

Algorithmic approaches to ecological rationality in humans and machines

A DISSERTATION PRESENTED
BY
ISHITA DASGUPTA
TO
THE DEPARTMENT OF PHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2020

©2014 – ISHITA DASGUPTA
ALL RIGHTS RESERVED.

Algorithmic approaches to ecological rationality in humans and machines

ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetur. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetur erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dic-

Thesis advisor: Professor Samuel J. Gershman

Ishita Dasgupta

tum elit, vel lacinia enim metus eu nunc.

Contents

THIS IS THE DEDICATION.

Acknowledgments

LOREM IPSUM DOLOR SIT AMET, consectetur adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetur. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

0

Introduction

This is the Introduction – to be written.

*Nulla facilisi. In vel sem. Morbi id urna in diam dignis-
sim feugiat. Proin molestie tortor eu velit. Aliquam erat
volutpat. Nullam ultrices, diam tempus vulputate egestas,
eros pede varius leo.*

Quoteauthor Lastname

1

Resource-limited approximation at run-time

In his preface to *Astronomia Nova* (1609), Johannes Kepler described how he struggled to find an accurate mathematical description of planetary motion. Like most of his contemporaries, he started with the hypothesis that planets move in perfect circles. This necessitated extraordinary labor to

reconcile the equations of motion with his other assumptions, “because I had bound them to millstones (as it were) of circularity, under the spell of common opinion.” It was not the case that Kepler simply favored circles over ellipses (which he ultimately accepted), since he considered several other alternatives prior to ellipses. Kepler’s problem was that he failed to generate the right hypothesis.¹

Kepler is not alone: the history of science is replete with examples of “unconceived alternatives”², and many psychological biases can be traced to failures of hypothesis generation, as we discuss below. In this paper, we focus on hypothesis generation in the extensively studied domain of probabilistic inference. The generated hypothesis are a subset of a tremendously large space of possibilities. Our goal is to understand how humans generate that subset.

In general, probabilistic inference is comprised of two steps: hypothesis generation and hypothesis evaluation with feedback between these two processes. Given a complete set of hypotheses \mathcal{H} and observed data d , optimal evaluation is prescribed by Bayes’ rule, which assigns a posterior probability $P(b|d)$ to each hypothesis $b \in \mathcal{H}$ proportional to its prior probability $P(b)$ and the likelihood of the observed data under b , $P(d|b)$:

$$P(b|d) = \frac{P(d|b)P(b)}{\sum_{b' \in \mathcal{H}} P(d|b')P(b')}. \quad (1.1)$$

Many studies have found that when \mathcal{H} is supplied explicitly, humans can come close to the Bayesian

¹In fact, Kepler had tried fitting an oval to his observations only to reject it, and then labored for another seven years before finally trying an ellipse and realizing that it was mathematically equivalent to an oval. As he recounted, “The truth of nature, which I had rejected and chased away, returned by stealth through the back door, disguising itself to be accepted... Ah, what a foolish bird I have been!”

ideal e.g.,¹.² However, when humans must generate the set of hypotheses themselves, they cannot generate them all and instead generate only a subset, leading to judgment biases³. Some prominent biases of this kind are listed in Table ??.

Most previously proposed models of hypothesis generation rely on cued recall from memory based on similarity to previously observed scenarios c.f.². The probability of a generated hypothesis depends on the strength of its memory, and the number of such hypotheses generated is constrained by the available working memory resources. However, in most naturally encountered combinatorial hypothesis spaces, the number of possible hypotheses is vast and only ever sparsely observed.² showed that, when inferring Boolean concepts, people can generate previously unseen hypotheses by using compositional rules, instead of likening the situation to previously observed situations. So it seems that humans do not generate hypotheses only from the manageably small subset of previously observed hypotheses in memory and instead are able to generate hypotheses from the formidably large combinatorial space of all the conceivable possibilities. Given how large this space is, resource constraints at the time of inference suggest that only a subset are actually generated.

In this paper, we develop a normative theoretical framework for hypothesis generation in the domain of probabilistic inference, arguing that the brain copes with the intractability of inference by stochastically sampling hypotheses from the combinatorial space of possibilities. Although this sampling process is asymptotically exact, time pressure and cognitive resource constraints limit the

²This correspondence between human and Bayesian inference requires that the inference task must be one that is likely to have been optimized by evolution (e.g., predicting the duration of everyday events, categorizing and locating objects in images, making causal inferences); asking humans to reason consciously about unnatural problems like randomness or rare events tends to produce striking deviations from the Bayesian ideal see³ for discussion.

Table 1.1: Biases in human hypothesis generation and evaluation.

Name	Description	Reference
Subadditivity	Perceived probability of a hypothesis is higher when the hypothesis is described as a disjunction of typical component hypotheses (unpacked to typical examples).	?
Superadditivity	Perceived probability of a hypothesis is lower when the hypothesis is described as a disjunction of atypical component hypotheses (unpacked to atypical examples).	?, ?
Weak evidence effect	The probability of an outcome is judged to be lower when positive evidence for a weak cause is presented	?
Dud alternative effect	The judged probability of a focal outcome is higher when implausible alternatives are presented	?
Self-generation effect	The probability judgment over hypotheses that participants have generated themselves is lower as compared to the same hypotheses generated by others	??
Crowd within	The mean squared error of an estimate with respect to the true value reduces with the number of guesses. This reduction is more pronounced when the guesses are averaged across participants rather than within participants.	?
Anchoring and Adjustment	Generated hypotheses are biased by the hypothesis that is prompted at the start.	?

number of samples that can be generated, giving rise to systematic biases. We explore what sampler designs can reproduce the phenomena listed in Table ??, and then test our theory’s novel predictions in two experiments.

1.1 A RATIONAL PROCESS MODEL OF HYPOTHESIS GENERATION

Much of the recent work on probabilistic inference in human cognition has been deliberately agnostic about its underlying mechanisms, in order to make claims specifically about the subjective probability models people use in different domains¹. Because the posterior distribution $P(h|d)$ is completely determined by the joint distribution $P(h, d) = P(d|h)P(h)$, an idealized reasoner’s inferences can be perfectly predicted given this joint distribution. By comparing different assumptions about the joint distribution (e.g., the choice of prior or likelihood) under these idealized conditions, researchers have attempted to adjudicate between different models. Importantly, any algorithm that computes the exact posterior will yield identical predictions, which is what licenses agnosticism about mechanism. This method of abstraction is the essence of the “computational level of analysis”², and is closely related to the competence/performance distinction in linguistics and “as-if” explanations of choice behavior in economics.

The phenomena listed in Table ?? do not yield easily to a purely computational-level analysis, since different choices for the probabilistic model do not account for the systematic errors in approximating them. For this reason, we turn to “rational process” models see³ for a review, which make explicit claims about the mechanistic implementation of inference. Rational process models

are designed to be approximations of the idealized reasoner, but make distinctive predictions under resource constraints. In particular, we explore how sample-based approximations lead to particular cognitive biases in a large space of hypotheses, when the number of samples is limited. With an infinite number of samples, different sampling algorithms are indistinguishable as they all converge to the ideal response, but these algorithms display different behaviors at small sample sizes. We narrow the space of candidate sampling algorithms by studying these behaviors and comparing their predictions to observed cognitive biases.

1.1.1 MONTE CARLO METHODS

In their simplest form, sample-based approximations also known as *Monte Carlo* approximations;², take the following form:

$$P(b|d) \approx \hat{P}_N(b|d) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}[b_n = b], \quad (1.2)$$

where $\mathbb{I}[\cdot] = 1$ when its argument is true (0 otherwise) and b_n is a random hypothesis drawn from some distribution $Q_n(b)$.³ When $Q_n(b) = P(b|d)$, this approximation is unbiased, meaning $\mathbb{E}[\hat{P}_N(b|d)] = P(b|d)$, and asymptotically exact, meaning $\lim_{N \rightarrow \infty} \hat{P}_N(b|d) = P(b|d)$.

In general, a bounded reasoner cannot directly sample from the posterior, because the normalizing constant $P(d) = \sum_b P(b, d)$ requires the evaluation of the joint probabilities of each and every hypothesis and is intractable when the hypothesis space is large. In fact, sampling from the ex-

³This approach is straightforwardly generalized to sets of hypotheses: $\hat{P}_N(b \in H|d) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}[b_n \in H]$, where $H \subset \mathcal{H}$.

act posterior entails solving exactly the problem which we wish to approximate. Nonetheless, it is still possible to construct an asymptotically exact approximation by sampling from a Markov chain whose stationary distribution is the posterior; this method is known as *Markov chain Monte Carlo* (MCMC). Before presenting a concrete version of this method, we highlight several properties that make it suitable as a process model of hypothesis generation.

First, MCMC approximations are stochastic in the finite sample regime, producing “posterior probability matching”^{???}: hypotheses are generated with frequencies proportional to their posterior probabilities. Second, MCMC does not require knowledge of normalized probabilities at any stage and relies solely on an ability to compare the relative probabilities of two hypotheses. It has been shown in the literature[?] that humans have a better sense for relative rather than absolute probabilities. Third, MCMC allows for feedback between the generation and evaluation processes. The evaluated probability of already generated hypotheses influences if and how many new hypotheses will be generated, consistent with observations such as in ?. Fourth, Markov chains generate autocorrelated samples, consistent with autocorrelation in hypothesis generation^{???}. Correlation between consecutive hypotheses manifested as anchoring effects (where judgments are biased by the initial hypothesis[?]) are replicated by MCMC approximations that are also transiently biased (during the “burn-in” period) by their initial hypothesis, prior to reaching the stationary distribution[?]. Finally, work in theoretical neuroscience has shown how MCMC algorithms could be realized in generic cortical circuits^{???}.

We have emphasized properties that emerge in the finite sample regime because people tend to only generate a small number of hypotheses^{???}. Although this may seem to be manifestly sub-

optimal, it can be justified within a “resource-rational” framework^{???}. If generating hypotheses is costly (in terms of time and cognitive resources), then the rational strategy is to generate the minimum number of samples necessary to achieve a desired level of accuracy. This implies that incentives or uncertainty should have systematic effects on hypothesis generation. For example, ? showed that people generated more hypotheses when they were more uncertain. By the same token, cognitive load[?] or response time pressure[?] act as disincentives, reducing the number of generated hypotheses.

1.1.2 A SPECIFIC MARKOV CHAIN MONTE CARLO ALGORITHM

The space of MCMC algorithms is vast[?], but for the purposes of modeling psychological phenomena many of the algorithms generate indistinguishable predictions. Our goal in this section is to specify one such algorithm, without making a strong claim that people adhere to it in every detail. We focus on qualitative features of the algorithm that align with aspects of human cognition. Nonetheless, we shall see that the algorithm makes accurate quantitative predictions about human probabilistic judgments.

The most well-known and widely-used version of MCMC is the Metropolis-Hastings algorithm. Here, at step n in the Markov chain, new suggestions b' are drawn from a proposal distribution $Q(b'|b_n)$, where b_n is the hypothesis at step n . This proposal is accepted or rejected according to:

$$P(b_{n+1} = b'|b_n) = \min \left[1, \frac{P(d|b')P(b')Q(b_n|b')}{P(d|b_n)P(b_n)Q(b'|b_n)} \right]. \quad (1.3)$$

If the proposal is rejected, then the chain stays at the same hypothesis, $b_{n+1} = b_n$. Although the

posterior cannot be directly evaluated, we assume it can be computed up to a normalizing constant, since $P(b|d) \propto P(d|b)P(b)$. The acceptance function forces moves to higher probability hypotheses, while also stochastically exploring lower probability hypotheses. This process repeats until N samples have been generated. Once the burn-in period has elapsed, the amount of time the chain spends at a particular hypothesis is proportional to its posterior probability. The unique members of the set of accepted samples constitute the generated hypotheses, and the number of times they appear provides their judged probability.

We recap here two psychologically appealing properties of the algorithm mentioned in the previous section. First, we see that it relies solely on being able to gauge relative probabilities and not on having good estimates for any absolute probabilities. Second, the acceptance function engenders an interaction between generation and evaluation by ensuring that if one is at a high probability hypothesis, proposals are more likely to be rejected and therefore not generated. This follows the intuition that if one finds a good (high probability) hypothesis, one is less likely to generate more hypotheses. Conversely, if one is at a bad (low probability hypothesis), more proposals will be accepted.

The next step is to specify the proposal distribution. For simplicity, we assume that the proposal is symmetric, $Q(b'|b) = Q(b|b')$. This reduces the acceptance function to:

$$P(h_{n+1} = b'|h_n) = \min \left[1, \frac{P(d|b')P(b')}{P(d|h_n)P(h_n)} \right]. \quad (1.4)$$

We also assume that the proposal distribution is “local”: the proposal distribution preferentially

proposes hypotheses that are in some way “close” to the current one. This ensures that the next generated hypothesis is close to the current one with high probability. The alternate possibility is to instead have a “global” proposal distribution - for example one that proposes the next hypothesis uniformly at random from the space of all possible hypotheses, instead of favoring those closer to the current one.

MCMC algorithms always exhibit some autocorrelation irrespective of the proposal distribution, because the same state can occur consecutively several times if proposals are repeatedly rejected. However, we are also interested in the next *new* hypothesis that is generated, not exact repetitions of the same hypothesis. A more nuanced notion of autocorrelation takes into account the fact that sampled hypotheses can be “similar” (though not identical) when the proposal distribution is centered on a local neighborhood of the current hypothesis, as opposed to if the proposal is a “global” one. This kind of locality in determining the next state given the current one, has been studied previously in the context of traversing and searching semantic networks² and combinatorial spaces². This locality has been shown to be optimal as a foraging strategy² as well as consistent with human behavioral data. Since the generation of hypotheses is largely analogous to a search through the combinatorial space of conceivable possibilities, locality in the proposal distribution that moderates this search can be expected.

The question then is how we should define locality. This is relatively easy to answer in domains where the inference is over a one-dimensional continuous latent variable like in²; for example, one can use a normal distribution centered at the current hypothesis. For the discrete combinatorial hypothesis spaces studied in this paper, we assume that there is some natural clustering of the hypothe-

ses based on the observations they tend to generate (their centroids). We use the Euclidean distance between centroids as a measure of distance between clusters. In our simulations, we assume for simplicity that all hypotheses within a cluster are equidistant and that all clusters are equidistant from each other. The proposal distribution chooses hypotheses in the same cluster with a higher probability than those outside the cluster, but it treats all hypotheses within a cluster equiprobably. While this structure induces locality in the proposal distribution, we are not making a strong claim about the nature or role of clustering in hypothesis generation. We speculate about more sophisticated proposal distributions in the section on Future Work.

Finally, we need to specify how the chain is initialized. For cases where certain hypotheses are presented explicitly or primed in the query, we assume that the chain starts at one of those hypotheses. However, in cases where no hypotheses is explicitly prompted, we assume that the initial hypothesis is drawn from the prior over the hypotheses of interest. This assumption is consistent with evidence that hypotheses with high base rates are more likely to be generated[?]. There may also be initialization schemes that mix explicit prompts and sampling from the prior—for example a prompt that encourages sampling from a specific subset of the hypothesis space. We speculate about more sophisticated initialization schemes in the section on Future Work.

1.2 MODEL SIMULATIONS

In this section we apply our model to a range of empirical phenomena, using a disease-symptom Bayesian network as our running example. For each simulation, we run the Markov chain many

times and average the results, in order to emulate multiple participants in an experiment.

1.2.1 DIAGNOSTIC HYPOTHESES IN A DISEASE-SYMPTOM NETWORK

Our model is generally applicable to domains where the inference is carried out over a large space of possibilities that is sparsely observed and thus requires one to generate previously unobserved possibilities. A data set containing medical symptoms is a prototypical example of this problem: a patient could have any combination of more than one disease and many such combinations will not have been encountered before by an individual clinician. This combinatorial structure makes medical diagnosis computationally difficult—exact inference in a Bayesian network is known to be NP-hard[?]. To address this problem, approximate probabilistic inference algorithms (including Monte Carlo methods) are now widely-established e.g.,^{???}. It is thus reasonable to conjecture that diagnostic reasoning by humans could be captured by similar approximate inference algorithms. Suggestively, a number of the judgment biases listed in Table ?? have been documented in clinical settings^{???}; our goal is to investigate whether the MCMC model can reproduce these biases.

In the disease-symptom network, the observations are the presence or absence of symptoms and the latent variables are the presence or absence of diseases (S possible symptoms and D possible diseases). The diagnostic problem is to compute the posterior distribution over 2^D binary vectors, where each vector encodes the presence ($h_d = 1$) or absence ($h_d = 0$) of diseases $d = 1, \dots, D$. The

diseases are connected to the symptoms via a noisy-or likelihood, following² :

$$P(k_s = 1|h) = 1 - (1 - \varepsilon) \prod_{d=1}^D (1 - w_{ds})^{h_d}, \quad (1.5)$$

where $k_s = 1$ when symptom $s = 1, \dots, S$ is present (0 otherwise), $\varepsilon \in [0, 1]$ is a base probability of observing a symptom, and $w_{ds} \in [0, 1]$ is a parameter expressing the probability of observing symptom s when only disease d is present. Intuitively, the noisy-or likelihood captures the idea that each disease has an independent chance to produce a symptom.

As our goal is to use this set-up purely for illustrative purposes, we use a simplified fictitious disease-symptom data set designed to resemble real-world contingencies (Table ??). We designated two distinct clusters of four diseases each (gastrointestinal diseases and respiratory diseases); these two clusters have largely disjoint sets of symptoms, and the symptoms within a cluster are largely overlapping. We allow any combination of diseases to be present, making even this small number of diseases a fairly large space of 256 possible hypotheses.

1.2.2 SUBADDITIVITY

As described above, a resource-rational algorithm will arrest computation after a small number of samples, once accuracy is balanced against the cost of sampling². This gives rise to *subadditivity* (see Table ??): the probability of a disjunction (in “packed” form) is judged to be less than the probability of the same disjunction presented explicitly as the union of its sub-hypotheses (in “unpacked” form)^{2,3}, despite the fact that mathematically these are equal. For example, the probability of a gas-

Table 1.2: Parameters used for noisy-or model.

Diseases & Symp- toms	lung cancer	TB	respiratory flu	cold	gastro- enteritis	stomach cancer	stomach flu	food poisoning	base
Prior	0.001	0.05	0.1	0.2	0.1	0.05	0.15	0.2	1.0
cough	0.3	0.7	0.05	0.5	0.0	0.0	0.0	0.0	0.01
fever	0.0	0.1	0.5	0.3	0.0	0.0	0.1	0.2	0.01
chest pain	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.01
short breath	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.01
nausea	0.0	0.0	0.2	0.1	0.5	0.1	0.5	0.7	0.01
fatigue	0.0	0.0	0.2	0.3	0.1	0.05	0.2	0.4	0.01
stomach cramps	0.0	0.0	0.0	0.0	0.3	0.05	0.1	0.5	0.01
abdom. pain	0.0	0.0	0.01	0.0	0.1	0.5	0.0	0.0	0.01

gastrointestinal disease is judged to be less than the sum of the probabilities of each possible gastrointestinal disease.

Let us define a few terms here that we use in our simulations of these unpacking effects. The space of hypotheses that the disjunction refers to is called the *focal space* of the query. For example, when queried about the probability of a gastrointestinal diseases, the focal space is the set of all hypotheses that include at least one gastrointestinal disease. When unpacking this disjunction, we do not unpack to every single member of the focal space. Instead, we unpack to a few examples and to a *catch-all hypothesis* that refers to all other members of the focal space that were not explicitly unpacked. For example: "Food poisoning, stomach cancer or any other gastrointestinal disease" where a few example components of the focal space are explicitly unpacked (food poisoning and stomach

cancer) and presented along with a catch-all hypothesis (any other gastrointestinal disease).

Our model offers the following explanation of subadditivity: when a packed hypothesis is unpacked to typical examples and a catch-all hypothesis, the typical examples (that are part of the focal space) are explicitly prompted, causing the Markov chain to start there and thus include them in the cache of generated hypotheses. If the examples had not been explicitly prompted and instead a packed hypothesis had been presented, the chain initializes randomly and perhaps these high probability typical examples are never generated and so not included in the cache. Initializing the chain at a typical (high probability) state gives the chain a head-start in generating high probability hypotheses in the focal space and thus results in a larger probability judgment for that focal space.

To illustrate this effect in our medical diagnosis model, consider the following queries:

- Packed query: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of a gastrointestinal disease?
- Unpacked query (typical examples): Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of food poisoning, stomach flu, or any other gastrointestinal diseases?

The difference between the probability estimates between these two conditions is shown in Figure ??.

Experiments in⁷ show that the effect size of subadditivity decreases as the participants are given more time to answer the question. In our model, as more samples are taken, it becomes more and more likely that the packed chain also finds the high probability examples prompted in the unpacked scenario on its own. So the head-start given to the unpacked chain gets gradually washed out and the effect size of subadditivity decreases. If we assume that as more time passes, people

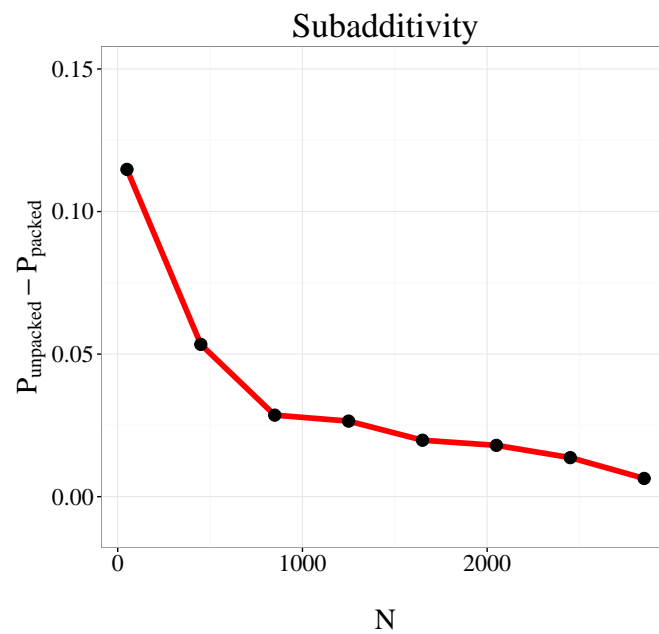


Figure 1.1: Subadditivity. MCMC estimates were made for the following queries: Given the symptoms *fever*, *nausea* and *fatigue*, (a) Packed: what is the probability that these symptoms were caused by the presence of a gastrointestinal disease? (b) Unpacked to typical examples: what is the probability that these were caused by the presence of food poisoning, stomach flu, or any other gastrointestinal diseases? The estimate for the unpacked condition is higher than that of the packed condition. The difference between these estimates is represented by the red line. This effect diminishes as the number of samples increases.

take more samples (up until a resource-rational limit on the number of samples), and that the time-points measured are before the resource-rational sample limit is met, our model replicates these time-dependence effects as seen in Figure ??.

1.2.3 SUPERADDITIVITY AND RELATED EFFECTS

Taking a limited number of samples with an MCMC sampler can also give rise to an effect opposite to the one described in the previous section, known as *superadditivity* (see Table ??): the probability of a disjunction (in “packed” form) is judged to be *greater* than the probability of the same disjunction presented explicitly as the union of its sub-hypotheses (in “unpacked” form)^{2,3}, despite the fact that mathematically they should be equal. This effect occurs when unpacking to atypical (low probability) examples and subadditivity prevails when unpacking to typical (high probability) examples.

The key feature that produces this effect is the acceptance function of the MCMC sampler and the feedback it causes between the generation and evaluation processes. If a chain is at a low probability hypothesis (such as when a low probability hypothesis is explicitly prompted in the form of an atypical unpacking), the chain is likely to accept more of the proposals made by the proposal distribution. Therefore this chain could generate many alternate hypotheses outside the focal space. In contrast, a chain at a higher probability hypothesis (for example, if it was randomly initialized in the focal space instead of being initialized at a particularly atypical example) will reject more of these proposals and remain at the initial hypothesis. So most of these proposals will not be generated. The probability estimate for the focal space \mathcal{A} is given by

$$\sum_{b \in \mathcal{A}} \hat{P}(b|d) = \sum_{b \in \mathcal{A}} \frac{1}{N} \sum_{n=1}^N \mathbb{I}[b_n = b] = \frac{\sum_{b \in \mathcal{A}} \sum_{n=1}^N \mathbb{I}[b_n = b]}{\sum_{b \in \mathcal{A}} \sum_{n=1}^N \mathbb{I}[b_n = b] + \sum_{b' \notin \mathcal{A}} \sum_{n=1}^N \mathbb{I}[b_n = b']} \quad (1.6)$$

Being in \mathcal{A} or not divides the total hypothesis space of \mathcal{H} into two mutually exclusive parts.

Therefore, the generation of more hypotheses outside the focal space (on average) when initialized at a consistently low probability (atypical) hypothesis in the focal space lowers the resulting probability estimate of the focal hypothesis space. This results in superadditive judgments.

To elucidate this effect in our medical diagnosis model, we use the following “unpacked to atypical examples” query: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease? The difference between the probability estimates from the two conditions is shown in Figure ??.

Previous accounts of subadditivity e.g.,^{2,3} cannot explain superadditivity; any unpacked example only increases the probability judgment of the unpacked query with respect to the packed query. This weakness of MINERVA-DM has been observed by³ in the context of its failure to model binary complementarity—an effect which their noise-based analysis can capture. However, their analysis still fails to completely capture superadditivity, as it constrains unpacked judgments to be greater than (and, only for binary complements, equal to) the packed judgment, never less than the packed judgment.³ explain superadditivity by suggesting that atypical examples divert attention from more typical examples and thus lower the probability estimate. But an explanation at the level of a ratio-

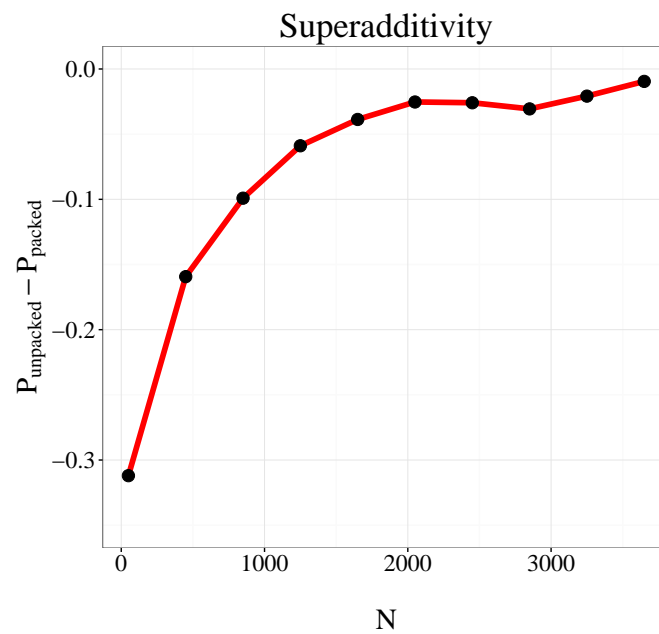


Figure 1.2: Superadditivity. MCMC-estimates were made for the following queries: Given the symptoms *fever*, *nausea* and *fatigue*, (a) Packed: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Unpacked to atypical examples: What is the probability is that these symptoms were caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease? The estimate for the unpacked condition is lower than that of the packed condition. The difference between these estimates is represented by the red line. This effect diminishes as the number of samples increases.

nal process model is, to our knowledge, lacking in the literature.

Some other cognitive effects can also be modeled by the same mechanism that gives rise to superadditivity. One example is the *weak evidence effect*: the perceived probability of an outcome is lowered by the presence of evidence supporting a weak cause³. While the added evidence increases the net posterior probability of the queried focal space, it also results in initialization at a very low probability hypothesis (an atypical or weak hypothesis). This initialization lowers the probability estimate as in the superadditivity effect and overwhelms the effects of the added evidence. This causes the ultimate perceived probability to be lower than if the positive evidence had not been presented and the chain was initialized randomly (on average at a higher probability hypothesis than the presented weak one) in the focal space.

To elucidate this effect in our medical diagnosis model, we use the following query:

- Control: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of gastrointestinal disease?
- Evidence for a weak cause: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of gastrointestinal disease, assuming the patient's grandmother was diagnosed with stomach cancer?

