeded binary responses (defect/loyal) using the keyboard.

### Results

The results of Experiments 1, as well as the experimental design, suggest that one condition (S1) provided opportunity for reuse of a sub-inference, and the other condition (S2), did not. In particular, the sub-query C was presented before F in S1, but after F in S2. To quantify the advantage of reuse in S1, we computed a "facilitation effect," defined as $\log \mathrm{RT}_D - \log \mathrm{RT}_F$, where $\mathrm{RT}_F$ is the response time (in milliseconds) for Query E. We use the RT for Query D as a baseline, since it has the same underlying structure as Query E, but lacks a sub-query among the other trials. A larger facilitation effect means that responses are faster for the target query relative to the baseline.

To reduce sensitivity to outliers, we compared the median (rather than the mean) facilitation effect in the two conditions, as shown in Figure 8. Consistent with our hypothesis, the facilitation effect was larger in condition S1 compared to S2 ($z = 2.09, p < 0.05$, rank sum test), indicating a speed advantage when reuse is available. Furthermore, the median facilitation effect was significantly greater than 0 in S1 ($z = 3.68, p < 0.005$, signed rank test), but not in S2 ($p = 0.16$, signed rank test). We conclude that the availability of reuse facilitates the speed of inference.

## Discussion

Most algorithms for probabilistic inference assume that queries are processed independently. Our experiments show that this assumption is incorrect as a description of human reasoning: Multiple related queries are *not* processed independently. Specifically, queries influence each other when the answer to one query supplies a memoizable sub-computation for another query. Experiment 1 showed that the answer to a complex query can be systematically predicted from a person's response to a simpler query if the simpler query was presented first and entails a sub-computation of the more complex query. Experiment 2 showed that the same conditions lead to faster inferences for complex queries, consistent with the idea that retrieving an inference from memory is faster than recomputing it.

Our results place new constraints on rational process models of cognition (see Griffiths et al., 2012, for a review). In particular, these models need to be augmented with storage and retrieval mechanisms for reusing inferences. However, we are still far from a detailed computational understanding of amortized inference in human reasoning. One open question is the inference algorithms people use even for isolated inferences in tasks like those in our experiments. Currently one of the most promising hypotheses is that the brain uses some form of sample-based approximate inference, such as Markov chain Monte Carlo (MCMC; Gershman et al., 2012; Lieder et al., 2012) or importance sampling (Shi et al., 2010). We showed that a simple sample-based algorithm augmented with sample reuse can give rise to the observed correlation and RT effects. Stuhlmüller et al. (2013) proposed a more sophisticated framework for reusing samples within MCMC. More research will be required to directly investigate the inference algorithms and methods of reuse in human reasoning.

A key challenge going forward will be determining the flexibility of inference reuse—when can stored inferences be used compositionally to answer larger questions? It will require more complex probabilistic models than the simple Bayesian network we used in our experiments to investigate this question. However, training people on complex Bayesian networks is difficult; it is well known that people show systematic biases in their interpretation of probabilities (Kahneman & Tversky, 1982), and complex networks may also tax working memory. One alternative would be to present the network in a frequency format by generating samples. Another alternative would be to exploit complex probabilistic models that people have already learned, such as intuitive physics (Battaglia et al., 2013).

More complex models also raise the question of which inferences to store, since the memory cost of storing all inferences may be prohibitive. A number of trade-offs are involved: storing more complex inferences provides greater savings in computation time, but incurs a larger memory cost; complex inferences should not be stored if they can be decomposed into simpler inferences that are already stored, and conversely simple inferences should be stored if they can be composed into larger inferences; storage of frequent inferences should be preferred to storage of rare inferences. Optimally balancing these intuitive trade-offs is subtle; it may addressable via resource-rational analysis, which dictates how the cost of computation is balanced against the accuracy of inference (Howes et al., 2009; Vul et al., in press).

Memory mechanisms have figured prominently in exemplar models of inductive reasoning (e.g., Dougherty et al., 1999; Estes, 1994; Heit, 1992; Heit & Hayes, 2011; Juslin & Persson, 2002), which assert that inductive judgments are formed by taking a similarity-weighted average of exemplars

stored in memory. These models draw support from experiments showing correlations between measures of reasoning and memory (Hayes & Heit, 2013; Heit & Hayes, 2011). It has even been suggested that exemplar models may provide a general method for performing probabilistic inference (Shi et al., 2010), based on the idea that exemplars correspond to samples from a generative model. The framework of amortized inference makes rather different claims about the role of memory in inductive reasoning: *inferences* (rather than stimulus exemplars or samples from the prior) are stored in memory and reused in sophisticated ways. For example, stored inferences may be composed together, along with freshly computed inferences, to answer a more complex query (it is also possible to construct memory-based inference systems that lack this form of compositionality).

Amortized inference also suggests a number of ways in which ideas from memory research can be applied to inductive reasoning (see Heit & Hayes, 2011, for another perspective). For example, do stored inferences interfere with each other proactively and retroactively? Is there a temporal gradient—analogous to a forgetting function—such that older inferences are less likely to be reused? Can we prime particular inferences using contextual reminders?

Our experiments provide initial evidence that human inference is an active and ongoing process of reuse and recombination, jointly solving myriad related questions over time—*amortized inference*. Thus the astonishing efficiency of human probabilistic reasoning may be explained partly by interactions between inference and memory.

## Acknowledgments

We thank Andreas Stuhlmüller for helpful discussions. This work was supported by the MIT Intelligence Initiative, a John S. McDonnell Foundation Scholar Award, grants from the ONR, the IARPA ICARUS program, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

## References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327–18332.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180–209.

Estes, W. K. (1994). *Classification and Cognition*. Oxford University Press.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

Gregory, R. L. (1970). *The Intelligent Eye*. Oxford University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Hayes, B. K., & Heit, E. (2013). How similar are recognition memory and inductive reasoning? *Memory & Cognition*, 1–15.

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology*, *24*, 341–380.

Heit, E., & Hayes, B. K. (2011). Predicting reasoning from memory. *Journal of Experimental Psychology: General*, *140*, 76.

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological Review*, *116*, 717.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123–141.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, 2699–2707.

Michie, D. (1968). Memo functions and machine learning. *Nature*, *218*, 19–22.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*, 1170–1178.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443–464.

Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. In *Advances in neural information processing systems* (pp. 3048–3056).

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (in press). One and done? optimal decisions from very few samples. *Cognitive Science*.