The increase in support of the weak cause (stomach cancer), by making available the presence of familial history, is implemented in our model by increasing the prior probability of stomach cancer in this patient from 0.05 to 0.06 (see Table ??). While this small change isn't expected to elicit a large difference in the probability of of gastrointestinal diseases between the two cases, it certainly does make it more (rather than less) probable compared to the control. However, it also causes the chain to be initialized at the weak hypothesis of stomach cancer by prompting it, resulting in the genera-

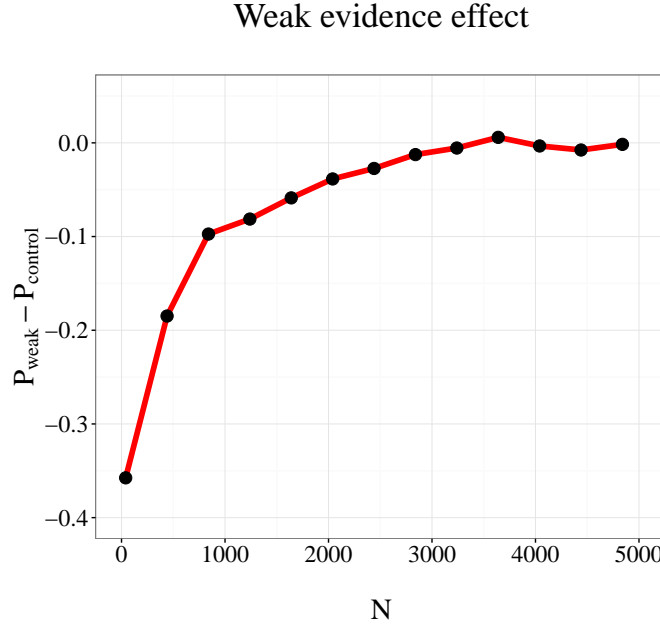


Figure 1.3: Weak evidence effect. MCMC estimates were made for the following queries: Given the symptoms *fever*, *nausea* and *fatigue*, (a) Control: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Evidence for a weak cause: What is the probability that these symptoms were caused by the presence of gastrointestinal disease, assuming the patient’s grandmother was diagnosed with stomach cancer? The increase in support of the weak cause (stomach cancer) is modeled by increasing the prior probability of stomach cancer from 0.05 to 0.06. The estimate from the weak evidence chain is lower than that from the control chain. The difference between these estimates is represented by the red line. The effect diminishes as the number of samples increases.

tion of more alternative hypotheses outside the focal space and a lower probability judgment than in the first case (Figure ??).

Another such bias is the *Dud alternative effect*: presenting low probability (or “dud”) alternate hypotheses increases the perceived probability of the focal space of hypotheses². This can be viewed as the superadditivity effect in the complement (alternate) hypothesis space. The queries being contrasted here are initialized in the space complementary to the focal space—i.e., the space of alternatives. Initialization at a low probability alternative when it is explicitly prompted in the question

results in a superadditive judgment (i.e., a lower probability judgment) of the complement space.

This lower probability estimate for the complement space entails a higher probability estimate for the focal space \mathcal{A} .

To elucidate this effect in our medical diagnosis model, we use the following queries:

- Control: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of a gastrointestinal disease)?
- Dud alternative: Given the symptoms *fever*, *nausea* and *fatigue*, what is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease)?

We see in Figure ?? that the model predicts that the scenario with dud alternatives produces higher probability judgments than the control. Findings in⁷ also suggest that the magnitude of this effect decreases with the amount of processing time given to participants. The model also replicates this phenomenon, if we assume that more time means more samples, and that the time points queried are before the resource-rational limit on the number samples is reached.

1.2.4 SELF-GENERATION OF HYPOTHESES

In this section, we focus on the *self-generation effect*: the probability judgment of a set of hypothesis that are generated and reported by a subject themselves is lower than when the same set of reported hypotheses is presented to a new subject^{7,8}. Our model provides the following explanation: Self-reported hypotheses generated by a chain are the modes it discovers after having explored the space and having generated several alternate hypotheses. However, in a situation where these high probability hypotheses are directly presented, the chain starts at the mode and is likely to get stuck—i.e.,

Dud alternative effect

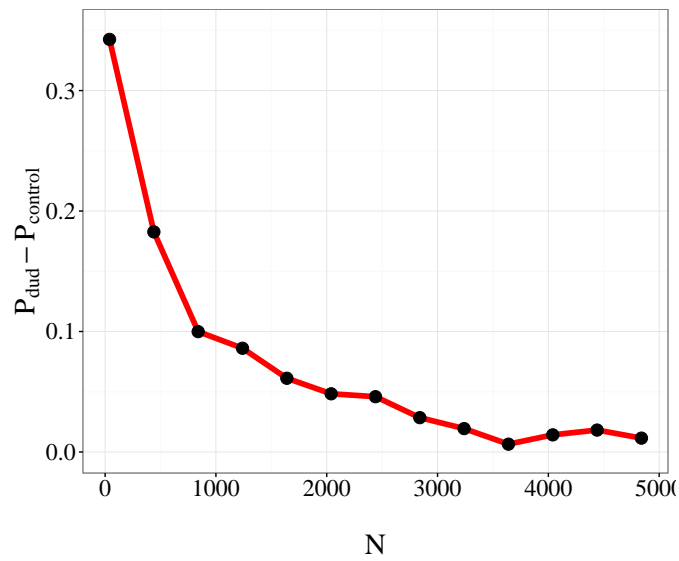


Figure 1.4: Dud alternative effect. MCMC estimates were made for the following queries: Given the symptoms *fever*, *nausea* and *fatigue*, (a) Control: What is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of a gastrointestinal disease)?, (b) With dud alternatives: What is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease)? The estimate from the control chain is higher than from the chain for which dud alternatives are presented. The difference between these estimates is represented by the red line and the effect diminishes as the number of samples increases

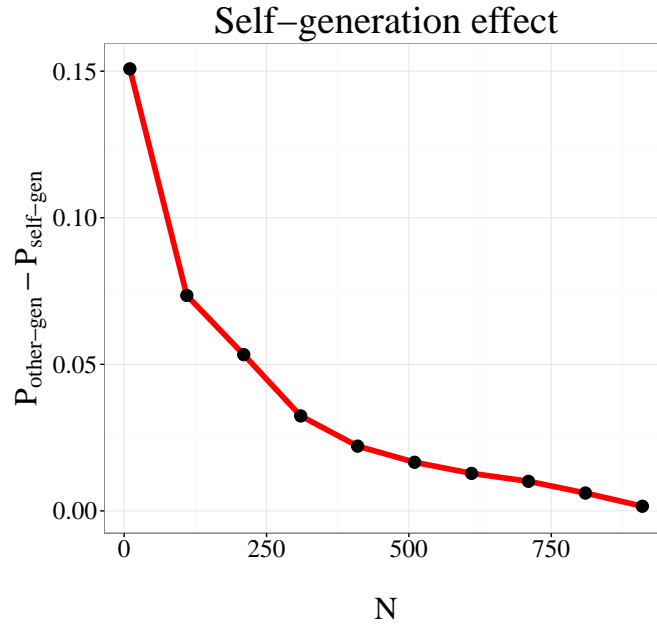


Figure 1.5: Self-generation effect. MCMC estimates for the following query: Given the symptoms *fever* and *fatigue*, (a) Self-generated: What are the two most likely respiratory diseases to have caused these symptoms? Estimate the probability that these symptoms are caused by either of these two diseases. (b) Other-generated: What is the probability that these symptoms were caused by the presence of a cold or respiratory flu (two most likely respiratory diseases to have caused these symptoms returned by the first chain)? The estimate from the other-generated chain is higher than from the self-generated chain. The difference between these estimates is represented by the red line and the effect decreases as the number of samples increases

not accept any of the proposals and thus not generate them at all. This, in the small sample limit, results in the generation of fewer alternate hypotheses. As in the previous section, fewer alternate hypotheses leads to a higher probability judgment.

We simulate an experiment analogous to the experiments in³ by querying the model as follows: Given the symptoms *fever* and *fatigue*, what are the two most likely respiratory disease to have caused these symptoms? To simulate the answer to this query, a randomly initialized “self-generated” chain is run and the 2 hypotheses over which this chain returns the highest probabilities are returned.

In this case, these are *a cold* and *respiratory flu*. The net probability estimate of the generated hypotheses *cold or respiratory flu* is tracked over time for the chain that generated them. A separate “other-generated” chain is queried as follows: Given the symptoms *fever* and *fatigue*, What is the probability that these symptoms were caused by the presence of a *cold* or *respiratory flu*? Thus, this chain is initialized at these high probability hypotheses of cold and respiratory flu. The difference between the probability estimates from these two chains is shown in Figure ??.

While this effect has previously been understood in terms of the generation of alternatives², a rational process model specifying a mechanism for this differential generation of alternatives is novel. Our explanation of this effect is largely contingent upon a link between generation and evaluation. In both self-generated and other-generated scenarios, the same hypothesis was generated, but evaluated differently depending on how many alternatives were generated. An MCMC chain can “get stuck” at a high probability hypothesis when initialized there by rejecting most of the new proposals, resulting in fewer generated alternatives.

1.2.5 ANCHORING AND ADJUSTMENT

In a classic experiment, ? had participants observe a roulette wheel that was predetermined to stop on either 10 or 65. Participants subsequently guessed the percentage of African countries in the United Nations. Participants who saw the wheel stopping on 10 guessed lower values than participants whose wheel stopped at 65. This and other findings led ? to hypothesize the “anchoring and adjustment” heuristic, according to which people anchor on a salient reference (even if it is irrelevant) and incrementally adjust away from the anchor towards the correct answer.

? showed that the anchoring and adjustment heuristic is a basic consequence of MCMC algorithms, due to the inherent autocorrelation of samples. Consistent with this account, our model posits that anchors, even when irrelevant, can serve to initialize the Markov chain. Locality guarantees that the chain will adjust incrementally away from the initial state, though anchoring will occur more generally as long as the rejection probability is non-zero. An MCMC algorithm with global proposals will capture anchoring to some extent because of its non-zero rejection probability and resulting auto-correlation of samples. However, without locality, estimates would not adjust incrementally away from the initial state. In other words, any MCMC algorithm will over-represent the initial anchoring hypothesis in the small sample limit, but only an MCMC algorithm with local proposals will also over-represent other hypotheses *close* to the initial anchoring hypothesis.

We illustrate this effect in Figure ?? using MCMC with local proposals on the disease-symptom network. The space of diseases in our example is clustered into respiratory and gastrointestinal diseases. The given symptoms are *fever* and *fatigue*. Chains initialized in different clusters show an initial within-cluster bias (i.e. not just a bias towards the initial anchoring hypothesis, but also to other hypotheses in its cluster), and this bias diminishes with the number of samples.

1.2.6 THE CROWD WITHIN

Error in estimates of numerical quantities decrease when the estimates are averaged across individuals, a phenomenon known as the *wisdom of crowds*³. This is expected if the error in the estimate of one individual is statistically independent from the error of the others, such that averaging removes the noise. Any unbiased stochastic sampling algorithm replicates this result, because taking more

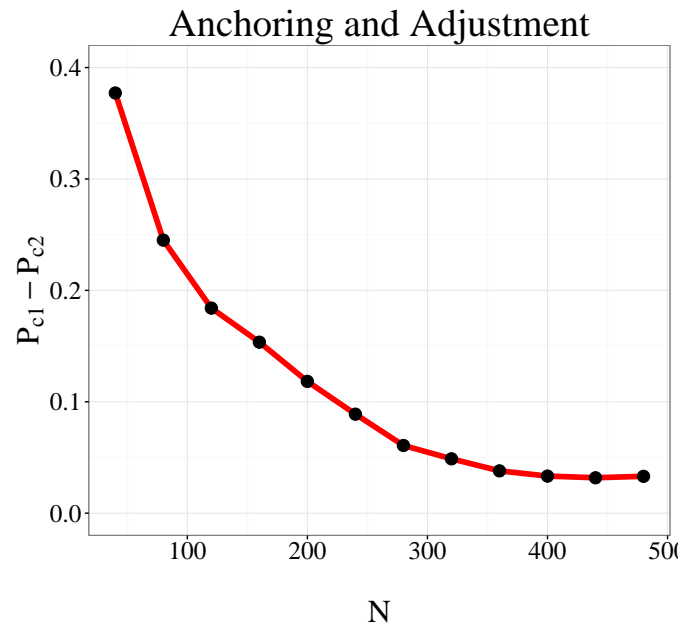


Figure 1.6: Anchoring and adjustment. The y axis represents the difference in the probabilities of respiratory flu and stomach flu given the symptoms *fever* and *fatigue* as returned by two different chains that are initialized differently. The chains are initialized in the two different clusters, at hypotheses other than the focal hypotheses of *respiratory* or *stomach flu*. Before reaching convergence, the chain initialized in cluster 1 of respiratory diseases places higher probability on respiratory flu, than the chain initialized in cluster 2 of gastrointestinal diseases. The net difference between the two chains diminishes as the number of samples increases.

samples gets one closer to the asymptotic regime, where the estimates are exact and the error tends to zero.

This error analysis was extended by³ to the effects of averaging across multiple estimates from a single individual. They found that averaging estimates reduced error—a phenomenon they named the *crowd within*. However, they also found that this error reduction was less compared to the reduction obtained by averaging across individuals. One explanation for this observation is that the error in the estimates given by the same individual are not entirely independent. We propose that the dependence between multiple estimates arises from an autocorrelated stochastic sampling algorithm like MCMC. This effect is illustrated in Figure ?? . We presented the following query to the model: Given symptoms are *fever*, *nausea* and *fatigue*, what is the probability that these symptoms are caused by the presence of a respiratory disease rather than a gastrointestinal disease? We ran several chains ($N_c = 24$) initialized randomly in the space of all possible diseases, with each run generating the same number of samples ($N_s = 200$). Each chain is initialized at the last sample of the previous chain⁴, for another N_s steps and a new set of N_c estimates are obtained, corresponding to the second guesses of the N_c individuals. This process is continued until we have 7 estimates from each of the $N_c = 24$ participants. The samples are then averaged either within or across individuals (chains). We find results analogous to those in³ —the error of the responses monotonically declines with the number of samples, and the error reduction is greater when averaging across (compared to

⁴We could also induce correlation between consecutive estimates by continuing the chain—i.e., carrying over the estimates from the first guess to the second one, instead of re-initializing. However, if we continue the chain, the second estimate is made with more samples and will always be lower error on average than the first one.³ find this to not be the case empirically.

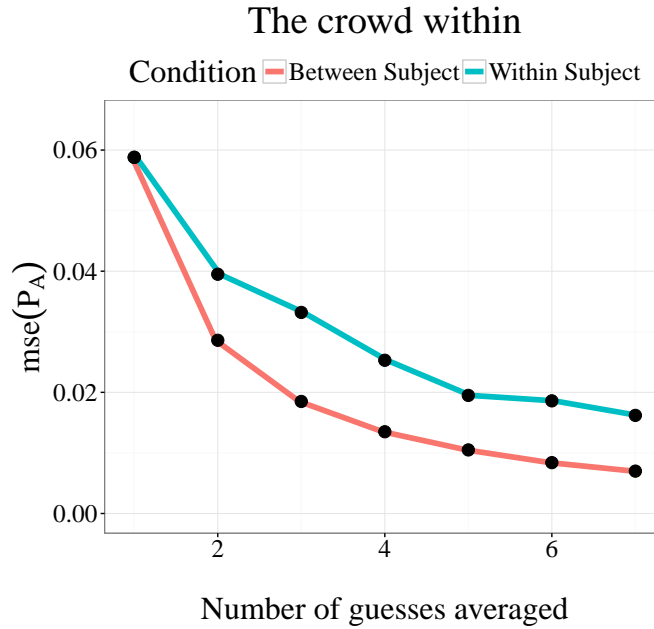


Figure 1.7: The crowd within. Errors in the MCMC estimates for the following query: Given the symptoms *nausea* and *shortness of breath*, what is the probability that these were caused by the presence of a respiratory disease? The estimates are averaged either over samples from the same individual (blue) or over samples from different individuals (red)

within) individuals.

Our MCMC model can replicate this effect because it generates auto-correlated samples. The last sample from one estimate is where the chain for the next estimate is initialized. As the sampling process is auto-correlated, subsequent samples in the second chain (in the small sample size limit) are correlated to its initial sample. Similarly, earlier samples from the first chain are correlated to its last sample. Because the samples from the two chains are correlated via the common sample, the probability estimates they generated are correlated as well. This auto-correlation exists irrespective of proposal distribution because of the non-zero rejection probability, but is strengthened by locality in

the proposals because this increases correlation.

1.2.7 SUMMARY OF SIMULATION RESULTS AND COMPARISON WITH IMPORTANCE SAMPLING

To highlight the distinctive predictions of MCMC, it is useful to compare it with other sampling algorithms that have been explored in the psychological literature. *Importance sampling* also uses a proposal distribution $Q(b)$, but unlike MCMC it samples multiple hypotheses independently and in parallel. These samples are then weighted to obtain an approximation of the posterior:

$$\hat{P}_N(b|d) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}[b_n = d] w_n, \quad (1.7)$$

where w_n is an “importance weight” for sample n computed according to:

$$w_n \propto \frac{P(b_n, d)}{Q(b_n)}. \quad (1.8)$$

Intuitively, the importance weight corrects for the fact that the importance sampler draws samples from the wrong distribution. ? have shown how this algorithm can be used to simulate human performance on a wide range of tasks. They also identified a correspondence between importance sampling and exemplar models, which have been widely applied in psychology. In related work, ? demonstrated how importance sampling could be realized in a biologically plausible neural circuit see also? .

Some of the effects we have replicated in this work could also be captured by an importance sam-

pling algorithm with limited samples. ? have proposed a model, HyGene, that is similar in spirit to an importance sampler with limited samples, with a memory driven proposal distribution that selects the hypotheses to be generated. HyGene explains subadditivity in terms of a failure to retrieve all the relevant hypotheses from memory due to stochastic noise in the retrieval process and limited working memory capacity.

The self-generation effect can to some extent be reproduced by importance sampling because prompting a hypothesis causes it to be sampled an extra time. So the probability of the focal space will be slightly larger if hypotheses in it are explicitly prompted (other-generated and presented to the participant) than if they are generated without prompting (self-generated). However, Experiment 2 in³ shows that in a situation where all the alternatives are specified, prompting specific hypotheses (as in the other-generated scenarios), does not result in a higher probability judgment than when these hypotheses are not prompted (as in the self-generated scenarios). The MCMC algorithm captures this finding because in a small hypothesis space, the Markov chain will visit all the hypotheses with the right frequency irrespective of initialization. By contrast, the importance sampler predicts a higher probability for other-generated hypotheses, contrary to the empirical finding.

This brings us to a key difference between importance sampling and MCMC: Importance sampling generates all hypotheses in parallel—the generation of new hypotheses has no dependence on hypotheses that have already been generated. Without this dependence, there is no interaction between the generation and evaluation processes. MCMC captures this dependence by sequentially generating hypotheses. Our model’s explanation of the self-generation effect, superadditivity, the weak evidence effect and the dud alternative effect rests on this dependence. The Markov chain can

get ‘stuck’ (at least temporarily) by rejecting proposals, thus generating fewer alternatives. If, on the other hand, the current hypothesis has low probability, more alternatives are generated and the probability estimate of the focal space is reduced.

The important sampler does not produce these effects, because its mechanism for generating new hypotheses is independent of the probability of the current one. If anything, prompting a hypothesis within the focal space, no matter how atypical, causes it to be sampled, *increasing* the importance sampler’s estimate for the probability of the focal space, contradicting superadditivity.

Another key difference between MCMC and importance sampling is that MCMC generates correlated samples, whereas consecutive samples from an importance sampler are totally independent. This prevents the importance sampler from reproducing the effects in Table ?? that rely on correlated sampling, such as the anchoring effect and the crowd within.

Importance sampling has also been adapted to inference in dynamical systems by generating hypotheses online—a technique known as *particle filtering*. This approach has been fruitfully applied to a number of domains in psychology, such as multiple object tracking⁷ and change detection⁸. Although hypotheses are generated sequentially by particle filtering, it is important to note that this structure is dictated by the sequential nature of the generative process. For example, in multiple object tracking, the object positions are dynamic latent variables; particle filtering generates new hypotheses about the positions after each new data point is observed. In contrast, the sequences of hypotheses generated by MCMC algorithms are purely mental, in the sense that hypotheses are generated sequentially regardless of whether the generative process is itself sequential. In this paper, we focus on non-sequential generative models in order to stress this point.

1.3 OVERVIEW OF EXPERIMENTS

We now turn to novel experimental tests of our theory. As discussed in the Introduction, the primary impetus for considering rational process models based on approximate inference is that inference in many real-world problems is computationally intractable. However, studying complex inference problems experimentally is challenging because it becomes harder to control participants' knowledge about the generative model. In the case of medical diagnosis, we can rely on the extensive training of clinicians, but it is unclear whether conclusions from these studies are generalizable to non-expert populations. Thus, for our experiments we sought a more naturalistic inference problem.

One domain in which humans have rich, quantifiable knowledge is visual scene understanding. Extensive research suggests that the visual system encodes information about natural scene statistics^{2,3}. This suggests that we can use generative models of natural scene statistics as proxies for human scene knowledge. We can then leverage such models to test theories of hypothesis generation in this domain.

Since low-level scene statistics like the distribution of oriented edges are not consciously accessible, we focus on object-level co-occurrence statistics, which have recently been quantified by ?. Specifically, ? measured the co-occurrence frequency of objects in a database of labeled natural images. To obtain a generative model from these co-occurrence statistics, we fit the latent Dirichlet allocation (LDA) model², which captures the distribution of co-occurrences in terms of latent topics. Each topic specifies a distribution over objects in a scene, and each scene is modeled as a probabilistic

mixture of topics. LDA captures the fact that microwaves are likely to co-occur with toasters, and cars are likely to co-occur with mailboxes.

For our purposes, the important point is that we can use this model to compute conditional probabilities over hidden objects in a scene, given a set of observed objects. Formally, let $b \in \mathcal{H}$ denote a hypothesis about k hidden objects in a scene, among all such possible hypotheses \mathcal{H} . Given a set of observed objects d , the inference problem is to compute the conditional probability $P(b \in H|d)$ that b is in some set $H \subset \mathcal{H}$ (e.g., hypotheses in which at least one of the hidden objects is an electrical appliance, or hypotheses in which the name of at least one of the hidden objects starts with a particular letter). This conditional probability can be approximated using MCMC in the hypothesis space.

In our experiments, we present participants with a set of observed objects, and ask them to estimate the probability that the hidden objects belong to some subset of possible objects. By manipulating the query, we attempt to alter the initialization of participants' mental sampling algorithm, allowing us to quantitatively test some of the predictions of our model.

Due to the relative complexity of this domain (compared to the simplified fictitious disease-symptom domain we have used so far for illustrative purposes), we refrain from making claims about the structure of proposal locality here and only test the predictions of our model that are immune to the choice of proposal distribution. Specifically, we focus on subadditivity and superadditivity.

1.4 EXPERIMENT 1

Our first prediction is the occurrence of both superadditivity and subadditivity in the same domain. The key factor is the typicality of the examples prompted by the unpacked query. We predict that if the query prompts typical examples from the focal space, probability judgments of that focal space will be higher than in the packed condition where no hypotheses are prompted (subadditivity). By contrast, if the question prompts atypical examples from the focal space, probability judgments of that focal space will be lower than in the packed condition where no hypotheses are prompted (superadditivity).

Using LDA as the probabilistic model, the data consist of visible objects in a scene, and the hypotheses are hidden objects. The focal space of hypotheses is given by a query such as *all objects starting with 'c'*. The focal space was unpacked into several either highly probable (typical) examples or highly improbable (atypical) examples, as well as a catch-all hypothesis. In the packed condition, the focal space is queried without any unpacked examples.

PARTICIPANTS

59 participants (26 females, mean age=35.76, SD=11.63) were recruited via Amazon's Mechanical Turk and received \$1 for their participation plus a performance-dependent bonus.

MATERIALS AND PROCEDURE

Participants were asked to imagine playing a game with a friend in which the friend specifies an object in a scene that they cannot see themselves. The task is to estimate the probability of certain sets of other objects in the same scene. For example, the friend could specify “pillow”. In the unpacked condition, participants were then asked to estimate the conditional probability of a focal space presented as a few examples and a catch-all hypothesis (e.g., “an armchair, an apple, an alarm clock or any other object starting with an A”). In the packed condition, the query did not contain any examples.

Table 1.3: Queries in Experiment 1. The letter determines the focal space (e.g., all objects beginning with A), conditioned on the cue object. Typical and atypical unpackings are shown for each focal space.

Cue object	Letter	Unpacked-typical	Unpacked-atypical
Pillow	A	armchair, alarm clock, apple	arch, airplane, artichokes
Rug	B	book, bouquet, bed	bird, buffalo, bicycle
Table	C	chair, computer, curtain	cannon, cow, canoe
Telephone	D	display case, dresser, desk	drinking fountain, dryer, dome
Computer	E	envelope, electrical outlet, end table	eggplant, electric mixer, elevator door
Armchair	F	fireplace, filing cabinet, fan	fire hydrant, fountain, fish tank
Stove	L	light, lemon, ladle	leavers, ladder, lichen
Chair	P	painting, plant, printer	porch, pie, platform
Bed	R	rug, remote control, radio	railroad, recycling bins, rolling pin
Kettle	S	stove, shelves, soap	suitcase, shoe, scanner
Sink	T	table, towel, toilet	trumpet, toll gate, trunk
Lamp	W	window, wardrobe, wine rack	wheelbarrow, water fountain, windmill

Each participant responded to one query for each of 9 different scenarios shown in Table ??, with 3 unpacked-atypical, 3 unpacked-typical, and 3 packed questions. We randomized the order of the scenarios as well as the assignment of scenarios to condition for each participant.

On every trial, participants first saw the cue object, followed by a hypothesis (either packed, unpacked-typical or unpacked-atypical). Participants had 20 seconds to estimate the probability of the hypothesis on a scale from 0 (not at all likely) to 100 (certain). For every timely response per trial they gained an additional reward of \$0.1. A screenshot of the experiment is shown in Figure ??.

The screenshot shows a web-based experimental interface. In the top right corner, there is a green square timer icon displaying '07' and the word 'Time' below it. The main text area contains the following elements: 'Number of trials left: 11', 'I see a **table**.', 'What is the probability that I also see', a list of options ('**chair**', '**computer**', '**curtain**') followed by 'or any other object with **C**.', a horizontal slider bar with a small grey handle at the left end and the number '0' below it, and a 'Submit' button at the bottom left.

Figure 1.8: Experimental setup. Participants were asked to estimate the conditional probability using a slider bar within a 20-second time limit.

MODEL FITTING

Our model has two free parameters: the number of hidden objects in the scene (k) and the number of samples (N). These parameters were fit to the behavioral data from both Experiment 1 and Experiment 2 combined, using a coarse grid search to optimize the mean-squared error between the

experimental probability estimates and the probability estimates from the model. The value of k that best fit the data was $k = 6$, and the number of samples $N = 230$. This value of k is in the same ballpark as values found for average number of uniquely labeled objects in natural scenes from data collected in². This value for N as the number of samples is higher than numbers found in some previous work like² etc, but it is important to note that each unique hypothesis can appear several times in the sample set. So even if the number of samples is larger than in previous studies, the number of unique hypotheses is comparable.

RESULTS AND DISCUSSION

We compared the mean probability judgments for each condition (Figure ??). Consistent with our hypothesis, we found subadditivity in the unpacked-typical condition, with significantly higher probability estimates compared to the control condition [$t(116) = 4.08, p < 0.001$], and superadditivity in the unpacked-atypical condition, with significantly lower probability estimates compared to the control condition [$t(116) = -3.42, p < 0.001$]. This pattern of results was captured by our MCMC model.

Our results confirm the prediction that subadditivity and superadditivity will occur within the same paradigm, depending on the typicality of unpacking. A related result was reported by², who found subadditivity only when the definition of the focal space was fuzzy and typical unpacking may have led to the consideration of a larger focal space. We consider this study in more detail in the General Discussion.

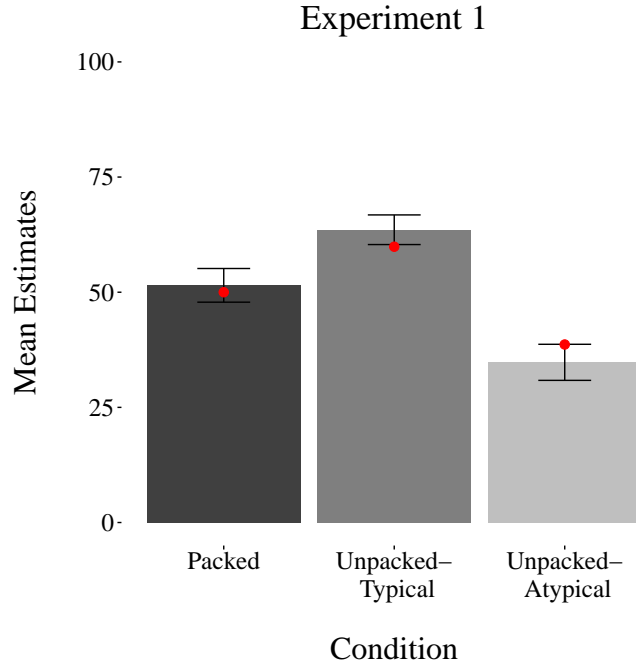


Figure 1.9: Experiment 1 results. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene.

1.5 EXPERIMENT 2

In Experiment 1, we demonstrated that the typicality of unpacked examples has a powerful effect on biases in probability estimation. In Experiment 2, we provide converging evidence by showing that different biases can be induced for the same unpacked examples by changing the cue object.

Typicality depends on an interaction between the cue and the examples: in the presence of a road, a crosswalk is typical and a coffee-maker is atypical, but the opposite is true in the presence of a sink. Our model predicts that subadditivity will occur when unpacked examples are typical for a given cue object, whereas superadditivity will occur when the same examples are atypical for a different cue

object.

PARTICIPANTS

180 participants (84 females, mean age = 34.25, SD = 11.16) were recruited via Amazon's Mechanical Turk web service and received \$0.5 for their participation plus a performance-dependent bonus.

MATERIALS AND PROCEDURE

The experimental procedure was identical to Experiment 1, except for the choice of scenarios (Table ??). Each participant responded to one unpacked-typical, one unpacked-atypical and one packed scenario in random order.

Table 1.4: Queries in Experiment 2. The letter determines the focal space (e.g., all objects beginning with A), conditioned on the cue object. Conditioned on cue object 1, unpacking 1 is predicted to cause subadditivity and unpacking 2 is predicted to cause superadditivity. These predictions reverse for cue object 2.

Cue object 1	Cue object 2	Letter	Unpacking 1	Unpacking 2
Pillow	Faucet	B	bed skirt, bedspread	bucket, bread
Road	Sink	C	cabin, crosswalk	cup, coffee maker
Cabinet	Road	T	toothpaste, tray	terrace, tunnel

RESULTS AND DISCUSSION

As shown in Figure ??, we observed a superadditivity effect: probability estimates were significantly higher in the packed condition compared to the atypical unpacking for both cue object 1 [$t(165) = 3.31, p < 0.01$] and cue object 2 [$t(162) = 4.31, p < 0.01$]. We did not observe a subadditivity effect for either cue object 1 [$t(171) = 0.73, p > 0.05$] or cue object 2 [$t(168) = 0.08, p > 0.05$].

Importantly, we found a significant interaction between the cue-object and the unpacking of the objects [$F(498, 2) = 12.69, p < 0.001$]. In particular, when conditioning on cue object 2, using “Unpacking 1” (see Table ??) leads to significantly lower estimates than using “Unpacking 2” [$t(251) = 2.52, p < 0.01$]. Additionally, when conditioning on cue object 1, using “Unpacking 2” produces significantly lower estimates than using “Unpacking 1”; [$t(165) = -3.31, p < 0.001$]. These results show that typicality of the unpackings and, by proxy the sub- and super-additive effects, crucially depend on the conditioned cue object.

Our fitted model matches the experimental data well ($r = 0.96, p < 0.001$), only slightly underestimating the superadditive effect with cue object 2 and unpacking 1. We can conclude from the fact that this cue-dependent swap can be even partially carried out—for example, the superadditivity effect certainly does get swapped—indicates that these effects are not modulated solely by the prior typicality or inherent availability of the unpacked examples. The same unpacking that induces superadditivity in the presence of one cue object, does not induce it in the presence of the second cue object. Furthermore, a new unpacking can be chosen such that it induces superadditivity in the presence of the second cue object but not in the presence of the first. These results support a sampling process that is modulated by the cue objects, i.e. the observed data.

1.6 GENERAL DISCUSSION

We have presented a rational process model of inference in complex hypothesis spaces. The main idea is to recast hypothesis generation as a Markov chain stochastically traversing the hypothesis

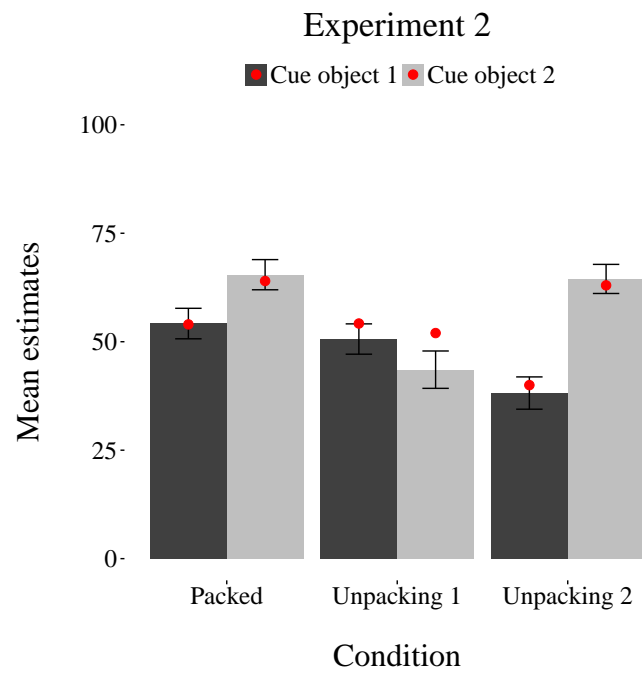


Figure 1.10: Experiment 2 results. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene. Unpacking 1 is typical for cue object 1 and atypical for cue object 2; unpacking 2 is typical for cue object 2 and atypical for cue object 1.

space, such that hypotheses are visited with a long-run frequency proportional to their probability. Our simulations demonstrated that this model reproduces many observed biases in human hypothesis generation. Finally, we confirmed in two experiments the model’s prediction that subadditivity and superadditivity depend critically on the typicality of unpacked examples.

Our work extends a line of research on using rational process models to understand cognitive biases. Most prominently, ? have attempted in their HyGene model to explain a wide range of hypothesis generation phenomena by assuming that Bayesian inference operates over a small subset of hypotheses drawn from memory. We follow a similar line of reasoning, but depart in the assumption that hypotheses may be generated *de novo* through stochastic exploration of the hypothesis space. This assumption is important for understanding how humans can generate hypotheses in complex combinatorial spaces where it is impossible to store all relevant hypotheses in memory.

Prior studies suggest that—when averaged over long time periods or across individuals—probability estimates converge roughly to the Bayesian ideal[?]. Like other models based on Monte Carlo methods e.g.,^{???}, our model predicts exact Bayesian inference in the limit of large sample sizes. However, cognitively bounded agents are expected to be *computationally rational*[?]: sampling takes time and energy, and hence the optimal sampling strategy will tend to generate relatively few hypotheses[?].

Our model recreates several cognitive biases exhibited by humans: subadditivity, superadditivity, anchoring and adjustment, weaker confidence in self-generated hypotheses, the crowd within, and the dud alternative and weak evidence effects. While some of these biases have been accounted for by other models, ours is the first unified rational process account. Table ?? provides a systematic comparison of which phenomena are accounted for by different models.

Our simulation results rest on two key features of the model. First, our model posits an interplay between generation and evaluation of hypotheses: when a low probability hypothesis has been generated, the sampler is more likely to accept new proposals compared to when a high probability hypothesis has been generated. This property of MCMC allows us to understand superadditivity and related effects (such as the dud alternative and weak evidence effects), where unpacking a query into low probability examples causes a reduction in the probability estimate for the focal space. This feature also explains why participants give lower probability estimates to hypotheses that are self-generated compared to those generated by others and presented to them. A shortcoming of previous models based on importance sampling² or cued recall² is that the generation and the evaluation processes are largely decoupled; the probabilities of the hypotheses already in the cache of generated hypotheses do not affect whether or not new hypotheses are generated.

The second key property of our model is the autocorrelation of hypotheses in the Markov chain. This autocorrelation arises from two sources: the non-zero rejection rate (which ensures that the chain sometimes stays at its current hypothesis for multiple time steps) and the locality of the proposal distribution (which ensures that proposed hypotheses are in the vicinity of the previously generated hypothesis). Previous models based on importance sampling or cued recall generate new candidate hypotheses independently of the hypotheses that have already been generated (i.e., the previously generated hypotheses have no impact on future hypotheses).² argued that autocorrelation and locality of proposals in MCMC models can account for the anchoring and adjustment phenomena. They analyzed a one-dimensional continuous hypothesis space for numerical estimation, and we generalized this idea to combinatorial spaces. More broadly, several findings in the literature

suggest hypothesis autocorrelation^{2,2,2}. For example, the “crowd within” phenomenon², which we also simulate, demonstrates that errors in numerical guesses are correlated in time, and this error is reduced if the guesses are spread out.

MCMC models with global proposal distributions will show much weaker autocorrelation compared to those with local proposal distributions, because any autocorrelation will depend entirely on rejection of proposals. Since efficient samplers have relatively low rejection rates², there is reason to believe that human probability estimation makes use of local proposal distributions. Evidence for locality has been found in domains analogous to that of hypothesis generation^{2,2}, further suggesting that humans use local proposal distributions.

Table 1.5: Comparison of stochastic sampling algorithms

Effect	Stochastic Sampling Variants		
	Importance Sampling	Global proposal MCMC	Local proposal MCMC
Subadditivity	✓	✓	✓
Superadditivity	×	✓	✓
Weak Evidence effect	×	✓	✓
Dud Alternative effect	×	✓	✓
Self-generation effect	× ⁵	✓	✓
Crowd within	×	✓	✓
Anchoring & adjustment	×	×	✓

Previous work demonstrating the effect of superadditivity² did not find subadditivity except in situations where the search was over an ill-defined fuzzy category, such that unpacked typical exam-

⁵While an importance sampler does reproduce this effect, we have elaborated in the section comparing our MCMC model to importance sampling how its explanation does not extend to follow-up studies on this effect in²

ples lead participants to consider a larger hypothesis space than entailed by the packed query. However, this effect was driven by a single item: *Guns that you buy at a hardware store with staple gun* as the unpacked typical example. Excluding this item, typical unpackings were not subadditive. Our experiments demonstrated that subadditivity can be obtained in well-defined (non-fuzzy) domains like “words starting with the letter A”, and where typical unpackings do not extend the hypothesis space. A possible explanation for this is that, as opposed to the studies in³, we time constrain the responses from our participants. This time pressure could lead to fewer samples being taken and an amplification of the subadditivity effect.

1.7 FUTURE WORK

Our model can be improved in several ways. First, we adopted relatively simple assumptions about initialization of the Markov chain. Recent work suggests that humans might use a fast, data-driven proposal distribution learned from previous experience^{2,3}. Second, our simplistic assumptions about the proposal distribution could likewise be made more sophisticated by using data-driven methods. Finally, we have assumed that the number of samples is constrained solely by the available time, but the computational rationality perspective argues that this number is chosen adaptively to balance the benefits of taking more samples against their costs in time and energy^{3,4,5}. Investigating cognitive algorithms for meta-control of sampling is an interesting avenue for future research.

Our experiments and simulations only studied two domains (medical diagnosis and scene understanding), but there exist many real-world domains that impose a severe computational burden on

mental inference. For example, it has been shown that humans are capable of simulating physical trajectories that they have never directly observed before and make fairly accurate inferences when predicting the motion of a projectile², judging mass in collisions², and judging the balance of block towers². Furthermore, research also suggests that humans sample noisy simulations of future trajectories^{2,2}, but the precise sampling mechanisms are currently unknown. The number of possible trajectories is exponentially large in this domain, and thus approximate inference schemes like MCMC may come into play.

ACKNOWLEDGMENTS

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We thank Kevin Smith for helpful comments and discussions.

This is some random quote to start off the chapter.

Firstname lastname

2

Re-use of previous solutionf for faster inference

Many theories of probabilistic reasoning assume that human brains are equipped with a general-purpose inference engine that can be used to answer arbitrary queries for a wide variety of probabilistic models². For example, given a joint distribution over objects in a scene, the inference engine can

be queried with arbitrary conditional distributions, such as:

- What is the probability of a microwave given that I've observed a sink?
- What is the probability of a toaster given that I've observed a sink and a microwave?
- What is the probability of a toaster and a microwave given that I've observed a sink?

The nature of the inference engine that answers such queries is still an open research question, though many theories posit some form of approximate inference using Monte Carlo sampling e.g.,^{??????}. According to these theories, probability distributions are mentally represented with a set of samples, which are generated using a general-purpose inference engine that can operate on arbitrary probability distributions.

The flexibility and power of such a general-purpose inference engine trades off against its computational efficiency: by treating each query distribution independently, an inference engine forgoes the opportunity to reuse computations across queries, thus reducing time complexity (but possibly increasing space complexity). Every time a distribution is queried, past computations are ignored and answers are produced anew—the inference engine is memoryless, a property that makes it statistically accurate but inefficient in environments with overlapping queries.

Continuing the scene inference example, answering the third query should be easily computable once the first two queries have been computed. Mathematically, the answer can be expressed as:

$$P(\text{toaster} \wedge \text{microwave} | \text{sink}) = P(\text{toaster} | \text{sink}, \text{microwave})P(\text{microwave} | \text{sink}). \quad (2.1)$$

Even though this is a trivial example, standard inference engines do not exploit these kinds of regu-

larities because they are memoryless—they have no access to traces of past computations.

An inference engine may gain efficiency by incurring some amount of bias due to reuse of past computations—a strategy we will refer to as *amortized inference*². For example, if the inference engine stores its answers to the “toaster” and “microwave” queries, then it can efficiently compute the answer to the “toaster or microwave” query without rerunning inference from scratch. More generally, the posterior can be approximated as a parametrized function, or *recognition model*, that maps data in a bottom-up fashion to a distribution over hypotheses, with the parameters trained to minimize the divergence between the approximate and true posterior.¹ By sharing the same recognition model across multiple queries, the recognition model can support rapid inference, but is susceptible to “interference” across different queries, a property that we explore below.

One way to construct a recognition model is using Monte Carlo sampling: the recognition model can be viewed as a kind of data-driven sampler whose parameters are optimized so that the samples resemble the true posterior. In an amortized architecture, these parameters are shared across different inputs (i.e., data) and hence the samples will be correlated, introducing a systematic bias. If the sampling process corresponds to a Markov chain Monte Carlo algorithm (see below), this bias will disappear with a sufficiently large number of samples, but since humans appear to take a relatively small number of samples³, we expect this bias to be measurable.

Amortization has a long history in machine learning; the *locus classicus* is the Helmholtz ma-

¹Formally, this is known as *variational inference*³, where the divergence is typically the Kullback-Leibler divergence between the approximate and true posterior. Although this divergence cannot be minimized directly (since it requires knowledge of the true posterior), a bound (variational free energy) can be tractably optimized for some classes of approximations.

chine^{??}, which uses samples from the generative model to train a recognition model. More recent extensions and applications of this approach e.g.,^{????} have ushered in a new era of scalable Bayesian computation in machine learning. We propose that amortization is also employed by the brain see[?] for a related proposal, flexibly reusing past inferences in order to efficiently answer new but related queries. The key behavioral prediction of amortized inference is that people will show correlations in their judgments across related queries.

We report 3 experiments that test this prediction using a variant of the probabilistic reasoning task previously studied by ?. In this task, participants answer queries about objects in scenes, much like in the examples given above. Crucially, the hypothesis space is combinatorial because participants have to answer questions about sets of objects (e.g., “All objects starting with the letter S”). This renders exact inference intractable: the hypothesis space cannot be efficiently enumerated. In our previous work[?], we argued that people approximate inference in this domain using a form of Monte Carlo sampling. Although this algorithm is asymptotically exact, only a small number of samples can be generated due to cognitive limitations, thereby revealing systematic cognitive biases such as anchoring and adjustment, subadditivity, and superadditivity see also^{???}.

We show that the same algorithm can be generalized to reuse inferential computations in a manner consistent with human behavior. First we describe how amortization might be used by the mind. We consider two crucial questions about how this might be implemented: what parts of previous calculations do people reuse —all previous memories or summaries of the calculations— and when do they choose to reuse their amortized calculations. Next we test these questions empirically. In Experiment 1, we demonstrate that people *do* use amortization by showing that there is a lingering

influence of one query on participants’ answers to a second, related query. In Experiment 2, we explore what is reused, and find that people use summary statistics of their previously generated hypotheses, rather than the hypotheses themselves. Finally, in Experiment 3, we show that people are more likely to reuse previous computations when those computations are most likely to be relevant: when a second cue is similar to a previously evaluated one.

2.1 HYPOTHESIS GENERATION AND AMORTIZATION

Before describing the experiments, we provide an overview of our theoretical framework. First, we describe how Monte Carlo sampling can be used to approximate Bayesian inference, and summarize the psychological evidence for such an approximation. We then introduce amortized inference as a generalization of this framework.

2.1.1 MONTE CARLO SAMPLING

Bayes’ rule stipulates that the posterior distribution is obtained as a normalized product of the likelihood $P(d|h)$ and the prior $P(h)$:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \quad (2.2)$$

where \mathcal{H} is the hypothesis space. Unfortunately, Bayes’ rule is computationally intractable for all but the smallest hypothesis spaces, because the denominator requires summing over all possible hypotheses. This intractability is especially prevalent in combinatorial space, where hypothesis spaces

are exponentially large. In the scene inference example, $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_K$ is the product space of latent objects, so if there are K latent objects and \mathcal{M} possible objects, $|\mathcal{H}| = \mathcal{M}^K$. If we imagine there are $\mathcal{M} = 1000$ kinds of objects, then it only takes $K = 26$ latent objects for the number of hypotheses to exceed the number of atoms in the universe.

Monte Carlo methods approximate probability distributions with samples $\mathcal{S} = \{b_1, \dots, b_N\}$ from the posterior distribution over the hypothesis space. We can understand Monte Carlo methods as producing a recognition model $Q(b|d)$ parametrized by \mathcal{S} see³ for a systematic treatment. In the idealized case, each hypothesis is sampled from $P(b|d)$. The approximation is then given by:

$$P(b|d) \approx Q(b|d) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}[b_n = b], \quad (2.3)$$

where $\mathbb{I}[\cdot] = 1$ if its argument is true (and 0 otherwise). The accuracy of this approximation improves with N , but from a decision-theoretic perspective even small N may be serviceable^{3,22}.

The key challenge in applying Monte Carlo methods is that generally we do not have access to samples from the posterior. Most practical methods are based on sampling from a more convenient distribution, weighting or selecting the samples in a way that preserves the asymptotic correctness of the approximation². We focus on Markov chain Monte Carlo (MCMC) methods, the most widely used class of approximations, which are based on simulating a Markov chain whose stationary distribution is the posterior. In other words, if one samples from the Markov chain for long enough, eventually b will be sampled with frequency proportional to its posterior probability.

A number of findings suggest that MCMC is a psychologically plausible inference algorithm.

First, MCMC does not require knowledge of normalized probabilities at any stage and relies solely on an ability to compare the relative probabilities of two hypotheses. This is consistent with evidence that humans represent probabilities on a relative scale². Second, MCMC allows for feedback between the generation and evaluation processes. The evaluated probability of already-generated hypotheses influences if and how many new hypotheses will be generated, consistent with experimental observations². Finally, Markov chains also generate autocorrelated samples. This is consistent with autocorrelation in hypothesis generation^{2,2,2}.

Many implementations use a form of local stochastic search, proposing and then accepting or rejecting hypotheses. For example, the classic Metropolis-Hastings algorithm first samples a new hypothesis b' from a proposal distribution $\phi(b'|b_n)$ and then accepts this proposal with probability

$$P(h_{n+1} = b'|b_n) = \min \left[1, \frac{P(d|b')P(b')\phi(b_n|b')}{P(d|b_n)P(b_n)\phi(b'|b_n)} \right]. \quad (2.4)$$

Intuitively, this Markov chain will tend to move from lower to higher probability hypotheses, but will also sometimes “explore” low probability hypotheses. In order to ensure that a relatively high proportion of proposals are accepted, $\phi(b'|b_n)$ is usually constructed to sample proposals from a local region around b_n . This combination of locality and stochasticity leads to a characteristic pattern of small inferential steps punctuated by occasional leaps, much like the processes of conceptual discovery in childhood² and creative insight in adulthood². Even low-level visual phenomena like perceptual multistability can be described in these terms^{2,2}.

Another implication of MCMC, under the assumption that a small number of hypotheses are

Table 2.1: Unpacking induced biases in human hypothesis generation and evaluation.

Name	Description	References
Subadditivity	Perceived probability of a hypothesis is higher when the hypothesis is described as a disjunction of typical component hypotheses (unpacked to typical examples). $P(A_{\text{typical}} \cup B) < P(A_{\text{typical}}) + P(B)$?, ?
Superadditivity	Perceived probability of a hypothesis is lower when the hypothesis is described as a disjunction of atypical component hypotheses (unpacked to atypical examples). $P(A_{\text{atypical}} \cup B) > P(A_{\text{atypical}}) + P(B)$??

sampled, is that inferences will tend to show anchoring effects (i.e., a systematic bias towards the initial hypotheses in the Markov chain). Lieder and colleagues have shown how this idea can account for a wide variety of anchoring effects observed in human cognition^{??}. For example, priming someone with an arbitrary number (e.g., the last 4 digits of their social security number) will bias a subsequent judgment (e.g., about the birth date of Gandhi), because the arbitrary number influences the initialization of the Markov chain.

In previous research[?], we have shown that MCMC can account for many other probabilistic reasoning “fallacies,” suggesting that they arise not from a fundamental misunderstanding of probability, but rather from the inevitable need to approximate inference with limited cognitive resources. We explored this idea using the scene inference task introduced in the previous section. The task facing subjects in our experiments was to judge the probability of a particular set of latent objects

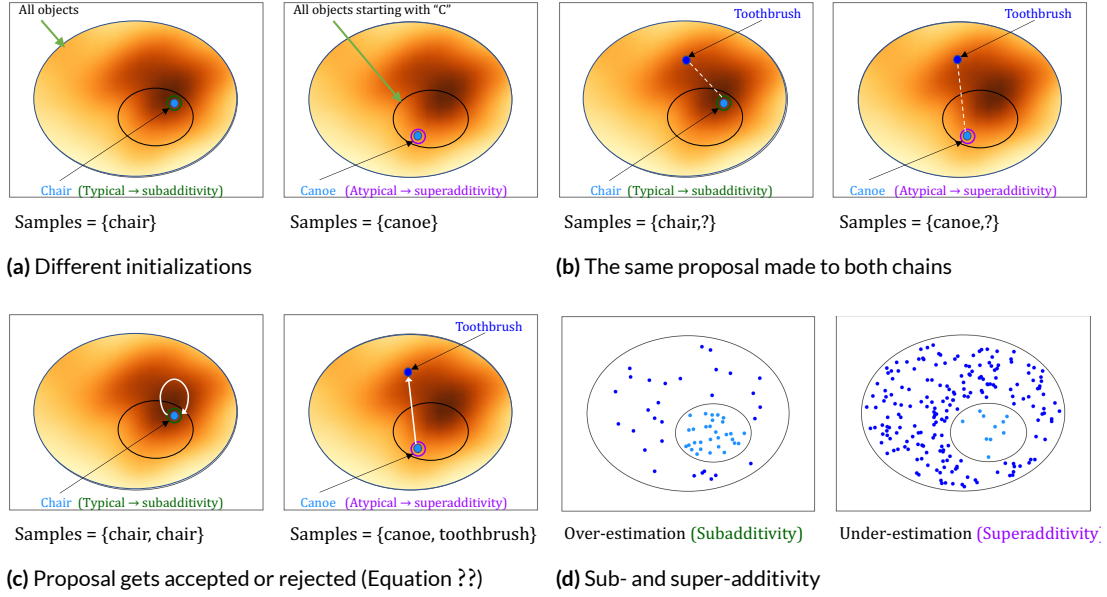


Figure 2.1: Demonstration of how MCMC sampling can give rise to sub- and super-additivity for different unpacked versions of the question : “In the presence of a table, what is the probability that there is also another object starting with C?”. The color gradient indicates probability density. (a) The chain initialized with a typical unpacking starts at ‘chair’, a high probability hypothesis, denoted by a darker shading, while the chain initialized with an atypical unpacking starts at ‘canoe’, a low probability hypothesis, denoted by a lighter shading. (b) For the purposes of illustration we show the same new intermediate probability proposal of ‘toothbrush’ being made to both chains. In the model, this proposal is randomly generated for each chain. (c) Since the probability of ‘toothbrush’ is significantly higher than ‘canoe’ the proposal is accepted by the atypically unpacked chain. But conversely since it is significantly less probable than ‘chair’, is likely rejected by the typically unpacked chain. (d) The tendency for the typically unpacked chain to tarry in the high probability region of the queried object set, gives rise to sub-additivity, whereas the tendency for the atypically unpacked to get easily derailed into regions outside the queried object set gives rise to super-additivity.

(the hypothesis, h) in a scene conditional on observing one object (the cue, d). By manipulating the framing of the query, we showed that subjects gave different answers to formally equivalent queries (see Table ??). In particular, by partially unpacking the queried object set (where fully unpacking an object set means to present it explicitly as a union of each of its member objects) into a small set of exemplars and a “catch-all” hypothesis (e.g., “what is the probability that there is a chair, a computer, or any other object beginning with C?”), we found that subjects judged the probability to be higher when the unpacked exemplars were typical a “subadditivity” effect; cf.² and lower when the unpacked exemplars were atypical a “superadditivity” effect; cf.³ compared to when the query was presented without any unpacking.

To provide a concrete example, in the presence of the cue “table,” the typically unpacked query “what is the probability that there is also a chair, a computer, or any other object beginning with C?” generates higher probability estimates relative to the packed query “what is the probability that there is another object beginning with C?”, whereas the atypically unpacked query “what is the probability that there is also a cow, a canoe, or any other object beginning with C?” generates lower probability estimates compared to the packed query.

The generative model for this scene inference task is approximated by fitting the database of natural scenes with hand-labeled objects, provided in ?, to a latent Dirichlet allocation (LDA) model². Specifically, the database consists of object co-occurrence statistics in natural scenes, which we model with a set of underlying “topics” (probability distributions over objects). This model allows us to analytically compute the joint probability of any combination of different objects. Finding the exact normalized conditional probabilities is still intractable due to the combinatorially large number of

possible hypotheses to normalize over, but Monte Carlo sampling methods like MCMC can approximate these probabilities.

We were also able to account for the sub- and super-additivity effects using MCMC under the assumption that the unpacked exemplars initialize the Markov chain that generates the sample set of query objects conditioned on the given cue object⁷. Because the initialization of the Markov chain transiently determines its future trajectory, initializing with typical examples causes the chain to tarry in the high probability region of the queried object set, thereby increasing its judged probability (subadditivity). In contrast, initializing with atypical examples causes the chain to get more easily derailed into regions outside the queried object set. This decreases the judged probability of the queried object set (superadditivity). The strength of these effects theoretically diminishes with the number of samples, as the chain approaches its stationary distribution. Accordingly, experimental manipulations that putatively reduce the number of samples, such as response deadlines and cognitive load, moderate this effect⁷. The experiments reported in this paper build on these findings, using subadditivity and superadditivity in the scene inference paradigm to detect behavioral signatures of amortized inference.

2.1.2 AMORTIZED INFERENCE

As defined in the previous section, Monte Carlo sampling is memoryless, approximating $P(h|d)$ without reference to other conditional distributions that have been computed in the past; all the hypothesis samples are specific to a particular query, and thus there can be no cumulative improvement in approximation accuracy across multiple queries. However, a moment's reflection suggests that

people are capable of such improvement. Every time you look out your window, you see a slightly different scene, but it would be wasteful to recompute a posterior over objects from scratch each time; if you did, you would be no faster at recognizing and locating objects the millionth time compared to the first time. Indeed, experimental research has found considerable speed-ups in object recognition and visual search when statistical regularities can be exploited⁷.

Amortized inference is a generalization of the standard memoryless framework. We will formulate it in the most general possible terms, and later explore more specific variants. Figure ?? illustrates the basic idea. In the standard, memoryless framework, an inference engine inverts a generative model $P(d, h)$ over hypothesis h and data d to compute a recognition model $Q(h|d)$ parametrized by ϑ . For example, Monte Carlo methods use a set of samples to parametrize the recognition model. Importantly, the answer to each query is approximated using a different set of parameters (e.g., independent samples)— $Q_1(h|d_1)$, $Q_2(h|d_2)$, etc. In the amortized framework, parameters are shared across queries. The parameters are selected to accurately approximate not just a single query, but a *distribution* of queries. If cognitive resources are unbounded, then the optimal solution is to parametrize each query separately, thereby recovering the memoryless framework. Under bounded resources, a finite number of parameters must be shared between multiple queries, leading to memory effects: the answer to one query affects the answer to other, similar queries.

While reuse increases computational efficiency, it can cause errors in two ways. First, if amortization is deployed not only when two queries are identical but also when they are similar, then answers will be biased due to blurring together of the distributions. This is analogous to interference effects in memory. Second, the answer to the first query might itself have been inaccurate or biased,

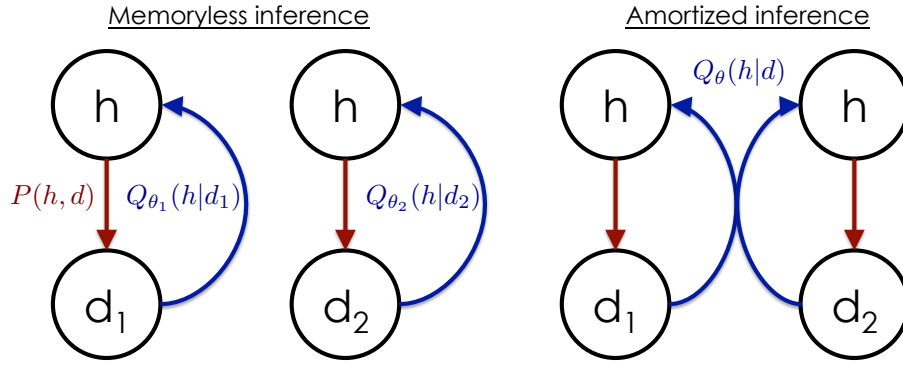


Figure 2.2: Theory schematic. (Left) Standard, memoryless framework in which a recognition model $Q(h|d)$ approximates the posterior over hypothesis h given data d . The recognition model is parametrized by \mathcal{D} (e.g., a set of samples in the case of Monte Carlo methods). Memoryless inference builds a separate recognition model for each query. (Right) Amortized framework, in which the recognition model shares parameters across queries. After each new query, the recognition model updates the shared parameters. In this way, the model “learns to infer.”

so its reuse will propagate that inaccuracy to the second query’s answer. Our experiments focus on the second type of error. Specifically, we will investigate how the over- or underestimation of unpacked probabilities resulting from approximate inference for one query will continue to influence responses to a second query.

2.1.3 TWO AMORTIZATION STRATEGIES

In our experiments, we ask participants to sequentially answer pairs of queries (denoted Q_1 and Q_2).

In Experiment 2, both queries are conditioned on the same cue object (d), but with varying query object sets (h). That is, both questions are querying the same probability distribution over objects, but eliciting the probabilities of different objects in each case. So in theory, all samples taken to answer query 1, can be reused to answer query 2 (they are both samples from the same distribution).

This *sample reuse* strategy allows all computations carried out for query 1 to be reused to answer

query 2.² However, it is expensive, because each sample must be stored in memory. A less memory-intensive solution is to store and reuse summary statistics of the generated samples, rather than the samples themselves. This *summary reuse* strategy offers greater efficiency but less flexibility. Several more sophisticated amortization schemes have been developed in the machine learning literature e.g.,^{3,4}, but we focus on sample and summary reuse because they make clear experimental predictions, which we elaborate below.

In the context of our experiments, summary reuse is only applicable to problems where the answer to Q_2 can be expressed as the composition of the answer to Q_1 and another (putatively simpler) computation. In Experiment 2, Q_2 queries a hypothesis space that is the union of the hypothesis space queried in Q_1 and a disjoint hypothesis space. For example if Q_1 is “What is the probability that there is an object starting with a C in the scene?”, Q_2 could be “What is the probability that there is an object starting with a C or an R in the scene?”. In this case, samples generated in response to Q_1 are summarized by a single number (“the probability of an object starting with C”), new samples are generated in response to a simpler query (“the probability of an object starting with R”), and these two numbers are then composed (in this case added) to give the final estimate for Q_2 (“the probability of an object starting with C or R”). This is possible because both queries are functions of the same probability distribution over latent objects.

These strategies are simplifications of what the brain is likely doing. Reusing all the samples ex-

²We focus on sampling-based amortization strategies because our earlier experiments support the idea that human probability judgment is sample-based³. However, amortization strategies can be realized without any form of sampling. These typically reduce time complexity by re-using a feedforward mapping from inputs to probabilities that replaces a more expensive form of iterative computation (e.g., message passing).

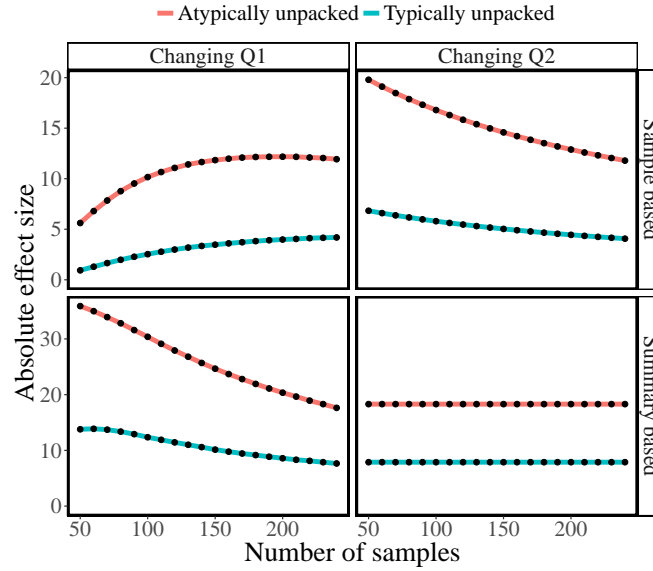


Figure 2.3: Simulation of subadditivity and superadditivity effects under sample-based (top) and summary-based (bottom) amortization strategies. In all panels, the y-axis represents the unstandardized effect size for Q_2 . Left panels show the effects of changing the sample size for Q_1 ; right panels show the effects of changing the sample size for Q_2 . When sample size for one query is changed, sample size for the other query is held fixed at 230 the sample size estimated by[?].

actly would involve their storage and is very intensive in its use of memory – in this aspect they are similar to exemplar models of categorization^{??}. While reusing only the summary statistic is much less memory intensive, it is unreasonably inflexible to restrict reuse of only the exact statistic in the few places that the second query can be expressed as a composition of the first query and a simpler computation. We do not claim that either extreme is plausible, but —to a first approximation— they capture the key ideas motivating our theoretical framework, and more importantly, they make testable predictions which can be used to assess which extreme pulls more weight in controlled experiments.

In particular, sample-based and summary-based amortization strategies make different predic-

tions about how subadditivity and superadditivity change as a function of the sample size (Figure ??, details of these implementations can be found in the Appendix). For sample-based amortization, as the sample size for Q_1 grows, the effect for Q_2 asymptotically *diminishes* and eventually vanishes as the effect of biased initialization in Q_1 washes out. However, initially increasing the sample size for Q_1 also *amplifies* the effects for Q_2 under a sample-based scheme, because this leads to more biased Q_1 samples being available for reuse. The amplification effect dominates up to a sample size of around 230 estimate for the number of samples taken for inference in this domain, reported in². This effect can be counteracted by increasing the sample size for Q_2 . These are unbiased samples, since Q_2 is always presented as a packed query. More such samples will push the effect down by drowning out the bias with additional unbiased samples.

Under a summary-based strategy, increasing the sample size for Q_1 will only *diminish* the effects for Q_2 , because the bias from Q_1 is strongest when the chain is close to its starting point. The effect of early, biased samples on the summary statistic disappears with more samples. We see also that changing the number of samples for Q_2 does not influence the effect size because the initialization of the chain for Q_2 is not influenced by the samples or summary statistic from the answer to Q_1 . Under the summary-based strategy, the subadditivity and superadditivity effects for Q_2 derive entirely from the same effects for Q_1 , which themselves are driven by the initialization see³.

We test the different predictions of these strategies by placing people under cognitive load during either Q_1 or Q_2 in Experiment 2, a manipulation that is expected to reduce the number of produced samples^{3,4}. In this way, we can sample different parts of the curves shown in Figure ??.

2.1.4 ADAPTIVE AMORTIZATION

Amortization is not always useful. As we have already mentioned, it can introduce systematic bias into probabilistic judgments. This is especially true if samples or summary statistics are transferred between two dissimilar distributions. This raises the question: are human amortization algorithms adaptive? This question is taken up empirically in Experiment 3. Here we discuss some of the theoretical issues.

Truly adaptive amortization requires a method to assess similarities between queries. Imagine as an example the situation in which there is a “chair” in the scene and you have to evaluate the probability of any object starting with a “P”. If afterwards you are told that there is a “book” in another scene, and the task is again to evaluate the probability of any object starting with a “P”, it could be a viable strategy to reuse at least some of the previous computations. However, in order to do so efficiently, you would have to know how similar a chair is to a book, i.e. if they occur with a similar set of other objects on average. One way to quantify this similarity is by assessing the induced posterior over all objects conditioned on either “book” or “chair”, and then comparing the two resulting distributions directly. Cues that are more similar should co-occur with other objects in similar proportions.

To assess the similarity of two distributions over objects induced by two different cues, we will need a formal similarity measure. One frequently used measure of similarity between two probability distribution is the Kullback-Leibler (KL) divergence. For two discrete probability distributions Q

and P , the KL divergence between P and Q is defined as

$$D_{\text{KL}}(P||Q) = \sum_b P(b) \log \frac{P(b)}{Q(b)}. \quad (2.5)$$

The KL divergence is minimized to 0 when Q and P are identical. We will use this measure in Experiment 3 to select queries that are either similar or dissimilar, in order to examine whether our participants only exhibit signatures of amortization when the queries are similar.³ Note, however, that the exact calculation of these divergences cannot be part of the algorithmic machinery used by humans to assess similarity, since it presupposes access to the posterior being approximated. Our experiments do not yet provide insight into how humans might algorithmically achieve tractable adaptive amortization, a problem we leave to future research.

2.2 EXPERIMENT 1

In Experiment 1, we sought initial confirmation of our central hypothesis: human inference is not memoryless. To detect these “remembrances of inferences past”, we asked participants to answer pairs of queries sequentially. The first query was manipulated (by packing or unpacking the queried hypothesis) in such a way that subadditive or superadditive probability judgments could be elicited². Crucially, the second query is always presented in packed form, so any differences across the experimental conditions in answers to the second query can only be attributed to the lingering effects of

³Our findings do not strongly depend on the use of the KL divergence measure and all of our qualitative effects remained unchanged when we applied a symmetric distance measure such as the Jensen-Shannon divergence.

the first query.

2.2.1 PARTICIPANTS

84 participants (53 males, mean age=32.61, SD=8.79) were recruited via Amazon’s Mechanical Turk and received \$0.50 for their participation plus an additional bonus of \$0.10 for every on-time response. The sample size for this and all of the following experiments was determined before data collection commenced. We decided to collect more participants than in our earlier work³ as the sub- and superadditivity effects might be weaker for the amortized answers to the second query.

2.2.2 PROCEDURE

Participants were asked to imagine playing a game in which their friend sees a photo and then mentions one particular object present in the photo (the cued object). The participant is then queried about the probability that another class of objects (e.g., “objects beginning with the letter B”) is also present in the photo.

Each participant completed 6 trials,⁴ where the stimuli on every trial corresponded to the rows in Table ???. On each trial, participants first answered Q_1 given the cued object (for example, “I see a lamp in this photo. What is the probability that I also see a window, a wardrobe, a wine rack, or any other object starting with a W?”), using a slider bar to report the conditional probability using values between 0 (not at all likely) to 100 (very likely, see also Figure ??).

⁴Note that participants were not directly informed that two consecutive trials are related and were therefore instructed that there would be 12 trials in total as there are two queries per query pair.

Figure 2.4: Experimental setup. Participants were asked to estimate the conditional probability using a slider bar within a 20-second time limit.

Table 2.2: Experimental stimuli and queries for Experiment 1.

Cue	Q_1	Q_1 -Typical	Q_1 -Atypical	Q_2
Table	C	chair, computer, curtain	cannon, cow, canoe	C or R
Telephone	D	display case, dresser, desk	drinking fountain, dryer, dome	D or L
Rug	B	book, bouquet, bed	bird, buffalo, bicycle	B or S
Chair	P	painting, plant, printer	porch, pie, platform	P or A
Sink	T	table, towel, toilet	trumpet, toll gate, trunk	T or E
Lamp	W	window, wardrobe, wine rack	wheelbarrow, water fountain, windmill	W or F

The Q_1 framing (typical-unpacked, atypical-unpacked or packed) was chosen randomly. Participants then completed the same procedure for Q_2 (immediately after Q_1), conditional on the same cued object. The framing for Q_2 was always packed and Q_2 was always presented as a conjunction (for example, “What is the probability I see an object starting with a W or F?”), where the order of the letters was determined at random.

Data for this experiment and all subsequent experiments reported in this article were submitted along with the final manuscript.

2.2.3 RESULTS

Six participants were excluded from the following analysis, four of whom failed to respond on time in more than half of the questions, and two of whom entered the same response throughout.

We applied one-sided hypothesis testing for all hypothesis involving sub- and superadditivity effects as these effects only make sense when assessed directionally.

Consistent with our previous studies², we found both subadditivity and superadditivity effects for Q_1 , depending on the unpacking: probability judgments were higher for unpacked-typical queries than for packed queries (a subadditivity effect; 59.35 vs. 49.67; $t(77) = 4.03, p < 0.001$) and lower for unpacked-atypical than for packed queries (a superadditivity effect; 31.42 vs. 49.67; $t(77) = -6.44, p < 0.001$).

Next we calculated the difference between each participant's response to every query and the mean packed response to the same queried object. This difference was then entered as a dependent variable into a linear mixed effects regression with random effects for both participants and queried objects as well as a fixed effect for the condition. The resulting estimates revealed both a significant subadditivity effect (difference = 12.60 ± 1.25 , $t(610.49) = 10.083, p < 0.0001$) and superadditivity effect (difference = -15.69 ± 1.32 , $t(615.46) = -11.89, p < 0.0001$).

Additionally, we found evidence that participants reused calculations from Q_1 for Q_2 : even though all Q_2 queries were presented in the same format (as packed), the estimates for that query differed depending on how Q_1 was presented. In particular, estimates for Q_2 were lower when Q_1 was unpacked to atypical exemplars (46.38 vs 56.83; $t(77) = 5.08, p < 0.001$), demonstrating a

superadditivity effect that carried over from one query to the next. We did not find an analogous carry-over effect for subadditivity (58.47 vs. 56.83; $t(77) = 0.72, p = 0.4$). This is possibly due to the subadditivity effect “washing out” more quickly (i.e. with fewer samples) than superadditivity, as has been observed in this domain before see⁵.

We calculated the difference between each participant’s response for every Q_2 and the mean response for the same object averaged over all responses to Q_2 conditional on Q_1 being packed. The resulting difference was again entered as the dependent variable into a linear mixed effects regression with both participants and cued object as random effects as well as condition as a fixed effect. The resulting estimates showed both a significant subadditivity effect (difference = 4.39 ± 1.14 , $t(606.40) = 3.83, p < 0.001$) and superadditivity effect (difference = -7.86 ± 1.21 , $t(610.41) = -6.50, p < 0.0001$).

We calculated each participant’s mean response to all packed hypotheses for Q_2 over all trials as a baseline measure and then assessed the difference between each condition’s mean response and this mean packed response. This resulted in a measure of an average effect size for the Q_2 responses (how much each participant under- or overestimates different hypotheses as compared to an average packed hypothesis). Results of this calculation are shown in Figure ??.

The superadditivity effect was significantly greater than 0 ($t(77) = 5.07, p < 0.001$). However, the subadditivity effect did not differ significantly from 0 ($t(77) = -0.42, p > 0.6$; see also⁵).

⁵The extent and direction of this asymmetry depends on the difference between how many samples it takes on average to get out of modes once the chain is in them (the root cause of subadditivity), and how many samples it takes on average to find high probability areas when the chain is far away from them (the root cause of superadditivity).

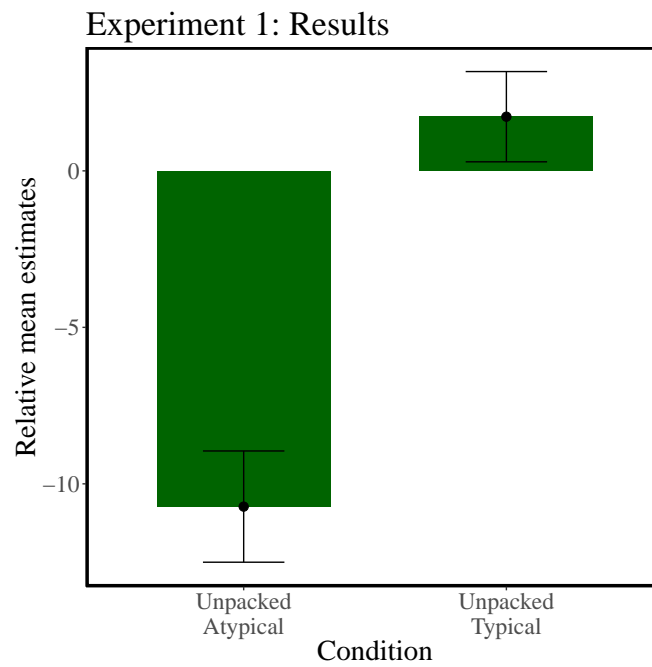


Figure 2.5: Experiment 1: Differences between Q_2 responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

Next, we explored whether responses to Q_1 predicted trial-by-trial variation in responses to Q_2 . Figure ?? shows the difference between participants' estimates for Q_1 and the true underlying probability of the query (as derived by letting our MCMC model run until convergence) plotted against the same difference for Q_2 .⁶ If participants do indeed reuse computations, then how much their estimates deviate from the underlying truth for Q_1 should be predictive of the deviance of their estimates for Q_2 .

We found significant positive correlations between the two queries in all conditions when aggregating across participants (average correlation: $r = 0.67, p < 0.01$). The same conclusion can be drawn from analyzing correlations within participants and then testing the average correlation against 0 ($r = 0.55, p < 0.01$). Moreover, the within-participant effect size (the response difference between the unpacked conditions and the packed condition) for Q_1 was correlated with responses to Q_2 for both atypical ($r = 0.35, p < 0.01$) and typical ($r = 0.21, p < 0.05$) unpacking conditions. This means that participants who showed greater subadditivity or superadditivity for Q_1 also showed correspondingly greater effects for Q_2 .

2.2.4 DISCUSSION

Experiment 1 established a memory effect in probabilistic inference: answers to a query are influenced by answers to a previous query, thereby providing evidence for amortization. In particular, both a sub- and a superadditivity effect induced at Q_1 carried over to Q_2 , and participants showing

⁶Although we did not initially plan to perform the analysis using difference scores, we believe that this is the correct way to report our results as it takes into account the mean differences between the judgments. In fact, performing the correction actually lead to smaller correlations and weaker effects overall as compared to using the raw values.

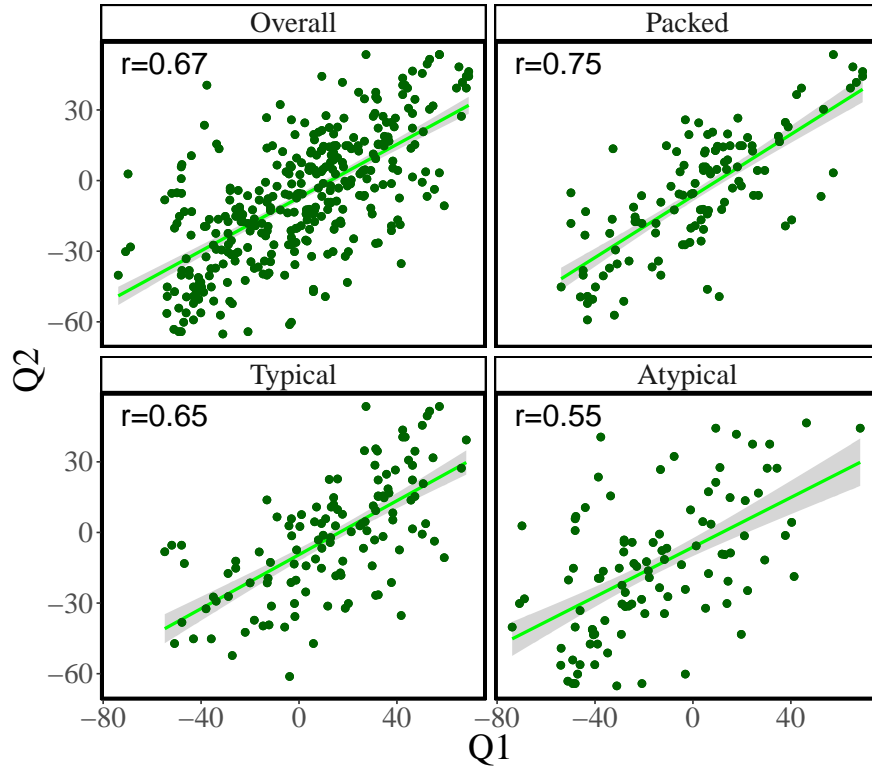


Figure 2.6: Trial-by-trial analyses of Experiment 1. Difference between Q_1 responses and true probability (as assessed by our MCMC model) plotted against the same quantity for Q_2 . Lines show the least-squares fit with standard error bands.

stronger effect sizes for both sub- and superadditivity for Q_1 also showed greater effects for Q_2 .

2.3 EXPERIMENT 2

Our next experiment sought to discriminate between sample-based and summary-based amortization strategies. We follow the logic of the simulations shown in Figure ??, manipulating cognitive load at Q_1 and Q_2 in order to exogenously control the number of samples seen^{2,2} for a similar approach.

In addition to cognitive load, we manipulate the “overlap” of Q_1 with Q_2 , by creating a new set of queries with no overlap between the hypothesis spaces of the query pairs. We predicted that we would only see a memory effect for queries with overlapping pairs. This manipulation allows us to rule out an alternative trivial explanation of our results: numerical anchoring (high answers to the first query lead to high answers to the second query). If the apparent memory effect was just due to anchoring, we would expect to see the effect regardless of query overlap, contrary to our predictions.

2.3.1 PARTICIPANTS

80 participants (53 males, mean age=32.96, SD=11.56) were recruited from Amazon Mechanical Turk and received \$0.50 as a basic participation fee and an additional bonus of \$0.10 for every on time response as well as \$0.10 for every correctly classified digit during cognitive load trials.

Table 2.3: Experimental stimuli and queries for Experiment 2.

Cue	Q_1	Q_1 -Typical	Q_1 -Atypical	Q_2 -Partial overlap	Q_2 -No overlap
Table	C	chair, computer, curtain	cannon, cow, canoe	C or R	T or R
Telephone	D	display case, dresser, desk	drinking fountain, dryer, dome	D or L	G or L
Rug	B	book, bouquet, bed	bird, buffalo, bicycle	B or S	D or S
Chair	P	painting, plant, printer	porch, pie, platform	P or A	M or A
Sink	T	table, towel, toilet	trumpet, toll gate, trunk	T or E	F or E
Lamp	W	window, wardrobe, wine rack	wheelbarrow, water fountain, windmill	W or F	L or F

2.3.2 PROCEDURE

The procedure in Experiment 2 was largely the same as in Experiment 1, with the following differences. To probe if the memory effects arise from reuse or from numerical anchoring, we added several Q_2 queries to the list shown in Table ???. These Q_2 queries have no overlap with the queried hypothesis for Q_1 (for example, 'T or R' instead of 'C or R' in the trial shown in the first row in Table ???). In other words, these queries could not be decomposed such that the biased samples from Q_1 would be reflected in the answer to Q_2 , so the sub- and super-additive effects would not be seen to carry over to Q_2 were reuse to occur. We refer to these queries as “no overlap”, in contrast to the other “partial overlap” queries in which one of the letters overlapped with the previously queried letter. Half of the queries had no overlap and half had partial overlap, randomly interspersed. The stimuli used in Experiment 2 are shown in Table ???.

To probe if the memory effect arises from reuse of generated samples (sample-based amortization) or reuse of summaries (summary-based amortization), we also manipulated cognitive load: on half of the trials, the cognitive load manipulation occurred at Q_1 and on half at Q_2 . A sequence of 3 different digits was presented prior to the query, where each of the digits remained on the screen for 1 second and then vanished. After their response to the query, participants were asked to make a

same/different judgment about a probe sequence. Half of the probes were old and half were new.

We hypothesized that partial overlap would lead to stronger amortization effects, whereas no overlap would lead to weaker effects. Furthermore, if participants are utilizing sample-based amortization, then cognitive load during Q_2 should increase the amortization effect: if more samples are generated during Q_1 (which are the samples that contain the sub- or superadditivity biases) and these samples are concatenated with fewer unbiased samples during Q_2 , then the combined samples will be dominated by biased samples from Q_1 and therefore show stronger effects. Vice versa, if participants are utilizing summary-based amortization, then cognitive load during Q_1 should increase the amortization effect: if less samples are generated during Q_1 , then a summary of those samples will inherit a stronger sub- or superadditivity effect such that the overall amortization effect will be stronger if the two summaries are combined (assuming that the summaries are combined with equal or close-to equal weights).

2.3.3 RESULTS

Analyzing only the queries with partial overlap (averaging across load conditions), we found that probability judgments for Q_1 were higher for unpacked-typical compared to packed conditions (a subadditivity effect; $t(79) = 4.38, p < 0.001$) and lower for unpacked-atypical compared to packed (a superadditivity effect; $t(79) = -4.94, p < 0.001$). These same effects occurred for Q_2 (unpacked-typical vs. packed: $t(79) = 2.44, p < 0.01$; unpacked-atypical vs. packed: $t(79) = -1.93, p < 0.05$).

We again calculated the difference between each participant's response to every query during Q_1 and the overall mean response for the same query object in the packed condition. This difference

was then used as the dependent variable in a linear mixed-effects regression model with participants and queried object as random effects and condition as fixed effect. The resulting estimates showed both a significant subadditivity effect (difference = 13.64 ± 1.57 , $t(396.95) = 8.70$, $p < 0.0001$) and superadditivity effect (-14.90 ± 1.56 , $t(395.48) = -9.55$, $p < 0.0001$). Afterwards, we repeated the same analysis for responses to Q_2 (as in Experiment 1). This analysis again showed significant indicators of amortization as both the subadditivity (difference = 5.37 ± 1.34 , $t(398.01) = 4.02$, $p < 0.001$) and the superadditivity effect (difference = -4.92 ± 1.34 , $t(398.01) = -3.69$, $p < 0.001$) were still present during Q_2 .

Next, we assessed how the memory effect was modulated by cognitive load and overlap (Figure ??). When cognitive load occurred during Q_2 and there was no overlap, none of the conditions produced an effect significantly different from 0 (all $p > 0.05$). When cognitive load occurred during Q_2 and there was partial overlap, only typically unpacked hypotheses produced an effect significantly greater than 0 ($t(38) = 2.14$, $p < 0.05$). When cognitive load occurred during Q_1 and there was no overlap, we again found no evidence that the conditions differ from 0 (all $p > 0.05$). Crucially, if cognitive load occurred during Q_1 and there was partial overlap, both conditions showed the expected subadditive ($t(38) = 4.18$, $p < 0.05$) and superadditive ($t(46) = -1.89$, $p < 0.05$) effects. Moreover, calculating the average effect size of amortization for the different quadrants of Figure ??, the partial overlap-cognitive load at Q_1 condition produced the highest overall effect ($d = 0.8$), followed by the partial overlap-cognitive load at Q_2 condition ($d = 0.56$) and the no overlap-cognitive load at Q_1 condition ($d = 0.42$). The no overlap-cognitive load at Q_2 condition did not produce an effect greater than 0. Partial overlap trials were also more strongly correlated

Experiment 2: Results

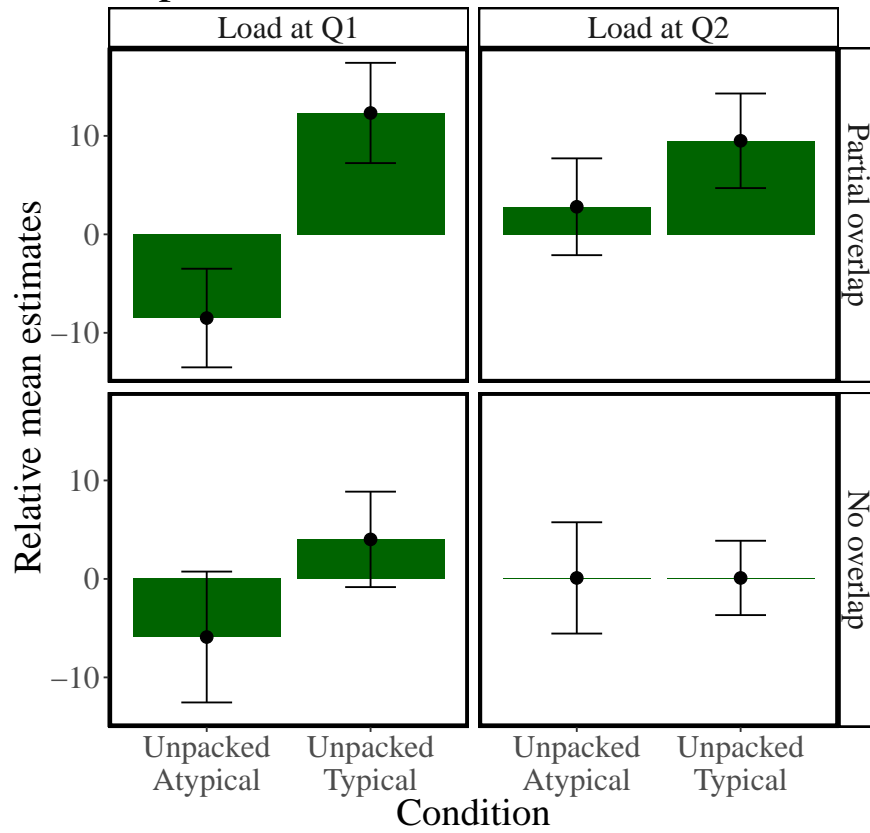


Figure 2.7: Experiment 2: Differences between Q_2 responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

with responses during Q_1 than trials with no overlap (0.41 vs 0.15 , $t(79) = -2.1$, $p < 0.05$).

Next, we calculated the difference between all responses to Q_2 and the mean responses to Q_2 over queried objects provided that Q_1 was packed. This difference was entered into a linear mixed-effects regression that contained overlap, cognitive load, and the presentation format of Q_1 as fixed effects, and participants and the queried objects as random effects. We then assessed the interaction between cognitive load and the sub- and superadditivity conditions while controlling for overlap. The resulting estimates showed that there was a significant subadditivity effect (difference = 5.25 ± 2.12 , $t(417.08) = 2.48$, $p < 0.05$) but no superadditivity effect (difference = -3.19 ± 2.17 , $t(419.23) = -1.47$, $p = 0.17$) when cognitive load was applied during Q_2 . Importantly, both the subadditivity (difference = 5.83 ± 2.25 , $t(418.91) = 2.59$, $p < 0.05$) and the superadditivity (difference = -6.86 ± 2.21 , $t(419.80) = -3.102$, $p < 0.01$) effects were present when cognitive load was applied during Q_1 . This finding points towards a larger amortization effect in the presence of cognitive load on Q_1 , thus supporting a summary-based over a sampled-based amortization scheme.

Further, on trials with cognitive load at Q_2 , participants were on average more likely to answer the probe correctly for partial overlap trials compared to no overlap trials ($t(36) = 3.16$, $p < 0.05$). This is another signature of amortization: participants are expected to have more resources to spare for the memory task at Q_2 if the computations they executed for Q_1 are reusable in answering Q_2 . This also indicates that these results cannot be explained by simply initializing the chain for Q_2 where the chain for Q_1 ended, which would not have affected the required computations.

Interestingly, there was no evidence for a significant difference between participants' responses to Q_2 under cognitive load in Experiment 2 as compared to participants' responses to Q_2 in Experi-

ment 1 when no cognitive load during either Q_1 or Q_2 was applied ($t(314) = -1.44, p = 0.15$).

Finally, we assessed how much the difference between responses for Q_1 and the true underlying probabilities were predictive of the difference between responses for Q_2 and the true underlying probabilities (Figure ??). There was a strong correlation between responses to Q_1 and Q_2 over all conditions ($r = 0.41, p < 0.001$), for the packed ($r = 0.44, p < 0.001$), the typically unpacked ($r = 0.36, p < 0.01$), as well as the atypically unpacked condition ($r = 0.40, p < 0.01$). Moreover, the differences of Q_1 and Q_2 responses from the true answer were also correlated within participants (mean $r = 0.31, p < 0.01$) and participants who showed stronger subadditivity or superadditivity effects for Q_1 also showed stronger effects for Q_2 overall ($r = 0.31, p < 0.001$), for the superadditive ($r = 0.3, p < 0.001$), and for the subadditive condition ($r = 0.29, p < 0.001$). This replicates the effects of amortization found in Experiment 1.

2.3.4 DISCUSSION

Experiment 2 extended the findings of Experiment 1, suggesting constraints on the underlying amortization strategy. Participants exhibited an intricate pattern of sensitivity to cognitive load and query overlap. Based on our simulations (Figure ??), we argue that the effect of cognitive load at Q_1 on Q_2 responses is more consistent with summary-based amortization than with sample-based amortization. Summary-based amortization is less flexible than sample-based amortization, but trades this inference limitation for an increase in memory efficiency, and is thus consistent with the idea that humans adopt cost-efficient resource-rational inference strategies^{???}. Further supporting this idea is our finding that performance on the secondary task was better in the partial overlap conditions,

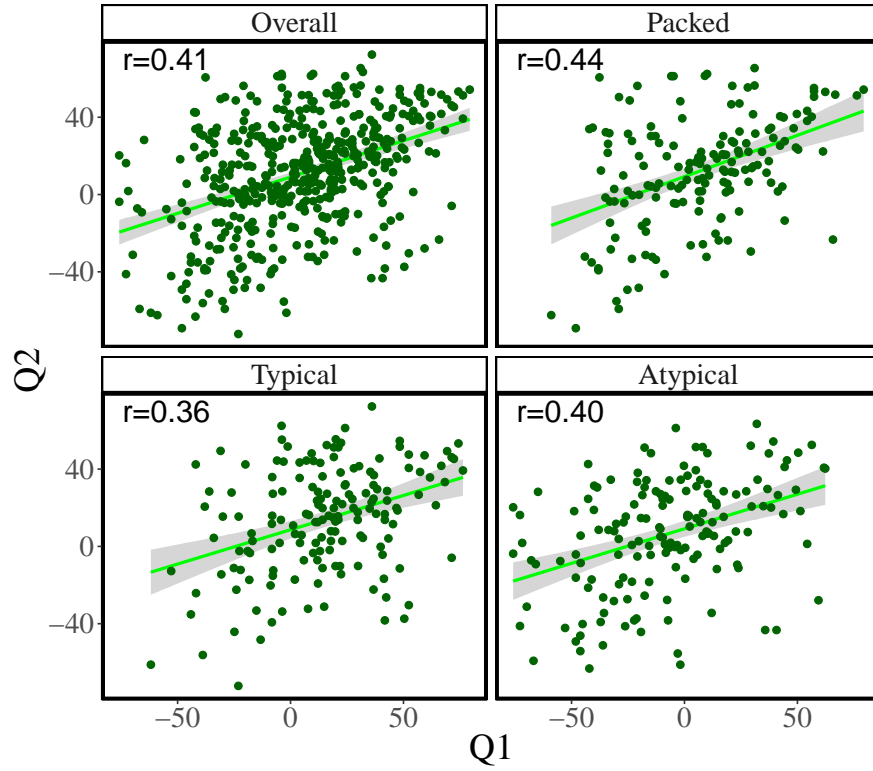


Figure 2.8: Trial-by-trial analyses of Experiment 2. Relationship between difference between Q_1 responses and true probability (as assessed by our MCMC model) and Q_2 responses and true probability. Lines show the least-squares fit with standard error bands.

indicating that more resources are available when computations can be amortized.

Our design allowed us to rule out a numerical anchoring effect, whereby participants would give high answers to the second query if they gave high answers to the first query. This effect should be invariant to the extent of overlap of the queried hypothesis spaces, but contrary to the anchoring hypothesis, the memory effect was stronger in the high overlap condition.

2.4 EXPERIMENT 3

In this experiment we further probe the strategic nature of amortization. So far, all generated hypotheses have been reusable, since both queries probe the same probability distribution, conditioned on the same cue object. By changing the cue object between Q_1 and Q_2 and manipulating the similarity between the cues, we can control how reusable the computations are. Note that this is in contrast to the notion of “overlap” in Experiment 2 where all the samples from Q_1 are always “reusable” in Q_2 since both query the same probability distribution, but in the no overlap conditions, the queried hypotheses spaces do not overlap resulting in the biased samples from Q_1 not being reflected in Q_2 judgments. The notion of reusability now allows us to test whether or not reuse always occurs, or if it occurs preferentially when it is more applicable (i.e., under high similarity between cues).

2.4.1 PARTICIPANTS

100 participants (41 females, mean age=35.74, SD=11.69) were recruited from Amazon Mechanical Turk and received \$0.50 as a basic participation fee and an additional bonus of \$0.10 for every on

Table 2.4: Experimental stimuli and queries for Experiment 3. Kullback-Leibler (KL) divergence between the posteriors for Q_1 and Q_2 are shown in parentheses.

Cue1	Q_1	Q_1 -Typical	Q_1 -Atypical	Cue2-sim (KL)	Cue2-diff (KL)
Rug	B	book, bouquet, bed	bird, buffalo, bicycle	Curtain (0.080)	Car (8.658)
Chair	P	painting, plant, printer	porch, pie, platform	Book (0.031)	Road (8.508)
Sink	T	table, towel, toilet	trumpet, toll gate, trunk	Counter (0.001)	Sidewalk (8.503)

time response.

2.4.2 PROCEDURE

The procedure was similar to Experiments 1 and 2. The only difference was that participants were shown a new cue word for Q_2 , asking them to judge the probability of objects starting with the same letter as the letter from Q_1 with no conjunction of letters (i.e., same query space, full overlap).

The query for Q_2 was always packed, as in previous experiments. The new cue words for Q_2 were generated to either have posterior with a low (similar cues) or a high (dissimilar cues) KL divergence from the Q_1 posterior. The range of KL divergences fell between 0 and 9; all similar cue words had conditional distributions with KL divergence of less than 0.1, and all dissimilar cue-words had a KL divergence of greater than 8.5. The exact KL divergences are reported in Table ??.

2.4.3 RESULTS

Seven participants did not respond on time to more than a half of all queries and were therefore excluded from the following analysis.

We again found that probability judgments for Q_1 in the typically unpacked queries were higher than in the unpacked condition (subadditivity effect: $t(92) = 4.67, p < 0.001$) and that proba-

bility judgments in the atypically unpacked condition were lower than in the unpacked condition (superadditivity effect: $t(92) = 3.25, p < 0.01$).

Analyzing the probability judgments for Q_2 , we found a significant subadditivity effect ($t(92) = 2.28, p < 0.05$) but not a significant superadditivity effect (56.06 vs. 55.31; $t(92) = 0.07, p = 0.94$).

As before, we calculated the difference between each participant's response to every query during Q_1 and the overall mean response for the same query object in the packed condition. This difference was entered as the dependent variable into a linear mixed-effects regression model with participants and queried object as random effects and condition as fixed effect. The resulting estimates showed both a significant subadditivity effect (difference = $14.39 \pm 1.97, t(189.84) = 7.31, p < 0.0001$) and a superadditivity effect ($-13.72 \pm 1.98, t(190.18) = -6.941, p < 0.0001$). Repeating this analysis for responses to Q_2 revealed a significant amortization effect for the typically unpacked condition (difference = $5.21 \pm 1.90, t(191) = 2.74, p < 0.05$) but not for the atypically unpacked condition (difference = $-2.49 \pm 1.91, t(191.52) = -1.303, p = 0.19$).

For the dissimilar cues, we did not find statistical evidence for an effect of subadditivity ($t(49) = 1.31, p = 0.19$) or superadditivity ($t(47) = -0.27, p = 0.79$). However, for the similar cues at Q_2 , the effect for the typically unpacked condition was significantly different from 0 (subadditivity effect: $t(47) = 3.30, p < 0.01$), whereas there was again no superadditivity effect ($t(48) = 0.54, p = 0.59$). The difference between the size of the subadditivity effect was marginally bigger for the similar cues as compared to the dissimilar cues ($t(36) = 1.83, p = 0.07$) and the overall effect size of the similar cues was $d = 0.17$, whereas the effect size for the dissimilar cues was $d = 0.11$.

The difference between judgments and the true probabilities was correlated between Q_1 and Q_2

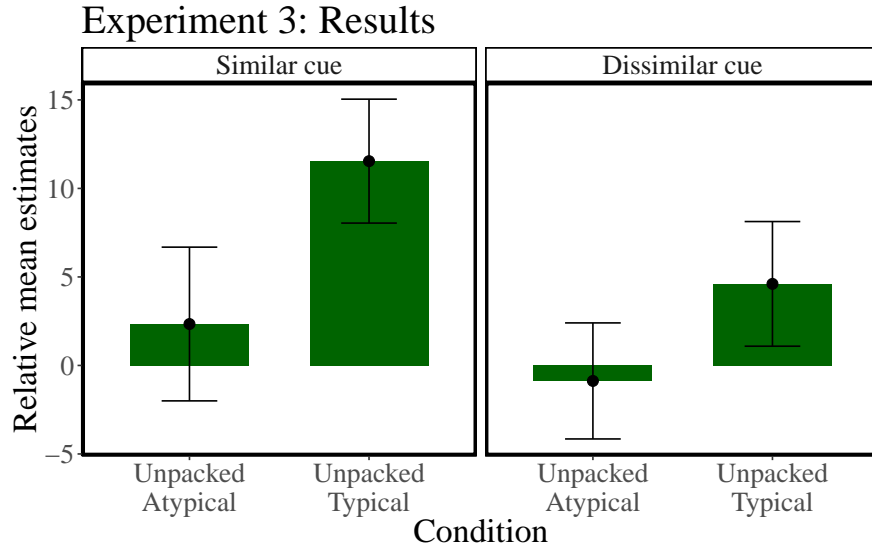


Figure 2.9: Experiment 3: Differences between Q_2 responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity effect and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

($r = 0.34, p < 0.001$), for the packed ($r = 0.43, p < 0.001$), the typically unpacked ($r = 0.43, p < 0.001$), but not the atypically unpacked condition ($r = 0.20, p = 0.3$); see Figure ??). Participants who showed higher subadditivity or superadditivity effects for Q_1 also showed higher effects for Q_2 overall ($r = 0.29, p < 0.001$), for the typically unpacked condition ($r = 0.39, p < 0.001$), but not for the atypically unpacked condition ($r = 0.11, p = 0.29$).

2.4.4 DISCUSSION

Experiment 3 assessed the strategic nature of amortization by manipulating the similarity between cues, which presumably affected the degree to which amortization is useful. We found a stronger subadditivity effect for similar cues compared to dissimilar cues, indicating that reuse is at least partially sensitive to similarity.

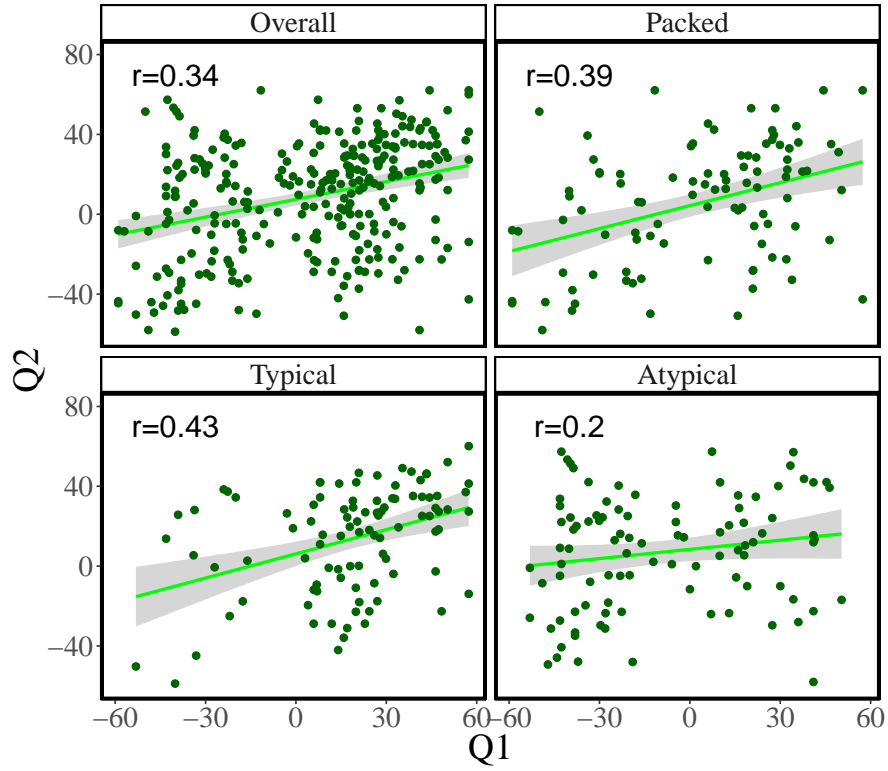


Figure 2.10: Trial-by-trial analyses of Experiment 3. Relationship between difference between Q_1 responses and true probability (as assessed by our MCMC model) and Q_2 responses and true probability. Lines show the least-squares fit with standard error bands.

An unexpected finding was that while the superadditivity effect in atypically-unpacked Q_1 was significant, neither the memory-based superadditivity effect (in Q_2) nor correlations across the queries for atypically-unpacked Q_1 were significant. This indicates that the answers to the atypically-unpacked Q_1 are not detectably being reused in Q_2 in this experiment. However, in Experiments 1 and 2, the atypically-unpacked answers seem to be reused (as indicated by a robust memory-based superadditivity effect, and correlations across the queries) *when the cue object remains the same*. This may be because the extent of rational reuse here (where the cues change) is smaller than in previous experiments (where the cues remained the same) and therefore harder to detect.

2.5 GENERAL DISCUSSION

We tested a model of amortized hypothesis generation across 3 experiments and found that participants not only exhibited subadditive and superadditive probability judgments in the same paradigm replicating³, but also carried over these effects to subsequent queries—a memory effect on inference. Experiment 2 demonstrated that this memory effect is some function of the hypotheses generated in the first query and made some inroads into trying to understand this function. We found that the effect is stronger when cognitive load is applied to the first query, suggesting that the memory effect is driven by a form of summary-based amortization, whereby a summary statistic of the first query is computed from the samples and then reused to answer subsequent queries, provided they can be expressed in terms of previous computations. Summary-based amortization gives up some flexibility (compared to reusing the raw samples generated by past inferences), in order to gain memory-

efficiency. Experiment 3 demonstrated that the memory effect selectively occurs when the queries are similar, indicating that reuse is deployed specifically when it is likely to be useful.

Building on earlier results², our findings support the existence of a sophisticated inference engine that adaptively exploits past computations. While reuse can introduce error, this error may be a natural consequence of a resource-bounded system that optimally balances accuracy and efficiency^{2,3,4}. The incorporation of reuse into a Monte Carlo sampling framework allows the inference engine to preserve asymptotic exactness while improving efficiency in the finite-sample regime.

2.5.1 RELATED WORK

This work fits into a larger nexus of ideas exploring the role of memory in inductive reasoning. Heit, Hayes and colleagues have carried out a number of studies that make this link explicit^{5,6,7}. For example, ⁵ developed a task in which participants studied a set of exemplars (large dogs that all possess “beta cells”) and then on a test set of exemplars (consisting of large and small dogs) made either property induction judgments (“does this dog have beta cells?”) or recognition memory judgments (“did this dog appear in the study phase?”). The key finding was that property induction and recognition memory judgments were strongly correlated across items, supporting the hypothesis that both judgments rely on a shared exemplar similarity computation: test exemplars are judged to be more familiar, and have the same latent properties, to the degree that they are similar to past exemplars. Heit and Hayes showed that both judgments could be captured by the same exemplar model, but with a broader generalization gradient for induction.

Another example of memory effects on inference is the observation that making a binary deci-

sion about a noisy stimulus (whether dots are moving to the left or right of a reference) influences a subsequent continuous judgment about motion direction[?]. Stocker and colleagues^{??} refer to this as “conditioned perception” or “self-consistent inference” because it appears as though observers are conditioning on their decision as they make a choice. ? have pushed this idea further, arguing that observers condition on their own confidence about the decision. Self-consistent inferences may reflect rational conditioning on choice or confidence information when a memory trace of the stimulus is unavailable or unreliable.

Another important consideration for the study of amortization is the utility conferred by reuse rather than simply the efficiency. Previous work has explored resource-rational solutions to balancing the utility of events with their probability of occurrence^{???}. These have successfully modeled effects such as the over-representation of low frequency events with extreme utilities, indicating a possible role for utility in availability for subsequent reuse.

An intriguing explanation of order effects has been reported by Wang and colleagues^{??}. The key idea, derived from a quantum probability model of cognition see also[?], is that answering a question will cause the corresponding mental state to linger and thus “superpose” with the mental state evoked by a second question. This superposition gives rise to a particular symmetry in the pattern of judgments when question order is manipulated, known as the *quantum question order equality* see[?] for details. Our amortization framework does not intrinsically make this prediction, but nor does it necessarily exclude it. Rather, we prefer to think about superposition states as arising from computational principles governing a computation-flexibility trade-off. Roughly speaking, states superpose in our framework because the inference engine is reusing information from past queries.

Recently, ? pointed out that the quantum question order equality could arise from rational probabilistic reasoning corrupted by correlated noise. In particular, answers to a probabilistic query will be corrupted by samples retrieved recently to answer another probabilistic query similar to the concept of “overgeneralization” in probabilistic estimation, as developed in[?]. ? view this as a kind of priming effect. Alternatively, correlated noise would arise in the amortized inference framework due to stochastic reuse. Thus, amortization might provide a complementary rational analysis for the “probability theory plus noise” model proposed by ?.

Most closely related to the present paper is the work of Dougherty and colleagues^{????}, who have pursued the idea that probability judgments depend on the generation of hypotheses from memory. In particular, they argue that subadditivity arises from the failure to generate hypotheses, much like the account offered by ?, and that this failure is exacerbated by cognitive load or low working memory capacity. The key difference from our account is the particular way in which memories are used to generate hypotheses. For combinatorial hypothesis spaces like the scene inference task used here and by ?, one cannot assume that all the relevant hypotheses are already stored in memory; rather, these must be generated on the fly—a function we ascribe to MCMC sampling,

where new hypotheses that have never been seen before can be generated from a probabilistic generative model, and only these generated samples need be stored for the purposes of inference. The present paper asserts a more direct role for memory within a sampling framework, by controlling the trade-off between computation and flexibility.

This trade-off mirrors a similar tension in reinforcement learning, where the goal is to estimate long-term reward^{???}. “Model-based” algorithms estimate long-term reward by applying tree search

or dynamic programming to a probabilistic model of the environment. This is flexible, but computationally expensive. “Model-free” algorithms avoid this cost by directly estimating long-term rewards by interacting with the environment, storing these estimates in a look-up table or function approximator. This is computationally cheap but inflexible. In other words, model-free algorithms trade time for space, much in the same way that amortized inference uses memory to reduce the cost of approximate inference. Analogous to our proposed summary-based amortization strategy, recent work has suggested that model-free value estimates can be incorporated into model-based tree search algorithms², thus occupying a middle ground in the time-space trade-off.

2.5.2 FUTURE DIRECTIONS

Our work has focused on fairly simple forms of amortization. There exists a much larger space of more sophisticated amortization strategies developed in the machine learning literature e.g.,²² that we have not yet explored. Finding behaviorally distinguishable versions of these algorithms is an interesting challenge. These versions could take the form of reuse in much more abstract ways, such as developing strategies and heuristics, instead of just local reuse in a sequence of queries. We believe that further examining established effects of heuristics and biases through the lens of computational rationality will continue to produce interesting insights into principles of cognition.

More broadly, we are still lacking a comprehensive, mechanistic theory of amortized inference. What objective function is being optimized by amortization? How are the computational trade-offs managed algorithmically? What are the contributions of different memory mechanisms (episodic, semantic, procedural, etc.)? Answering these questions will require a more general theoretical treat-

ment than the one offered here. Nonetheless, our experiments provide important constraints on any such theory.

ACKNOWLEDGMENTS

We thank Kevin Smith for helpful comments and discussions.

We would also like to thank Mike Oaksford and Michael Dougherty for useful feedback during the review process. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. E.S. was supported by a postdoctoral fellowship from the Harvard Data Science Initiative.

APPENDIX

TWO REUSE SCHEMES

The two schemes for reuse described in Figure ??, *summary-based* and *sample-based* amortization, are described below in greater detail.

In *sample-based amortization*, we simply add samples generated in response to one query (Q_1) to the sample set for another query (Q_2). So if N_1 samples were generated in response to Q_1 , and N_2 new samples are generated in response to Q_2 , in the absence of amortization, the responses to the two queries Q_1 and Q_2 would be generated as follows:

$$P(h_{Q_1}|d) \approx \frac{1}{N_1} \sum_{n=1}^{N_1} \mathbb{I}[h_n = h_{Q_1}]$$

$$P(h_{Q_2}|d) \approx \frac{1}{N_2} \sum_{n=1}^{N_2} \mathbb{I}[h_n = h_{Q_2}]$$

Under the sample-based amortization scheme, the response to Q_2 is given by a calculation carried out over all $N_1 + N_2$ equally weighted samples.

$$P(h_{Q_1}|d) \approx \frac{1}{N_1} \sum_{n=1}^{N_1} \mathbb{I}[h_n = h_{Q_1}]$$

$$P(h_{Q_2}|d) \approx \frac{1}{N_2 + N_1} \left(\sum_{n=1}^{N_1} \mathbb{I}[h_n = h_{Q_2}] + \sum_{n=1}^{N_2} \mathbb{I}[h_n = h_{Q_2}] \right)$$

Under this scheme, all the computations carried out for Q_1 are available for flexible reuse in the computation for Q_2 .

In *summary-based amortization*, we reuse a summary statistic computed from Q_1 . This strategy is only applicable to problems where the answer to Q_2 can be expressed as the composition of the answer to Q_1 , and an additional simpler computation. For example if Q_1 is “What is the probability that there is an object starting with a C in the scene?”, Q_2 could be “What is the probability that

there is an object starting with a C or an R in the scene?”. In this case, the N_1 samples generated in response to Q_1 are summarized into one probability (“the probability of an object starting with C”), N_2 new samples are generated in response to a simpler query (“the probability of an object starting with R”), and these two numbers are then composed (in this case simply added) to give the final estimate for Q_2 (“the probability of an object starting with C or R”).

$$P(h_{Q_1}|d) \approx \frac{1}{N_1} \sum_{n=1}^{N_1} \mathbb{I}[h_n = h_{Q_1}]$$

$$P(h_{Q_2}|d) \approx \frac{1}{N_1} \sum_{n=1}^{N_1} \mathbb{I}[h_n = h_{Q_1}] + \frac{1}{N_2} \sum_{n=1}^{N_2} \mathbb{I}[h_n = (h_{Q_2} - h_{Q_1})]$$

Under this scheme, only the final product of the computation carried out for Q_1 is reused in the calculations for Q_2 .

*Nulla facilisi. In vel sem. Morbi id urna in diam dignis-
sim feugiat. Proin molestie tortor eu velit. Aliquam erat
volutpat. Nullam ultrices, diam tempus vulputate egestas,
eros pede varius leo.*

Quoteauthor Lastname

3

Flexible amortization with a variational framework

Bayesian theories of cognition assume that people can integrate probabilities rationally. However, several empirical findings contradict this proposition: human probabilistic inferences are prone to systematic deviations from optimality. Puzzlingly, these deviations sometimes go in opposite

directions. Whereas some studies suggest that people under-react to prior probabilities (*base rate neglect*), other studies find that people under-react to the likelihood of the data (*conservatism*). We argue that these deviations arise because the human brain does not rely solely on a general-purpose mechanism for approximating Bayesian inference that is invariant across queries. Instead, the brain is equipped with a recognition model that maps queries to probability distributions. The parameters of this recognition model are optimized to get the output as close as possible, on average, to the true posterior. Because of our limited computational resources, the recognition model will allocate its resources so as to be more accurate for high probability queries than for low probability queries. By adapting to the query distribution, the recognition model “learns to infer.” We show that this theory can explain why and when people under-react to the data or the prior, and a new experiment demonstrates that these two forms of under-reaction can be systematically controlled by manipulating the query distribution. The theory also explains a range of related phenomena: memory effects, belief bias, and the structure of response variability in probabilistic reasoning. We also discuss how the theory can be integrated with prior sampling-based accounts of approximate inference.

3.1 INTRODUCTION

Studies of probabilistic reasoning frequently portray people as prone to errors^{???}. The cognitive processes that produce these errors is the subject of considerable debate^{???}. One influential class of models holds that rational probabilistic reasoning is too cognitively burdensome for people, who instead use a variety of heuristics^{???}. Alternatively, rational process models hold that errors arise from

principled approximations of rational reasoning, for example some form of hypothesis sampling^{???}. These different perspectives have some common ground; certain heuristics might be considered accurate approximations^{???}.

One challenge facing both heuristic and rational process models is that people appear to make different errors in different contexts. For example, some studies report *base rate neglect*^{???}, the finding that people under-react to prior probabilities relative to Bayes' rule. Other studies report *conservatism*^{??}, the finding that people under-react to evidence.¹

Heuristic models respond to this challenge by allowing heuristics to be context-sensitive, an example of *strategy selection*^{??}. Most models of strategy selection assume that people are able to assess the usefulness of a strategy, through cost-benefit analysis^{???}, reinforcement learning^{??}, or based on the strategy's applicability in a particular domain^{??}. All of these approaches require, either explicitly or implicitly, a feedback signal. This requirement poses a problem in inferential settings where no feedback is available. People can readily answer questions like "How likely is it that a newly invented machine could transform a rose into a blackbird?"[?] which lack an objective answer even in principle.

Most rational process models are based on domain-general algorithms, and thus struggle to explain the context-sensitivity of inferential errors see[?] for a similar argument. Some models explain why certain kinds of queries induce certain kinds of errors[?], but do not explain how errors can be modulated by other queries in the same context^{??}.

¹We will mostly avoid the term "conservatism" to denote under-reaction to data, because it is sometimes conflated with a bias to give "conservative" probability judgments (i.e., judgments close to uniform probability). These distinct phenomena make the same predictions only when the prior is uniform over hypotheses. We return to the second use of the term later in the article.

In this paper, we develop a new class of rational process models that explain the context-sensitivity of inferential errors. Specifically, we propose that people *learn to infer*. Instead of a domain-general inference algorithm that treats all queries equally, we postulate an approximate *recognition model*² that maps queries to posterior probabilities.² The parameters of this recognition model are optimized based on the distribution of queries, such that the output is on average as close as possible to the true posterior. This leads to learned biases in which sources of information to ignore, depending on which of these sources reliably co-vary with the true posterior.³ Importantly, this optimization is carried out without explicit feedback about the true posterior³.

Like other rational process models, our approach is motivated by the fact that any computationally realistic agent that performs inference in complex probabilistic models—in the real world, in real time—will need to make approximate inferences. Exact Bayesian inference is almost always impossible. “Learning to infer” refers to a particular approximate inference scheme, using a pattern recognition system (such as a neural network, but it could also be an exemplar generalization model) to find and exploit patterns in the conditional distribution of hypotheses given data (the posterior). We will argue that a relatively simple model of learned inference is both a good approximate inference scheme, purely on algorithmic terms, and also can account for a number of patterns of heuristic inference in the behavioral literature, where people have been observed to deviate from ideal Bayesian updating in ways that are otherwise hard to reconcile and even appear contradictory, because they

²When the recognition model is parametrized as a neural network, it is sometimes also referred to as an *inference network*^{2,3}.

³We focus on domains where we can control this covariance (of information sources with the posterior) within an experiment, to study the development of context-sensitive inferential errors. We also discuss how similar mechanisms could explain errors in more ‘real-world’ domains where this context is learned from experience before the experiment, based on ecological distributions of the relevant probabilities.

appear to deviate from Bayesian norms in different ways and in different contexts. Our theory of learning to infer explains why these contextual variations are observed, and why they *should* be observed, in a system designed to adapt efficient approximate inference to the environments it finds itself in.

The rest of the paper is organized as follows. We first summarize the empirical and theoretical literature on our motivating puzzle (under-reaction to prior vs. likelihood). We then introduce our new theory. In addition to addressing under-reaction, we show that the theory can explain a number of related phenomena: memory effects, belief bias, and the structure of response variability in probabilistic reasoning. In the Discussion, we connect our theory to previous accounts of approximate inference in human probabilistic reasoning.

3.2 UNDER-REACTION TO PROBABILISTIC INFORMATION

Given data d , Bayes' rule stipulates how a rational agent should update its prior probabilistic beliefs $P(h)$ about hypothesis h :

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}, \quad (3.1)$$

where $P(h|d)$ is the agent's posterior distribution, expressing its updated beliefs, and $P(d|h)$ is the likelihood, expressing the probability of the observed data under candidate hypothesis h .

The earliest studies of probabilistic belief updating, carried out by Ward Edwards and his students^{2,3}, asked subjects to imagine a set of 100 bags filled with blue and red poker chips. "Red" bags

were filled predominantly with red chips, and “blue” bags were filled predominantly with blue chips; the proportion of colors in each bag type was known to the subjects and manipulated experimentally. The subjects were told that one of the bags was randomly selected and a set of chips was randomly drawn from that bag. They then had to judge the probability that the observed chips came from each bag, by distributing 100 metal washers between two pegs. The proportion of washers on each peg was taken to be the subjective report of the corresponding probability. Closely related studies by Peterson and colleagues used a continuous slider as the response apparatus^{???}. It is important to emphasize that in these studies, subjects were given all the relevant information about the data-generating process necessary for computing the posterior. Thus, there should be no learning about the parameters of this process (i.e., the prior and likelihood).

Early on, it was evident that subjects were not exactly following Bayes’ rule in these experiments, despite being given all the information needed to compute it. In particular, subjects consistently under-reacted to the evidence, revising their beliefs less than mandated by Bayes’ rule (a phenomenon commonly referred to as “conservatism,” though we avoid this term for reasons explained in the Introduction). This phenomenon was robust across many variations of the basic experimental paradigm; later we will discuss a number of factors that influence the degree of under-reaction.

Several hypotheses about the origin of under-reaction were put forth for a comprehensive review, see[?]. One hypothesis held that subjects compute Bayes’ rule correctly, but had an inaccurate understanding of the underlying sampling distributions. Formally, subjects can be modeled as reporting

the following biased posterior $\pi(b|d)$:

$$\pi(b|d) = \frac{\pi(d|b)P(b)}{\sum_{b'} \pi(d|b')P(b')} \quad (3.2)$$

where biases in the posterior are driven by biases in the subjective sampling distribution $\pi(d|b)$. To accommodate the existence of under-reaction, subjects would need to assume subjective sampling distributions that were flatter (more dispersed) than the objective distributions. ? proposed that the subjective sampling distribution could be modeled as:

$$\pi(d|b) = \frac{[P(d|b)]}{\sum_{d'} [P(d'|b)]}. \quad (3.3)$$

The parameter ω controls the dispersion of the sampling distribution. When $\omega = 1$, the subjective and objective sampling distributions coincide. Under-reaction occurs when $\omega < 1$.

The biased sampling distribution hypothesis was supported by the observation that subjective sampling distributions were indeed flatter than the objective ones, and substituting these beliefs into Bayes' rule accorded well with reported posterior beliefs^{??}. On the other hand, a critical weakness of this hypothesis is that it cannot explain the existence of under-reaction with a sample size of 1, which would require that subjects disbelieve the experimenter when they are explicitly told the sampling distribution (i.e., the proportion of red chips in the bag). Moreover, even when subjective sampling distributions are entered into Bayes' rule, under-reaction is still sometimes observed e.g.,[?].

These weaknesses of the biased sampling distribution hypothesis motivated the alternative hy-

pothesis that subjects are systematically under-weighting the likelihood², what ? referred to as “conservatism bias.” This hypothesis can be formalized using a generalized version of Bayes’ rule:

$$P(b|d) \propto \frac{[P(d|h)]^\gamma P(b)}{\sum_{b'} [P(d|h')]^\gamma P(b')}, \quad (3.4)$$

where γ is a free parameter specifying the weighting of the likelihood. Note that this model is superficially similar to ?’s formalization of the biased sampling distribution hypothesis, and in fact $\omega = \gamma$ when the denominator of $\pi(d|h)$ ($\sum_{d'} [P(d'|h)]^\omega$) is constant as a function of h (for example, in symmetric problems, where the proportion of red chips in red bags is one minus the proportion of red chips in blue bags). However, the psychological interpretation is different: the biased sampling distribution hypothesis assumes that bias enters at the level of the sampling distribution representation, whereas the conservatism bias hypothesis assumes that bias enters when subjects combine the prior and likelihood. Thus, conservatism bias offers no explanation for why subjective sampling distributions should be biased. It can, however, accommodate the fact that under-reaction occurs for sample sizes of 1, because it posits that even explicit knowledge of the sampling distribution will not prevent biased updating. Likewise, it accommodates the observation that under-reaction is still sometimes observed when subjective sampling distributions are entered into Bayes’ rule.

A third hypothesis, first proposed by ?, is a form of “extreme belief aversion” see also³. If subjects avoid reporting extreme beliefs, then large posterior odds will be pulled towards 0. Consistent with this hypothesis, ? found that subjective odds coincided with the true posterior odds only for posterior odds between -1 and 1 ; outside this range, subjective odds were systematically less extreme than

posterior odds. A weakness of the extreme belief aversion hypothesis, at least in its most basic form, is that it assumes a fixed transformation of the true posterior, which means that it cannot account for experiments in which under-reaction changes across conditions while the true posterior is held fixed e.g.,^{???}.

The literature on under-reaction to evidence faded away without a satisfactory resolution, in part because research was driven towards the study of under-reaction to priors by the work of Kahneman and Tversky^{??}. Instead of using laboratory-controlled scenarios involving bags filled with poker chips, ? invoked more realistic scenarios such as the following:

Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

One group of subjects was told that Jack is one of 100 individuals, 30 of whom are lawyers, and 70 of whom are engineers. Another group of subjects was told that 70 of the individuals were lawyers and 30 were engineers. Kahneman and Tversky found that subjects were largely insensitive to this manipulation: subjects in the first group reported, on average, that the posterior probability of Jack being an engineer was 0.5, and subjects in the second group reported a posterior probability of 0.55. Thus, subjects clearly under-reacted to prior probabilities—i.e., they exhibited *base rate neglect*.⁴

Many subsequent studies have reported under-reaction to priors, though the interpretation of

⁴ Although base rate neglect was popularized by Kahneman and Tversky's work, it was in fact documented earlier using the poker chip paradigm[?], but this observation was mostly ignored by subsequent research using that paradigm.

these studies has been the focus of vigorous debate see²². It has been observed in incentivized experiments e.g.,²³, in real-world markets², and in highly trained specialists such as clinicians² and psychologists².

In addition to establishing the empirical evidence for under-reaction to priors, ² also proposed the most influential account of its psychological origin. They argued that instead of following Bayes' rule, people may use a *representativeness heuristic*, judging the probability of a hypothesis based on the similarity between the observed data and “representative” data under that hypothesis. For example, the vignette describing Jack is intuitively more representative of engineers than it is of lawyers. If people judge the probability of category membership based solely on representativeness, then they will neglect the prior probability of lawyers and engineers in the population, consistent with Kahneman and Tversky's results.

To capture under-reaction to priors formally, the model introduced in Eq. ?? can be generalized to allow insensitivity to the prior²:

$$P(h|d) \propto \frac{[P(d|h)]^{\alpha_L} P(h)^{\alpha_P}}{\sum_{h'} [P(d|h')]^{\alpha_L} P(h')^{\alpha_P}}, \quad (3.5)$$

As before, $\alpha_L < 1$ implies insensitivity to the likelihood; in addition, $\alpha_P < 1$ implies insensitivity to the prior (base rate neglect). ² referred to the case in which $\alpha_L > \alpha_P > 0$ as the *representativeness hypothesis*.

In the special case where $\alpha_L = 1$ and $\alpha_P = 0$, the posterior is simply the normalized likelihood. This corresponds to the model of representativeness judgments proposed by ² in the case where

there are two mutually exclusive hypotheses. This model accounts for why two observations can have the same likelihood but differ in their perceived representativeness. For example, a fair coin is equally likely to generate the sequences HHHH and HTHT (where “H” denotes heads and “T” denotes tails), but people intuitively perceive the latter sequence as more representative of a fair coin. Similarly, people perceive “being divorced 4 times” as more representative of Hollywood actresses than “voting Democratic,” even though the latter has a higher likelihood².

The model put forward by Tenenbaum and Griffiths formalizes the idea that representativeness is tied to *diagnosticity*: the extent to which the data are highly probable under one hypothesis and highly improbable under an alternative hypothesis. ³ offered a different formalization of representativeness that also captures the notion of diagnosticity. They model probability judgments based on consideration of data that are accessible in memory see also³. Judgmental biases arise when an agent engages in “local thinking” (retrieving data from memory based on its diagnosticity). This resonates with modern theories of episodic memory, which posit that the retrievability of information is related to its distinctiveness; under the assumption that information is stored and/or retrieved probabilistically, distinctiveness is directly related to diagnosticity^{3,4}. Consistent with the diagnosticity hypothesis, ⁵ showed greater under-reaction to the evidence when diagnosticity was higher see also^{3,4}. However, a meta-analysis by ⁶ showed that most studies actually find the opposite pattern: under-reaction to the evidence is positively correlated with diagnosticity. One goal of our theoretical account is to resolve this discrepancy.

While much of the work on under-reaction to the prior discussed above was largely driven by findings in more ‘realistic’ scenarios, such effects are also found in more laboratory-controlled paradigms

like those in ? and ?. In particular, when the parameters of the model in Equation ?? are fit to behavioral data from studies using such laboratory-controlled stimuli, the value of α_P is generally between 0 and 1 – indicating that subjects sometimes under-weight the prior in these cases as well, but do not neglect it completely⁷. This formulation therefore allows for the case where both α_P and α_L are less than 1, corresponding to a version of the “system neglect” hypothesis proposed by ?: both the likelihood and prior are neglected, producing an overall insensitivity to variations in the data-generating process. An important implication is that the two forms of under-reaction are compatible (one can under-react to both the likelihood and the prior) and could potentially be explained by a unified model, with similar mechanisms acting across these different domains. A goal of our theoretical account is to understand when under-reaction occurs and when such under-reaction to one source is more prominent than under-reaction to the other.

In summary, the literature on probabilistic belief updating has produced evidence for under-reaction to both prior probabilities and evidence. We now turn to the development of a theoretical account that will explain several aspects and properties of these and other errors.

3.3 LEARNING TO INFER

To understand why people make inferential errors, we need to start by understanding why inference is hard, and what kinds of algorithms people could plausibly use to find approximate solutions. We will therefore begin this section with a general discussion of approximate inference algorithms, identify some limitations of these algorithms (both computationally and cognitively), and then intro-

duce the *learning to infer* framework, which addresses these limitations. This framework provides the basic principles needed to make sense of under-reaction.

3.3.1 APPROXIMATE INFERENCE

The experiments discussed above involved very simple (mostly binary) hypothesis spaces where Bayes' rule is trivial. But in the more realistic domains that humans commonly confront, the hypothesis space can be vast.

For example, consider a clinician diagnosing a patient. A patient can simultaneously have any of N possible conditions. This means that the hypothesis space contains 2^N hypotheses. Or consider the segmentation problem, faced constantly by the visual system, of assigning each retinotopic location to the surface of an object. If there are K objects and N locations, the hypothesis space contains K^N hypotheses. Such vast hypothesis spaces render exact computation of Bayes' rule intractable, because the denominator (the normalizing constant, sometimes called the partition function or marginal likelihood) requires summation over all possible hypotheses.

Virtually all approximate inference algorithms address this problem by circumventing the calculation of the normalizing constant³. For example, Monte Carlo algorithms² approximate the posterior using M weighted samples $\{b^1, \dots, b^M\}$:

$$P(b|d) \approx \sum_{m=1}^M w^m \mathbb{I}[b^m = b], \quad (3.6)$$

where w^m is the weight attached to sample m , and $\mathbb{I}[\cdot] = 1$ if its argument is true (0 otherwise).

Markov chain Monte Carlo algorithms, generate these samples from a Markov chain whose stationary distribution is the posterior, and the weights are uniform, $w^m = 1/M$. The Markov chain is constructed in such a way that the transition distribution does not depend on the normalizing constant. Importance sampling algorithms generate samples simultaneously from a proposal distribution $\tilde{P}(b)$, with weights given by $w^m = P(d|b^m)P(b^m)/\tilde{P}(b^m)$.

Most cognitive theories of approximate inference have appealed to some form of Monte Carlo sampling, for several reasons. First, they can explain response variability in human judgments as arising from randomness in the sampling process^{???}. Second, they can explain a wide range of inferential errors, ranging from subadditivity to the conjunction fallacy^{??}. Third, they can be implemented in biologically plausible circuits with spiking neurons^{???}.

Monte Carlo algorithms can be thought of as procedures for generating an approximate posterior $Q(b|d)$ parametrized by the set of weights and samples, $\phi = \{w^m, b^m\}_{m=1}^M$. The superset Φ of all feasible sets (i.e., the sets that can be produced by a particular Monte Carlo algorithm) defines an approximation family. This leads us to a more general view of approximate inference as an optimization problem: find the approximation (parametrized by $\phi \in \Phi$) that gets “closest” to the true posterior,

$$\phi^* = \underset{\phi \in \Phi}{\operatorname{argmin}} \mathcal{D}[Q(b|d)||P(b|d)], \quad (3.7)$$

where dissimilarity between the two distributions is measured by a divergence functional \mathcal{D} . Most Monte Carlo algorithms do not directly solve this optimization problem, but instead randomly

sample ϕ such that, in the limit $M \rightarrow \infty$, they produce ϕ^* . It is however possible to design non-randomized algorithms that directly optimize ϕ in a sample-based approximation. Such optimization is an example of *variational inference*², because the solution can be derived using the calculus of variations. The most commonly used divergence functional is the Kullback-Leibler (KL) divergence (also known as the relative entropy):

$$\mathcal{D}_{\text{KL}}[Q(h|d)||P(h|d)] = \sum_h Q(h|d) \log \frac{Q(h|d)}{P(h|d)}. \quad (3.8)$$

The variational optimization view of approximate inference allows us to consider more general approximation families that go beyond weighted samples. In fact, the approximate posterior can be any parametrized function that defines a valid probability distribution over the relevant hypothesis space. For example, researchers have used deep neural networks as flexible function approximators^{3,4,5,6}. From a neuroscience perspective, this approach to approximate inference is appealing because it lets us contemplate complex, biologically realistic approximation architectures provided that the optimization procedures can also be realized biologically; see⁷. For example, particular implementations of variational inference have been used to model hierarchical predictive coding in the brain^{8,9}.

3.3.2 AMORTIZATION

Most approximate inference algorithms are *memoryless*: each time the system is queried (i.e., given data and asked to return the probability of a hypothesis or subset of hypotheses), the inference en-

gine is run with a fresh start, oblivious to any computations it carried out before. This has the advantage that the algorithm will be unbiased, and hence with enough computation the parameters can be fine-tuned for the current query. But memorylessness can also be colossally wasteful. Consider a doctor who sees a series of patients. She could in principle recompute her posterior from scratch for each set of observed symptoms. However, this would fail to take advantage of computational overlap across diagnostic queries, which would arise if multiple patients share symptom profiles. To address this problem, computer scientists have developed a variety of *amortized inference* algorithms that reuse computations across multiple queries^{????????}.

To formalize this idea, let the data variable d subsume not only the standard “observation” (e.g., symptoms in the diagnostic example) but also the information provided to the agent about the generative model $P(d, h)$ and the subset of the hypothesis space being queried (e.g., a particular diagnostic test, which is a subset of the joint diagnosis space). In the “classical” approximate inference setting, the inference engine computes a different approximate posterior for each query, with no memory across queries. In the amortized setting, we allow sharing of parameters across queries (Figure ??). Optimizing these parameters induces a form of memory, because changes to the parameter values in response to one query will affect the approximations for other queries. Put simply, the amortized inference engine *learns to infer*: it generalizes from past experience to efficiently compute the approximate posterior conditional on new data.

The optimization problem in the amortized setting is somewhat different from the classical setting. This is because we now have to think about a distribution of queries, $P(q)$. One way to formal-

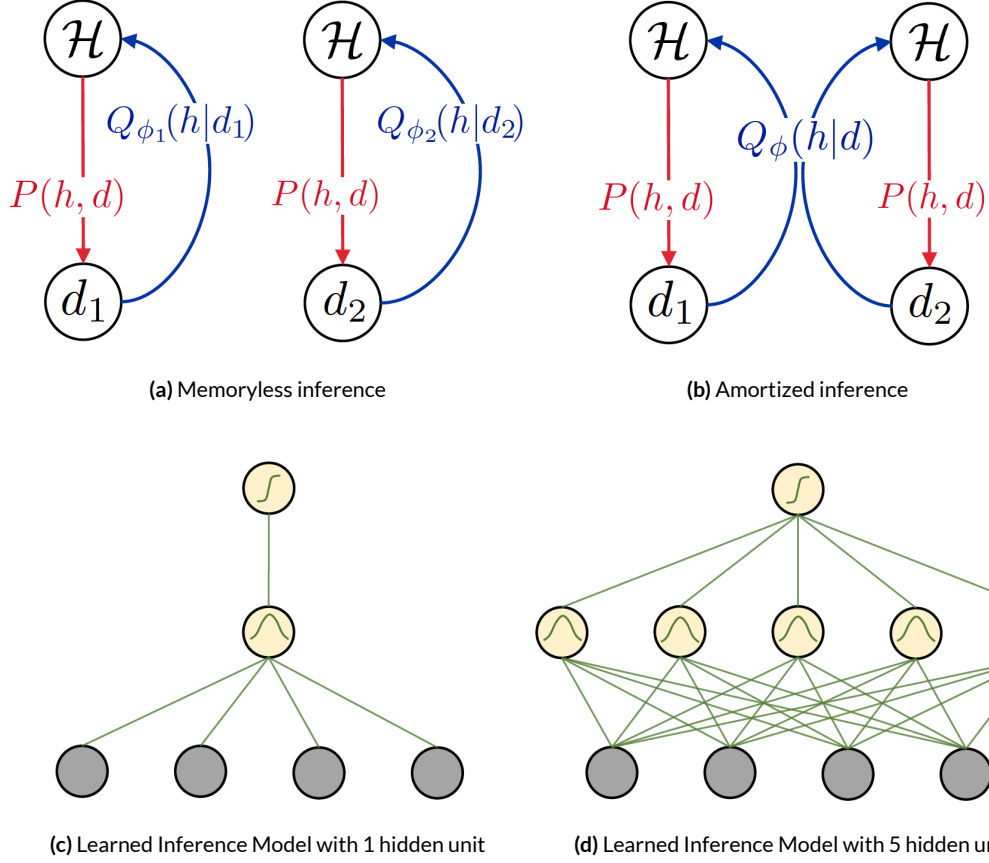


Figure 3.1: Schematics of different inference methods. (A) Memoryless inference recomputes the variational parameters ϕ from scratch for each new set of observations, resulting in an approximate posterior Q that is unique for each d . (B) Amortized inference allows some variational parameters to be shared across queries, optimizing them such that Q is a good approximation *in expectation* over the query distribution. (C) Schematic of how we implemented this framework with a neural network function approximator in the Learned Inference Model, with low capacity (1 hidden unit). (D) Schematic of a neural network function approximator in the Learned Inference Model, with high capacity (5 hidden units).

ize this problem is to define it as an expectation under the query distribution $P_{\text{query}}(d)$:

$$\phi^* = \operatorname{argmin}_{\phi \in \Phi} \mathbb{E}_{P_{\text{query}}} \{ \mathcal{D}[Q(\phi|d) || P(b|d)] \} \quad (3.9)$$

Under this objective function, high probability queries will exert a stronger influence on the variational parameters (see Figure ?? for an illustration). Note that $P_{\text{query}}(d)$ need not be identical to the true marginal probability of the data under the data-generating process, $P(d)$. For example, a child might ask you a series of questions about the reproductive habits of squirrels, but observations of these habits might be rare in your experience.

It is important to note that classical (non-amortized) approximate inference is a special case of amortized inference, and if there are no constraints on the amortization architecture, then the optimal architecture will not do any amortization. This means that amortization only becomes relevant when there are computational constraints that force sharing of variational parameters—i.e., limitations on the function approximator’s capacity. A key part of our argument is that the brain’s inference engine operates under such constraints see^{??}, which will produce the kinds of inferential errors we wish to explain.

3.3.3 THE LEARNED INFERENCE MODEL

We implement a specific version of this general framework, which we refer to as the *Learned Inference Model* (LIM). This model uses a three-layer feedforward neural network as the function approximator (see Figure ?? C-D, further details can be found in Appendix A). The inputs are all the

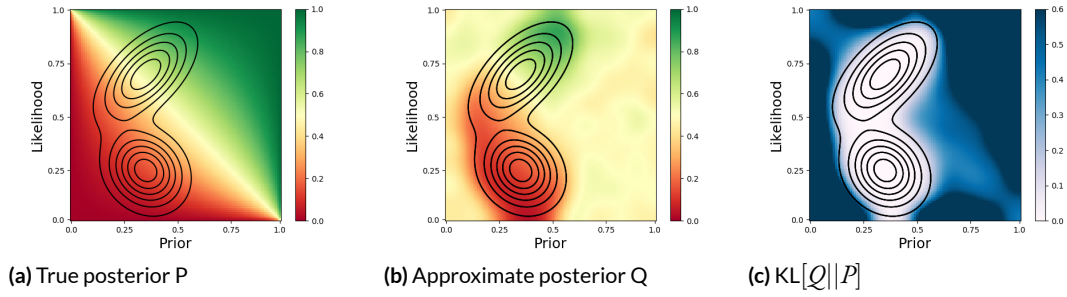


Figure 3.2: Schematic demonstration of how the approximate posterior depends on the query distribution. (A) The true posterior probability P (indicated by colors on the heatmap), as a function of the prior and likelihood for a generative model in which $b \sim \text{Bernoulli}(p_o)$ and $d|b \sim \text{Bernoulli}(p_l)$. The contour lines depict the query distribution. (B) The approximate posterior Q computed by the Learned Inference Model, averaged over the query distribution. The approximation is better for areas that are sufficiently covered by the query distribution. (C) The average KL divergence between the true and approximate posteriors. Higher divergence occurs in areas that are covered less by the query distribution.

relevant information about the query subsumed by the data variable d , and the outputs uniquely determine an approximate distribution $Q(b|d)$ over all hypotheses b . For example, if we want to model the posterior distribution $P(b|d)$ as a Bernoulli distribution over two hypotheses, then the inputs are the prior probabilities of the two hypotheses, the likelihood parameters, and observed data, while the output is a Bernoulli parameter that represents the approximate posterior. The same parameters of the network ϕ are used to generate the approximate distributions $Q(b|d)$ for all queries d (i.e., the approximation is amortized; Figure ?? A-B). The network encounters a series of queries d and outputs a guess for $Q(b|d)$. This guess is improved in response to each new d , with updates to the network parameters ϕ . This leads to query dependence (Figure ??) in the learned parameters ϕ , and therefore in the approximation $Q(b|d)$. The updates to ϕ are made using an algorithm that performs the optimization in Equation ?? using knowledge only of the joint distribution as a learning signal², see Appendix A for details. Since the joint distribution is

known, no external feedback is necessary for learning.

These implementational details were chosen for simplicity and tractability. Because many other choices would produce similar results, we will not make a strong argument in favor of this particular implementation. For our purposes, a neural network is just a learnable function approximator, utilizing the memory of previously sampled experience to approximate future posteriors. Several other memory-based process models for probability judgment for example:^{???} could also learn to infer. Nonetheless, the implementation fulfills several intuitive desiderata for a psychological process model. First, feedforward neural networks have been widely used to model behavioral and neural phenomena. Most relevant to the present approach is the work of ?, who showed how generic neural networks could be trained to implement probabilistic computation. Second, neural networks offer a natural way to specify the computational bottleneck in terms of a convergent pathway (the number of hidden units is smaller than the number of input units)⁵. Such convergence has played an important role in theorizing about other forms of cognitive bottlenecks e.g.,^{??}. Third, the learning rule (blackbox variational inference) can be applied incrementally, and does not require knowledge of the posterior normalizing constant, making it cognitively plausible. Fourth, as we discuss later, the model can be naturally integrated with Monte Carlo sampling accounts of approximate inference.

⁵In this parallel, the network in our LIM is not intended to represent an actual network of neurons in the brain per se, and the convergent bottlenecks induced are not intended as a literal number of neurons in a natural neural network. Real networks in the brain receive information in much higher dimensional format, where the relevant variables are yet to be isolated. Further, they have to cope with noise on these inputs, in the learning signal, and the even the neurons themselves are stochastic. Our model is a highly idealized version of the computations underlying probabilistic judgment, and specifics like the number of units in the bottleneck or the number of layers etc. cannot be directly compared to biologically realistic analogs.

All model parameters (number of hidden units in the bottleneck, the architecture of the network, properties of the optimization algorithm, etc.) are fixed across almost all the experiments (see Appendix A for details); any exceptions are noted where relevant. All the key predictions our model makes are qualitative in nature, and do not require fitting of free parameters to empirical results.

3.4 UNDERSTANDING UNDER-REACTION

We now apply the Learned Inference Model to our motivating question: what is the origin of under-reaction to prior probabilities and evidence? We argue that these inferential errors arise from an amortized posterior approximation. There are two key elements of this explanation. First, the amortized approximation has *limited capacity*: it can only accurately approximate a restricted set of posteriors, due to the fact that the approximation architecture has a computational bottleneck (in our case, a fixed number of units in the hidden layer). We will see how this leads to overall under-reaction to both priors and evidence. Second, the particular posteriors that can be accurately approximated are those that have high probability under the query distribution. We will see how this leads to differential under-reaction to either prior or evidence. In this section, we will focus on the first element (limited capacity), since most of the experiments that we focus on use near-uniform query distributions. We address the second element (dependence on the query distribution) in subsequent sections.

? presented a meta-analysis of studies using the classical balls-in-urns setup, or similar setups (e.g., poker chips in bags). For simplicity, we will use the ball-in-urns setup to describe all of these studies.

Subjects are informed that there are two urns (denoted R and B) filled with some mixture of blue and red balls. On each trial, an urn h is selected based on its prior probability $P(h)$, and then a data set $d = (N_r, N_b)$ of N_r red balls and N_b blue balls is drawn from $P(d|h)$ by sampling $N = N_r + N_b$ balls with replacement from urn h . The subject's task is to judge the posterior probability of urn R, $P(h = R|d)$. Urn R contains mostly red balls (red-dominant), and the urn B contains mostly blue balls (blue-dominant). Following ?, we focus on symmetric problems, where the proportion of the dominant color in both urns is denoted by ϑ , which is always greater than 0.5. We can also interpret ϑ as the *diagnosticity* of the likelihood: when ϑ is large, the urns are easier to tell apart based on a finite sample of balls.

In formalizing a model for subjective performance on this task,² follows³ in allowing separate parameters for sensitivity to the likelihood and the prior (Eq. ??). For analytical convenience, this model can be reformulated as linear in log-odds:

$$\log \frac{P(h = R|d)}{P(h = B|d)} = \alpha_P \log \frac{P(h = R)}{P(h = B)} + \alpha_L \log \frac{P(d|h = R)}{P(d|h = B)} + \epsilon, \quad (3.10)$$

where we have included a random response error term ϵ . This formulation allows us to obtain maximum likelihood estimates $\hat{\alpha}_P$ and $\hat{\alpha}_L$ using least squares linear regression applied to subjective probability judgments (transformed to the log-odds scale).³ first restricted the meta-analyses to studies with equal prior probabilities across the hypotheses, such that α_P is irrelevant. The estimates of α_L revealed three main findings: (i) Under-reaction to the likelihood is more prevalent ($\hat{\alpha}_L < 1$); (ii) the extent of under-reaction to the likelihood is greater ($\hat{\alpha}_L$ is lower) with larger sample size (high N);

and (iii) the extent of under-reaction is greater with higher diagnosticity (higher \mathcal{D}) of the likelihood.

We investigated whether the Learned Inference Model can capture these findings. For each experimental condition, collected from 15 experiments, we trained the model with 2 hidden units on the same stimuli presented to subjects. The conditions varied in likelihood diagnosticity (\mathcal{D}) and sample size (N). We additionally include some uniformly random sample sizes and diagnosticities in training as a proxy for subjects' ability to simulate other possible values for these query parameters, apart from the small set of specific ones chosen by the experimenters⁶. We found that the Learned Inference Model could successfully reproduce the 3 main findings from the ? meta-analysis (Figure ??).

We also applied the model to experiments in which the prior distribution was non-uniform (deviated substantially from 0.5). Figure ?? shows data aggregated by ? along with model simulations, demonstrating that both people and the model tend to be insufficiently sensitive to the prior odds ($\hat{\alpha}_p < 1$), consistent with base rate neglect.

We have shown that several of the main findings in the ? meta-analysis of inferential errors can be reproduced by the Learned Inference Model with limited capacity. We now build an intuition for how the model explains these phenomena. The key idea is that limited capacity forces the model to sacrifice some fidelity to the posterior, producing degeneracy: some inputs map to the same outputs see⁷ for a similar argument. This degeneracy can be seen in Figure ??, where posterior log-odds greater than +5 or less than -5 are mapped to almost the same approximate log-odds value. Degener-

⁶Crucially however, the stimuli actually used in the experiment are much better represented in the query distribution during training – leading to differences in the predictions made by the Learned Inference Models trained on different query distributions from each experiment. The uniformly random inputs primarily serve to add some noise to prevent the Learned Inference Model from overfitting in cases where the experimental stimuli only query a very small number of unique sample sizes and diagnosticities.

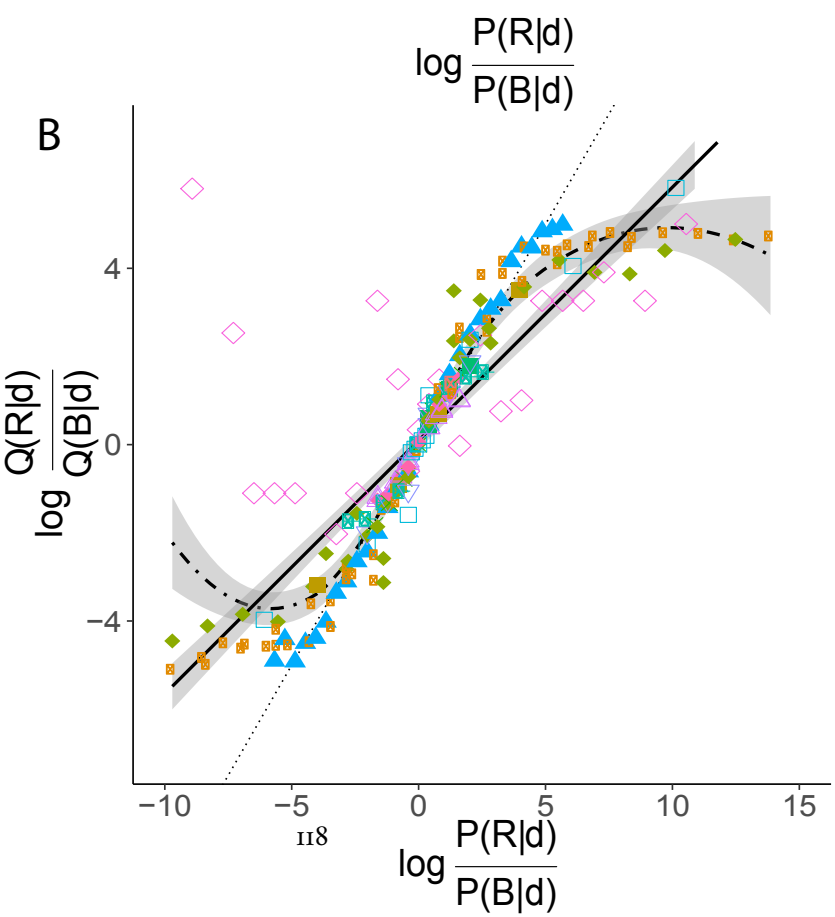
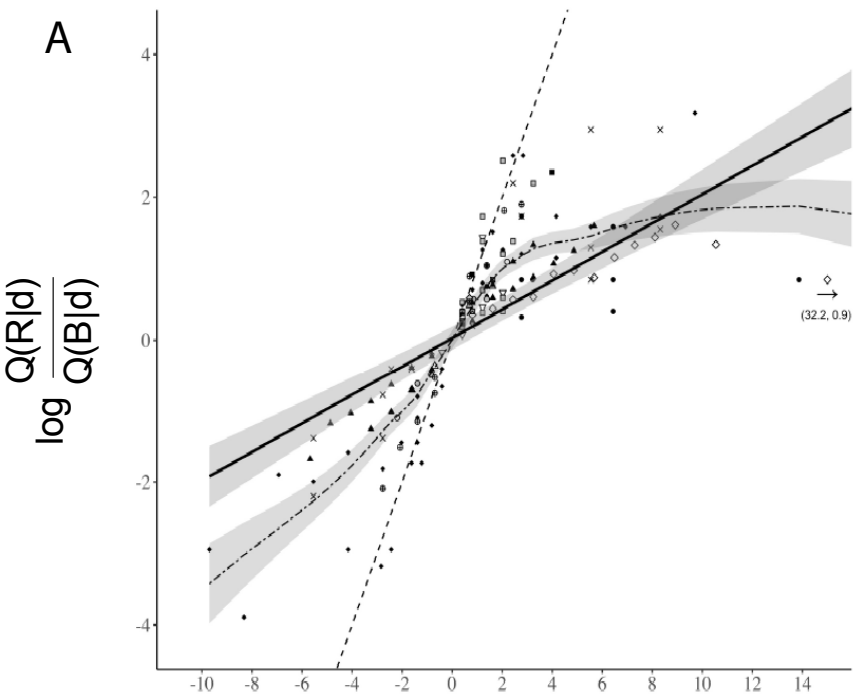
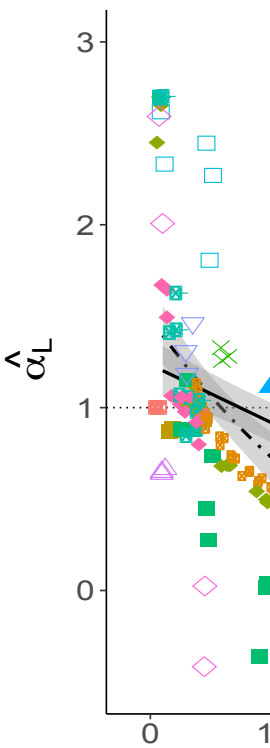
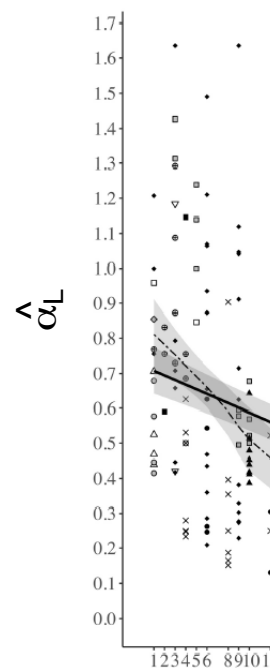


Figure 3.3: Simulation of inferential errors in binary symmetric problems with uniform priors. $P(h|d)$ represents true posterior probabilities, $Q(h|d)$ represents subjective posterior probabilities. (A) Data aggregated by β . (B) Learned Inference Model simulations. Left: subjective posterior log-odds vs. Bayesian posterior log-odds. Middle: estimated sensitivity to the likelihood $\hat{\alpha}_L$ vs sample size N . Right: estimated sensitivity to the likelihood vs. diagnosticity β . The shaded curves show the linear and nonlinear (LOESS) regression functions with 95% confidence bands

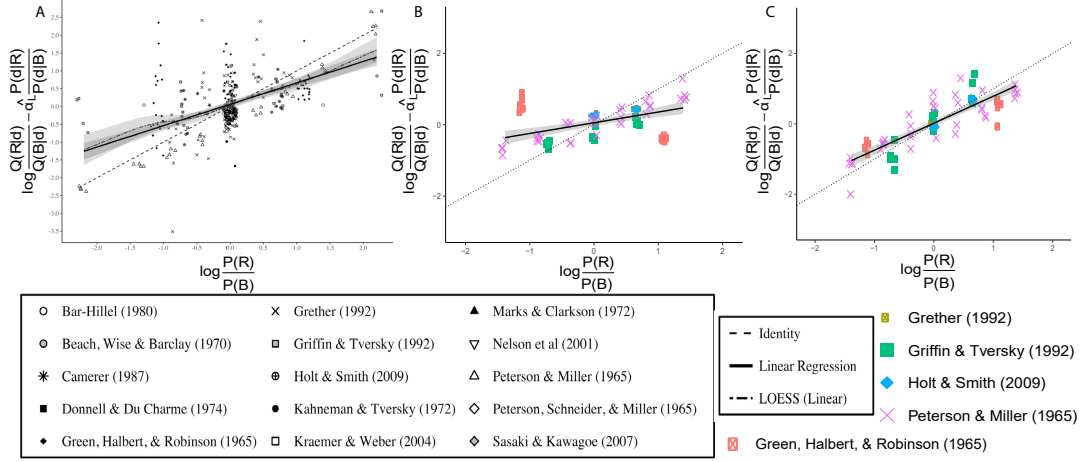


Figure 3.4: Simulation of inferential errors in binary symmetric problems with non-uniform priors. $P(b|d)$ represents true posterior probabilities, $Q(b|d)$ represents subjective posterior probabilities. Plots show prior log-odds on the x axis, and the subjective prior log-odds calculated as the subjective posterior log-odds adjusted for subjective response to the likelihood (as modulated by $\hat{\alpha}_L$). (A) Data aggregated by $\hat{\alpha}_L$. (B) Simulation with low-capacity (2 hidden nodes) Learned Inference Model. (C) Simulation with high-capacity (8 hidden nodes) Learned Inference Model. The shaded curves show the linear and nonlinear (LOESS) regression functions with 95% confidence bands.

acy causes under-reaction overall to sources of information (like sample size, prior and likelihood).

It also causes the approximate posteriors at extreme log-odds to suffer relatively greater deviations from the true posterior, in particular greater under-reaction to sources of information when the log-odds are extreme (e.g., with larger sample sizes and more diagnostic likelihoods). Intuitively, degeneracy causes the model to have a relatively flat response as a function of the posterior log-odds, which means that deviations will also increase with the posterior log-odds.

To demonstrate that these biases in our model are indeed caused by limited capacity in the network, we repeated the same simulations with greater capacity (8 hidden units instead of 2). In this case, we found that the approximate posterior mapped almost exactly to the true posterior (Figure ??A, left). Estimated sensitivity to the likelihood ($\hat{\alpha}_L$) across all diagnosticities and sample sizes was

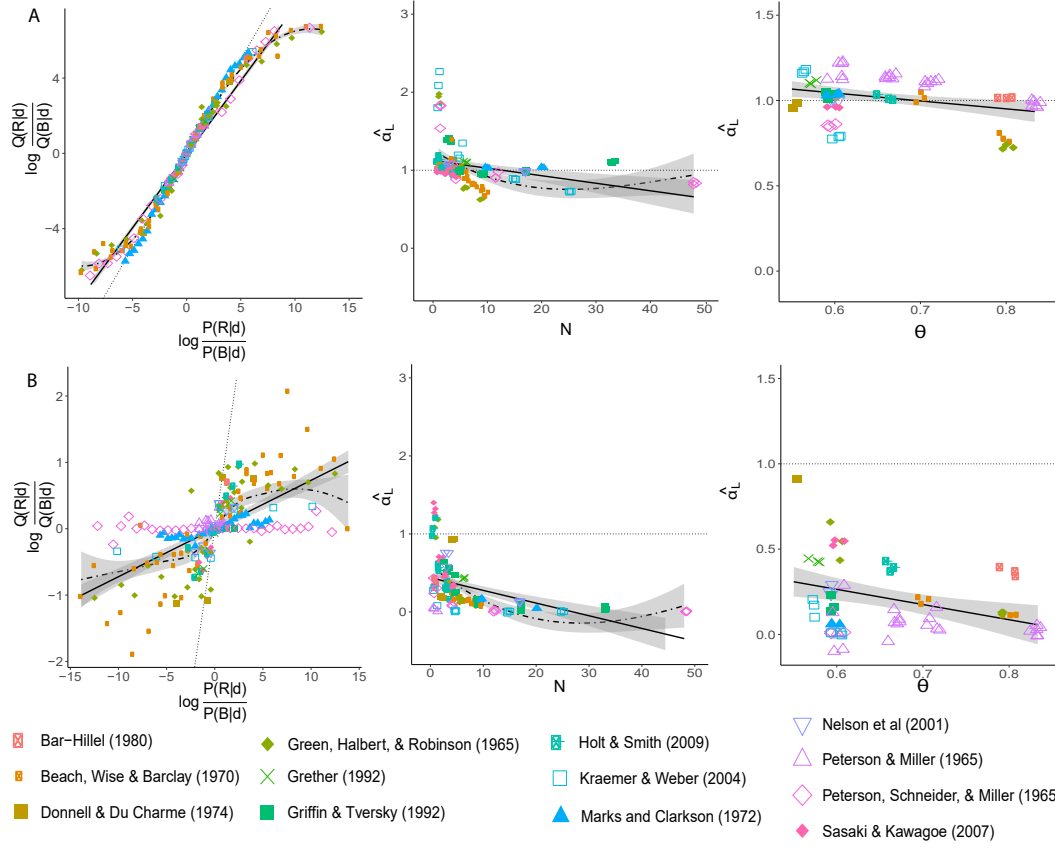


Figure 3.5: Simulations of inferential errors with high capacity and a biased query distribution. $P(b|d)$ represents true posterior probabilities, $Q(b|d)$ represents subjective posterior probabilities. (A) Simulation of high-capacity (8 hidden units) Learned Inference Model. (B) Simulation of low-capacity (2 hidden units) Learned Inference Model with biased query distribution. Left: subjective posterior log-odds vs. Bayesian posterior log-odds. Middle: estimated sensitivity to the likelihood $\hat{\alpha}_L$ vs. sample size N . Right: estimated sensitivity to the likelihood vs. diagnosticity θ . The shaded curves show the linear and nonlinear (LOESS) regression functions with 95% confidence bands.

very close to the Bayesian optimal of 1 (Figure ??A middle and right). We also found that higher capacity mostly abolished base rate neglect (Figure ??C).

What information is lost by a limited capacity approximation depends on the query distribution. To examine this point more closely, we simulated the Learned Inference Model (with 2 hidden units) trained on a biased query distribution, where the likelihood parameters, prior probabilities and sample sizes were the same as used in training previously, but the queries were manipulated such that 90% of the time the data were uninformative about which urn is more likely—i.e., the difference in the number of red and blue balls was close to zero. The query distribution therefore is very peaked around zero likelihood log-odds. We then tested the model on the same queries simulated in Figure ?. As shown in the left panel of Figure ??B (note the change in y axis scale), the approximation is still close to Bayes-optimal near zero posterior log-odds, but the extent of degeneracy is overall far greater, with all the true posterior log-odds being mapped to approximate posterior log-odds roughly between -1 and +1. This results in much greater under-reaction overall. This is also reflected in Figure ??B, middle and right, where the estimated sensitivity $\hat{\alpha}_L$ is closer to zero.

3.4.1 THE EFFECT OF SAMPLE SIZE

In this section, we consider the effect of sample size on the posterior distribution in greater detail, keeping the prior and likelihood parameters fixed. The most systematic investigation of sample size was reported by³, who suggested a specific decomposition of the posterior log-odds into the *strength* (sample proportion) and the *weight* (sample size) of the evidence. These are two sources of information that inform the posterior, and we can consider how strongly participants react to these

Number of heads (h)	Sample size (n)
2	3
3	3
3	5
4	5
5	5
5	9
6	9
7	9
9	17
10	17
11	17
19	33

Table 3.1: Stimuli used in².

the same way we consider their reactions to the prior and evidence in the previous section.

In one of their studies, they gave subjects the following instructions: Imagine that you are spinning a coin, and recording how often the coin lands heads and how often the coin lands tails. Unlike tossing, which (on average) yields an equal number of heads and tails, spinning a coin leads to a bias favoring one side or the other because of slight imperfections on the rim of the coin (and an uneven distribution of mass). Now imagine that you know that this bias is $3/5$. It tends to land on one side 3 out of 5 times. But you do not know if this bias is in favor of heads or in favor of tails. After being shown different sets of coin “spin” results that varied in the number of total spins and the number of observed heads (see Table ??), subjects were then asked to judge the posterior probability that the coin was biased towards heads rather than towards tails.

The two hypotheses in this task were that the biased coin either favors heads (denoted $h = \mathcal{A}$)

or that it favors tails (denoted $h = B$). The prior probabilities of both hypotheses were equal. The symmetric binomial probability was fixed at $\vartheta = 3/5$, and the observed data $d = (N_a, N_b)$ is the number of heads (N_a) and number of tails N_b . The posterior log-odds can then be written as:

$$\log \frac{P(h = A|d)}{P(h = B|d)} = N \left(\frac{N_a - N_b}{N} \right) \log \left(\frac{\vartheta}{1 - \vartheta} \right), \quad (3.11)$$

where $N = N_a + N_b$. Taking the log of this equation results in a linear function relating the log of the posterior log-odds to evidence “strength” $\log \left(\frac{N_a - N_b}{N} \right)$ and “weight” $\log N$. Following ?, ? allowed each component to be weighted by a coefficient (α_W for evidence weight, α_S for evidence strength), absorbed all constants into a fixed intercept term $\alpha_o = \log \log \left(\frac{\vartheta}{1 - \vartheta} \right)$, and allowed for random response error ε , arriving at the following regression model:

$$\log \left(\log \frac{P(h = A|d)}{P(h = B|d)} \right) = \alpha_o + \alpha_W \log(N) + \alpha_S \log \left(\frac{N_a - N_b}{N} \right) + \varepsilon. \quad (3.12)$$

The Bayes-optimal parametrization is $\alpha_W = \alpha_S = 1$. However, ? found that both α_W and α_S significantly smaller than 1. Furthermore, subjects tended to be less sensitive to the weight ($\hat{\alpha}_W = 0.31$) compared to the strength ($\hat{\alpha}_S = 0.81$).

We now turn to predictions from the Learned Inference Model. The actual stimuli presented to subjects in the original experiment were only a small subset of the possible data from the generative model implied by the instructions. Similarly to the previous section, we partially pre-trained the network with random samples from the generative model as follows: we sample the sample sizes from

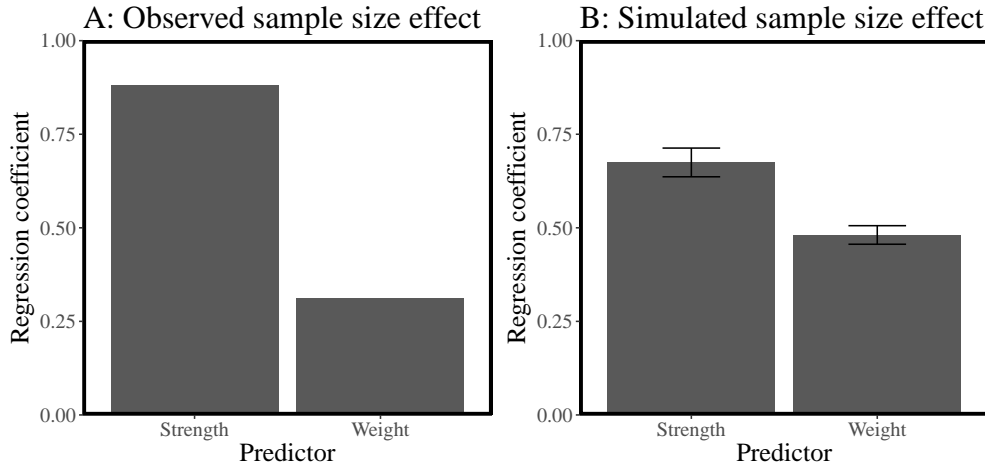


Figure 3.6: Strength and weight in probabilistic judgment. (A) Regression coefficients reported in². (B) Regression coefficients estimated from simulations of the Learned Inference Model. Error bars represent the standard error of the mean.

the set of stimuli used in the original experiment (Table ??), but did not fix the number of observed heads, which we sampled randomly from the generative distribution instead. This can be thought of as offline training on the generative process, which seems plausible based on the instructions given to the subjects, and serves to regularize the Learned Inference Model by preventing overfitting. We then trained exclusively on the specific stimuli used in the original experiment, and carried out our analyses on the model’s response to each query in Table ??.

Consistent with the experimental results, we found that the model was sub-optimally sensitive to both sources of information (Figure ??), with both $\hat{\alpha}_S$ and $\hat{\alpha}_W$ being less than 1. We also found that it was more sensitive to the strength than the weight ($\hat{\alpha}_S = 0.67$, $\hat{\alpha}_W = 0.48$).

Greater sensitivity to strength than to weight in our model can be explained by considering the amount of variance explained by each of these variables. We took random samples from the generative model and measured how much of the variance in the log of the true posterior log odds can

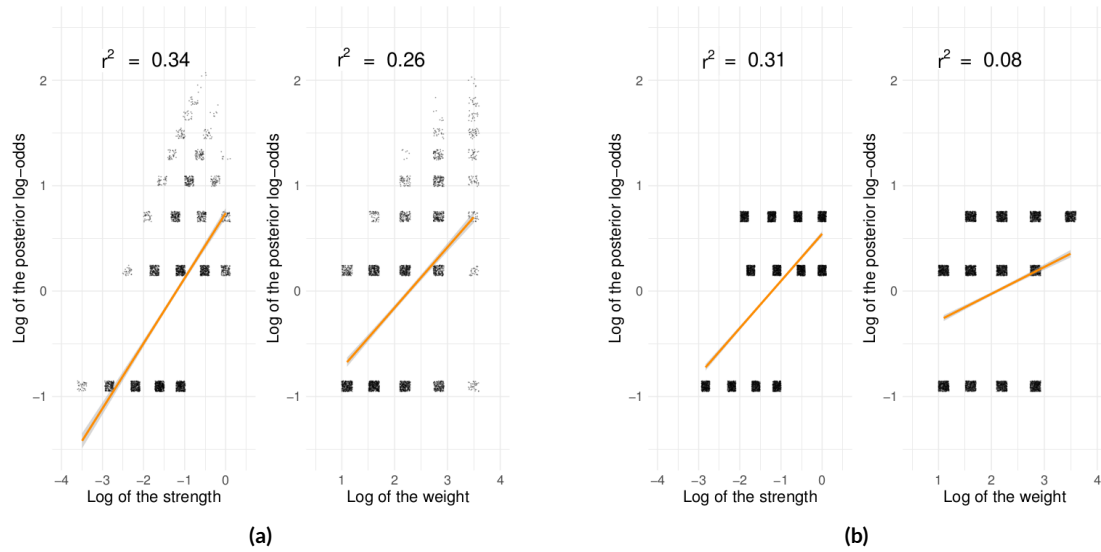


Figure 3.7: Variance explained by strength and weight independently. These plots show regressions between the log of the strength or weight of the evidence against the log of the posterior log-odds. (A) For samples drawn from the true generative process, the strength explains more variance in the posterior. (B) For the stimuli used in ³, the weight explains almost none of the variance in the log posterior log-odds, whereas the strength explains a much higher amount of the variance.

be explained by the log of the strength and the log of the weights separately. We found that the strength variable explains more of the variance in the true posterior than the weight variable (Figure ??A). A resource-limited approximation such as our Learned Inference Model picks up on this difference during pre-training and preferentially attends to the more informative source (i.e., the one that explains more of the variance). Moreover, we carried out these regressions with the specific stimuli used in the the experiment and found that this difference was exaggerated (Figure ??B), with the weight variable explaining very little of the variance in the true posteriors. Training and evaluation on a distribution where the weight explains so little of the variance in the posterior leads the model to react to the weight even less.

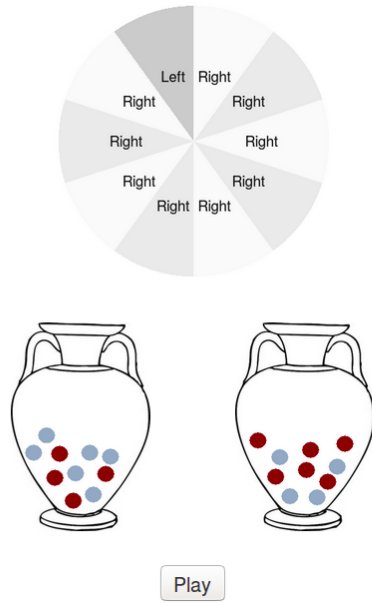
3.4.2 MANIPULATING THE QUERY DISTRIBUTION

In this section, we focus more directly on the role of the query distribution. A basic prediction of our model is that it will put more weight on either the prior or the likelihood, depending on which of the two has been historically more informative about the true posterior. We test this prediction empirically in a new experiment by manipulating the informativeness of the prior and the likelihood during a learning phase, in an effort to elicit over- and under-reaction to data in a subsequent test phase that is fixed across experimental conditions. Specifically, informativeness was manipulated through the diagnosticity of different information sources. In the informative prior/uninformative likelihood condition, the prior probabilities were more diagnostic across queries than the likelihoods, whereas in the uninformative prior/informative likelihood condition, the likelihoods were relatively more diagnostic.

SUBJECTS

We recruited 201 subjects (93 females, mean age=34.17, SD=8.39) on Amazon Mechanical Turk. Subjects were required to have at least 100 past completed studies with a historical completion rate of 99%. The experiment took 12 minutes on average and subjects were paid \$ 2 for their participation. The experiment was approved by the Harvard Institutional Review Board.

A: Informative prior, uninformative likelihood



B: Uninformative prior, informative likelihood

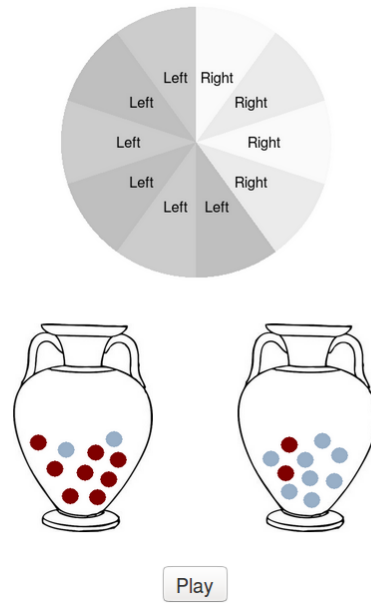


Figure 3.8: Screen shots of urn experiment. (A) In the condition with informative priors and uninformative likelihoods, the wheel of fortune had urn probabilities of 0.7, 0.8, or 0.9. The proportions of blue balls in the urns was 0.5 or 0.6. (B) In the condition with uninformative priors and informative likelihoods, the wheel of fortune had urn probabilities of 0.5 or 0.6. The proportions of blue balls in the urns was 0.7, 0.8, or 0.9.

DESIGN AND PROCEDURE

Subjects were told they would play 10 games with 10 trials each, in which they had to guess from which of two urns a ball was sampled (i.e., which urn was more probable *a posteriori*). On every round, they saw a wheel of fortune and two urns (Figure ??). They were then told that the game was played by another person spinning the wheel of fortune, selecting the resulting urn, and then randomly sampling a ball from the selected urn. The wheel of fortune thus corresponded to the prior and the balls in the urns to the likelihood on each trial. Subjects were told that each trial was independent of all other trials.

Subjects were randomly assigned to one of two between-subjects conditions. One group of subjects went through 8 blocks of 10 trials each with informative priors and uninformative likelihoods (Figure ??A); the other group went through 8 blocks of informative likelihoods and uninformative priors (Figure ??B). We manipulated the prior distribution by changing the number of options on the wheel labeled “left” or “right”. We manipulated the likelihood by changing the proportions of two different colors in both the left and the right urn. Both urns always contained 10 balls of the same colors and the proportion of colors was always exactly mirrored. For example, if the left urn had 8 red balls and 2 blue balls, then the right urn had 2 red and 8 blue balls. For the informative prior/uninformative likelihood condition, the wheel of fortune had urn probabilities (and diagnosticities θ) of 0.7, 0.8, or 0.9, and the proportions of blue balls in the urns was 0.5 or 0.6. For the uninformative prior/informative likelihood condition, the wheel of fortune had urn probabilities of 0.5 or 0.6, and the proportions of blue balls in the urns was 0.7, 0.8, or 0.9.

After the first 8 blocks, both groups of subjects went through the same test blocks. Each test block had either informative priors or informative likelihoods, with their order determined at random. We hypothesized that, if subjects learned to infer the posterior based on their experience during the training blocks, subjects who had experienced informative likelihoods would be more sensitive to the likelihood than subjects who had experienced informative priors, who would be relatively more sensitive to the prior.

BEHAVIORAL RESULTS

We fitted a regression to subjects' responses (transformed to log-odds) during the test blocks following Eq. ???. Thus, we entered the log-odds of the prior, the log-odds of the likelihood, the condition (coded as 'o' for the informative prior condition, and 'l' for the informative likelihood condition), as well as an interaction effect between condition and likelihood and between condition and prior.

As expected, subjects' judgments were influenced by both the prior ($\alpha_P = 0.77, t = 27.529, p < .001$) and the likelihood ($\alpha_L = 0.92, t = 32.68, p < .001$), indicating that they understand the key components of the generative process and therefore recognize and represent both of these as relevant to their final judgment. Crucially, subjects who had previously experienced informative priors reacted more strongly towards the prior than subjects who had experienced informative likelihoods (interaction effect of condition $\times \alpha_P = 0.10, t = 2.44, p = .01$, Figure ??A). Vice versa, subjects who had previously experienced informative likelihoods reacted more strongly towards the likelihoods than subjects who had experienced informative priors (interaction effect of condition $\times \alpha_L = -0.22, t = -5.31, p < .001$, Figure ??B). Furthermore, when estimating individual regressions for both conditions, the reaction to the prior was stronger than the reaction to the likelihood in the informative prior condition ($\hat{\alpha}_P = 0.88$ vs. $\hat{\alpha}_L = 0.70, p < .001$), whereas the reverse was true for the informative likelihood condition ($\hat{\alpha}_P = 0.78$ vs. $\hat{\alpha}_L = 0.92, p < .001$).

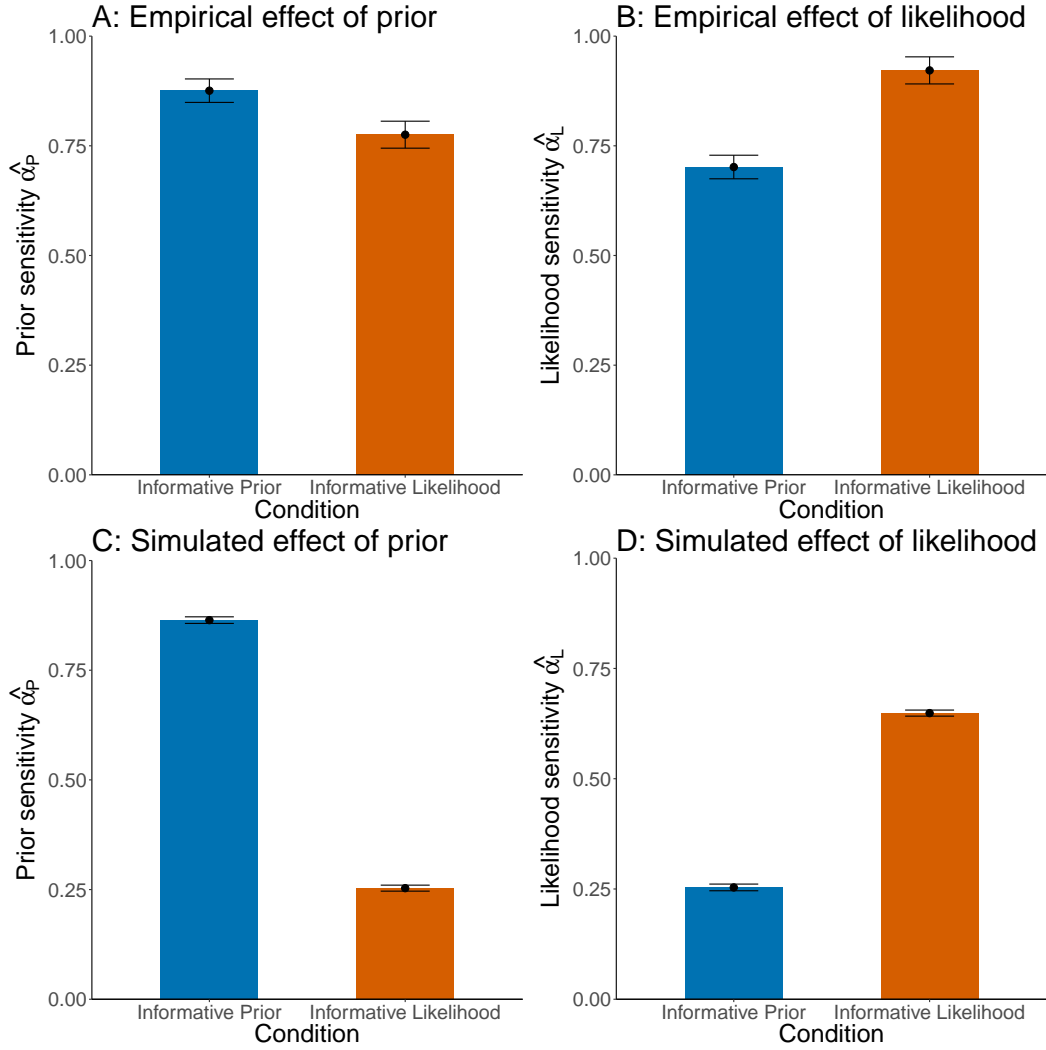


Figure 3.9: Results of urn experiment. The y-axis shows estimates for the regression coefficients α_L and α_P (see Equation ??), and the x-axis represents the experimental condition. (A) Subjects weighted the prior more in the informative prior than in the informative likelihood condition. (B) Subjects weighted the likelihood more in the informative likelihood than in the informative prior condition. (C) The Learned Inference Model weights the prior more in the informative prior condition as compared to in the informative likelihood condition. (D) The Learned Inference Model weights the likelihood more in the informative likelihood condition as compared in the informative prior condition. Error bars represent the standard errors of the regression coefficients.

MODELING RESULTS

We trained the Learned Inference Model to predict the posterior probability for each of the two urns, given the prior probability for each urn and the ratio of colored balls in each of the urns, and the color of the observed ball. We trained 40 different “simulated subjects”, 20 in each condition, each of which observed exactly the data that a subject in their condition had seen, and then tested them on the same test blocks that human subjects went through. We applied the same regression to our Learned Inference Model’s judgments that we applied to subject data. Our Learned Inference Model’s judgments were significantly influenced by both the prior ($\alpha_P = 0.27, t = 41.41, p < .001$) and the likelihood ($\alpha_L = 0.69, t = 104.98, p < .001$). Importantly, the simulated subjects in the informative prior condition reacted more strongly toward the prior (interaction effect condition $\times \alpha_P = 0.60, t = 64.83, p < .001$, Figure ??C), whereas the simulated subjects in the informative likelihood condition reacted more strongly toward the likelihood (interaction effect of condition $\times \alpha_L = -0.41, t = -44.27, p < .001$, Figure ??D). Estimating individual regressions for both conditions as before, the reaction to the prior was higher than the reaction to the likelihood in the informative prior condition ($\hat{\alpha}_P = 0.80$ vs. $\hat{\alpha}_L = 0.21, p < .001$), whereas the reverse was true for the informative likelihood condition ($\hat{\alpha}_P = 0.29$ vs. $\hat{\alpha}_L = 0.71, p < .001$). Our Learned Inference Model therefore reproduces the behavioral findings observed in our experiment.

3.4.3 MANIPULATING THE QUERY DISTRIBUTION BETWEEN VS. WITHIN SUBJECTS

The study reported in the previous section demonstrates that the weight of an information source (prior or likelihood) is correlated with its diagnosticity. An additional implication of the Learned Inference Model is that people will only be sensitive to the prior and likelihood if these parameters vary across queries during training of the recognition model. If the parameters are relatively constant (even if very diagnostic), then the recognition model will learn to “ignore” them. More precisely, the recognition model learns to amortize a fixed belief about the priors when they are held constant, and therefore will be relatively insensitive to surprising changes in the prior. This implication is relevant to a line of argument articulated by ?, that base rates are only ignored when they are manipulated between rather than within subjects.

Several lines of evidence support Koehler’s argument. ? found greater sensitivity to base rates using a within-subject design, and similar results have been reported by ? and ?, though see ? for evidence that base rate neglect occurs even using within-subject designs. ? pointed out an asymmetry in the experiments of ?, where individuating information was manipulated within subject and base rates were manipulated between subjects. He suggested that this may have focused subjects’ attention on individuating information at the expense of base rates. Using a full between-subjects design, ? found greater sensitivity to base rates, consistent with a reduction in the relative salience of individuating information compared to the mixed within/between-subject design.

For concreteness, we will consider this issue in the context of the well-known taxi cab problem, where subjects were asked to answer the following question:

Two cab companies, the Blue and the Green, operate in a given city. Eighty-five percent of the cabs in the city are Blue; the remaining fifteen percent are Green. A cab was involved in a hit-and-run accident at night. A witness identified the cab as a Green cab. The court tested the witness' ability to distinguish a Blue cab from a Green cab at night by presenting to him film sequences, half of which depicted Blue cabs, and half depicting Green cabs. He was able to make correct identification in 8 out of 10 tries. He made one error on each color of cab. What do you think is the probability (expressed as a percentage) that the cab involved in this accident was Green?

Note that the prior in this case is fairly diagnostic: it strongly favors Blue cabs. However, several studies of the taxi-cab and similar problems produced evidence for base rate neglect⁷. These studies manipulated the base rates in a between-subject design. In the taxi cab problem, this corresponds to telling one group of subjects that 85% of the cabs are Blue and telling another that 85% are Green. Therefore, while the prior information is diagnostic, as it appears to each subject, it never varies.

As mentioned above, β found greater base rate sensitivity using a within-subject manipulation of base rates in the taxi cab problem. Each subject was given two different base rates for the cab problem. We simulate the condition in which the base rates were either 85% or 15%. The Learned Inference Model reproduces the key finding of greater sensitivity to base rates using a within-subject design (Figure ??). In fact, the model exhibits total neglect of base rates in the between-subjects design, consistent with previous findings reported by⁷, though not all experiments show such extreme results.⁷ The Learned Inference Model naturally explains the difference between experimental designs

⁷The assumption that these are the only queries ever seen by participants would result in no covariance between prior and posterior in the between-subject design (since the prior never varies). This would give total base-rate neglect. This is an extreme assumption we make for illustrative purposes. More realistically, observ-

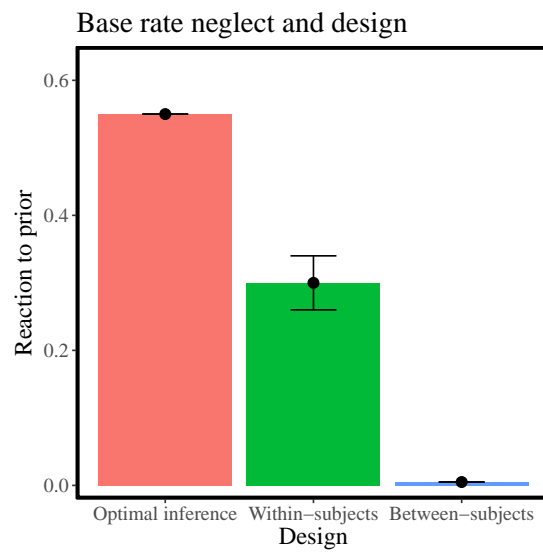


Figure 3.10: Base rate neglect within and between subjects. The y-axis shows the reaction to the prior as measured in predictions from the Learned Inference Model, the x-axis shows the different conditions. Reaction to the prior here is measured by the difference between the responses given to test queries in which the base rate was 85% and those in which the base rate was 15%. Thus, a greater difference indicates a stronger reaction to prior information. The model simulations of the within-subjects design show a stronger reaction to the base rates than the simulations of the between-subjects design (which shows no reaction to the base rate at all). Both of these conditions produce under-reaction to the base rate compared to the Bayes-optimal judgment.

as a consequence of the fact that limited capacity and biased query distributions cause the model to ignore sources of information that do not reliably covary with the posterior.

The differences in historical query distributions for each subject as determined by the experimental design also sheds light on discrepancies in the effects of diagnosticity on the extent of under-reaction. Studies that find that reactions to a source of information are stronger with increasing relative diagnosticity^{7, 22} of that source of information, used between-subject designs. This is analogous to our study in which subjects “attend” more to a source that was more informative in the experienced query distribution, leading to a stronger reaction to that source in future queries. However, studies reported in⁷ find greater under-reaction with increasing diagnosticity (Figure ??). We note that these studies predominantly used within-subject designs,⁸ in which the same subject has to make inferences across all levels of diagnosticity. This leads to a much broader query distribution, where no source has reliably higher diagnosticity. Imposing a limitation on the capacity of the approximation results in an inability to faithfully express this broad query distribution, and some neglect of the specific parameters⁷. This produces degeneracies in the response that manifest as greater under-reaction to more diagnostic sources of information. Our model therefore is able to replicate these seemingly contradictory findings, by taking into account the experienced query distribution of each subject.

ing these queries simply concentrates the query distribution in this space and reduces covariance between the prior and the posterior.

⁸Two exceptions to this pattern are⁷ and⁷.

3.4.4 EXTENSION TO A CONTINUOUS DOMAIN

In this section, we investigate the effect of informativeness in a continuous domain, re-analyzing a data set reported by ?. Subjects ($N = 117$) were recruited through Amazon Mechanical Turk to take part in an experiment in which they had to predict the pay-off of different slot machines. In total, they were shown 10 different slot machines and had to make 10 guesses per slot machine. Pay-offs varied between 0 and 100 and were noisy such that no slot machine gave the same pay-off every time. Subjects were assigned randomly to one of two groups in a between-subjects design. Each machine k was associated with a Gaussian distribution $\mathcal{N}(m_k, s)$ over outputs y_{kn} on each trial n . The variance s was fixed to 25 and the mean was drawn from a normal distribution $\mathcal{N}(m_o, v)$, with m_o set to 40 and the global variance v manipulated between groups. One group, in the low dispersion condition, experienced a global variance of $v = 36$. The other group, in the high dispersion condition, experienced a global variance of $v = 144$.

? used this paradigm to show how manipulating the dispersion produced faster or slower acquisition of abstract knowledge; we focus on a different aspect of the data here: subjects updating behavior. Figure ??A shows subjects' reaction to the incoming data, quantified as how much they update their predictions after observing a slot machine's output, plotted against the predicted update of a rational hierarchical model inferring the posterior mean payoffs for a machine.⁹ Subjects' updates are positively correlated with the model's predicted updates for both the high dispersion ($r(99) = 0.57$,

⁹This rational hierarchical model is assumed to know the true parameter values for s , v and m_o . However, in this experiment, these parameters for the full data-generating process were not explicitly shown to participants. We therefore also carry out an analysis using a hierarchical Bayesian model that additionally also infers these parameter values. This leads to similar results; see Appendix B for details.

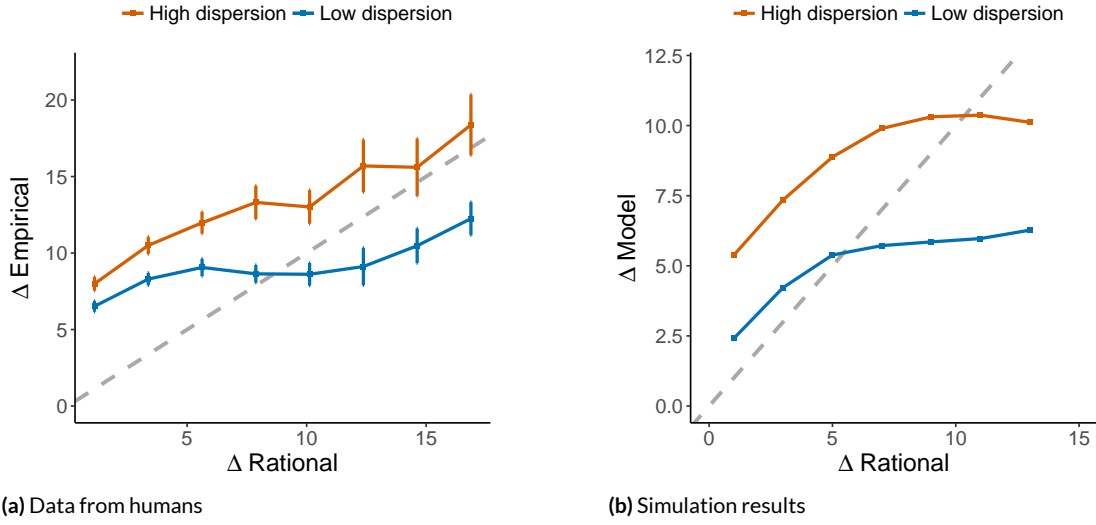


Figure 3.11: Inferential errors in a continuous domain. (A) Reanalysis of data from the payoff prediction task collected by ². (B) Simulations of the Learned Inference Model. Each panel shows subjective updates from prior to posterior (Δ Data) on the y-axis and the update of a rational (hierarchical) model (Δ Rational) on the x-axis. Error bars represent the standard error of the mean. Gray lines represent $y = x$.

$p < .001$) and the low dispersion condition ($r(116) = 0.36, p < .001$). This is expected as the hierarchical model is assumed to be a good first approximation of human behavior in this task. However, subjects updated their beliefs much more in the high dispersion than in the low dispersion condition – even for the same rational update ($t(214) = 9.24, p < .001$, after accounting for differences in rational updates between the conditions). This means that they were affected more strongly by the *same* incoming evidence in the high dispersion than in the low dispersion condition. As the higher dispersion group experienced a higher global variance, this also means that they experienced a less informative prior. Thus, the fact that they under-reacted to the prior when it is relatively less informative reproduces the effect observed in our urn experiment in a continuous domain.

To simulate these findings, we parametrized the outputs of the Learned Inference Model to re-

turn the mean and log standard deviation of a Gaussian posterior. The function approximator was a neural network with a single two-unit hidden layer and a tanh non-linearity, taking as input the last observation, the mean of the observations seen so far in that episode and the number of observations in that episode. We trained the model on the same generative process as was applied in the behavioral study. We then use the model’s predicted mean as the response on every trial.

The results, shown in Figure ??B, demonstrate that the model qualitatively matches the human data: a positive correlation between the hierarchical model’s predictions and our Learned Inference Model’s responses for both the low dispersion ($r(19) = 0.82, p < .001$) and the high dispersion condition ($r(19) = 0.82, p < .001$), but critically the update was stronger for the high dispersion condition than for the low dispersion condition ($t(38) = 7.40, p < .001$).

A discrepancy in the behavior of our model and the human data can be seen for large updates, where the model predictions flatten out significantly compared to human data. This is due to the degeneracy caused by limited capacity (see also figures ?? and ??). Different architectures and ways to parametrize the approximate distribution Q would lead to different kinds of degeneracies and might better model this aspect of the human data. Nonetheless, the effect we are primarily interested in in this study is that the updates in the high dispersion condition are greater than in the low dispersion condition (for both our model and the human data), for every value of the true Bayesian update. This validates our claim that reaction to data depends on the relative informativeness of the prior and the likelihood in past queries. This claim applies to both discrete and continuous domains.

3.5 FURTHER EVIDENCE FOR AMORTIZATION: BELIEF BIAS AND MEMORY EFFECTS

We now shift from our analysis of under-reaction to a broader evaluation of the Learned Inference Model, focusing on two predictions. First, the model predicts that the accuracy of human probabilistic judgment will depend not only on the “syntax” of the inference engine (how accurately the inference engine manipulates probabilistic information) but also on the “semantics” (how well the probabilistic information corresponds to prior experience and knowledge). The semantic dependence gives rise to a form of *belief bias*, in which people are more accurate when asked to make judgments about “believable” probabilistic information compared to “unbelievable” information, even when the syntactic demands (i.e., Bayes’ rule) are equated. Second, the model predicts that there will be *memory effects* (sequential dependencies): one probabilistic judgment may influence a subsequent judgment even when the two queries are different.

3.5.1 BELIEF BIAS

In studies of deductive reasoning, people appear to be influenced by their prior beliefs in ways that sometimes conflict with logical validity. Specifically, they tend to endorse arguments whose conclusions are believable, and reject arguments whose conclusions are unbelievable, regardless of the arguments’ logical validity e.g.,^{???}. This belief bias phenomenon has played a pivotal role in adjudicating between theories of logical reasoning.

Belief bias has also been observed in probabilistic reasoning tasks^{??}. Here we focus on the study reported by ?, which varied whether the posterior probabilities dictated by Bayes’ rule were close

to independently measured intuitive estimates of the corresponding real-world probabilities. Subjects were asked to perform Bayesian reasoning in real-world situations (e.g., medical diagnosis), with prior and likelihood information that was either consistent with (believable condition), or inconsistent with (unbelievable condition) observed real-world values. The authors found that subjects' responses correlated well with Bayesian posterior probabilities in the believable condition (Figure ??A), and were much less correlated in the unbelievable condition (Figure ??B).

An intuitive interpretation for these results is that people anchor to the experienced real-world values of the prior, likelihood, and resulting posterior, and adjust their computations inadequately to the parameters actually presented in the query. The final responses are therefore closer to the true posterior when this anchor is close to the experimental parameters presented, as in the believable condition. Anchoring has previously been modeled as the outcome of a resource-limited sampling algorithms^{??}, but has usually been studied in cases where the anchor is explicitly provided in the experimental prompt. Learned inference strategies account for memory of previous queries, and provide a model for what such an anchor for a new query could be, in the form of an *a priori* guess based on relevant past judgment experience. This interpretation of learned inference as augmenting or anchoring other run-time approximate inference strategies is discussed in greater detail in the section on Amortization as Regularization.

We model these effects by training the Learned Inference Model on a set of priors and likelihoods that result in a particular posterior distribution, P_A , and testing on a set of priors and likelihoods that result in posterior probabilities that either have the same distribution P_A (believable condition)

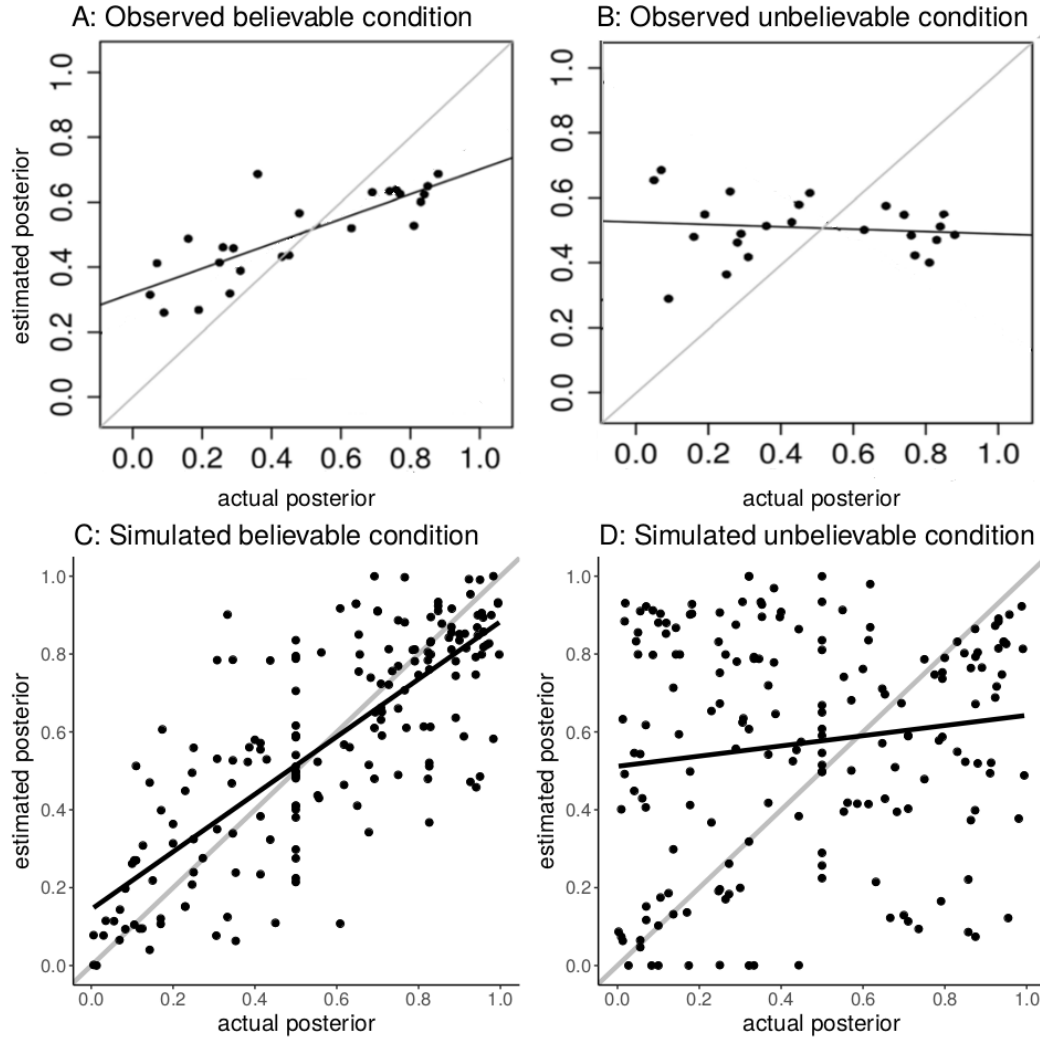


Figure 3.12: Belief bias. Top: experimental data. Bottom: simulations of the Learned Inference Model. (A) Empirical results for the believable condition³. (B) Empirical results for the unbelievable condition. (C) Simulated results for the believable condition. (D) Simulated results for the unbelievable condition. The correlation between the actual and estimated posterior is closer to 1 (i.e., exact Bayesian inference) in the believable condition than in the unbelievable condition. The Learned Inference Model reproduces this effect.

or a different distribution P_B (unbelievable condition)¹⁰. The model produces responses that are highly correlated with the true posterior probability in the believable condition (Figure ??C, $r = 0.78, p < .001$), but this correlation is much lower in the unbelievable condition (Figure ??D, $r = 0.14, p = .06$, comparative test: $z = 2.64, p = .004$). Our model therefore reproduces the belief bias effect reported by³.

3.5.2 MEMORY EFFECTS

In our own previous work³, we observed signatures of amortized inference in subjects' probability estimates. One such signature was that their answers to a question (Q_2) were predictably biased by their answers to a previous question (Q_1). This bias was stronger in cases where the two queries were more similar.

The experiments were carried out in the domain of scene statistics. We asked people to predict the probability of the presence of a “query object”, given the presence of a “cue object” in a scene. The query object was kept the same across both queries. In one condition, the cue object in Q_1 was “similar” to the one in Q_2 , measured by the KL divergence between the two posteriors over objects conditional on the cue object. In the other condition, the cue object in Q_2 was dissimilar from the one in Q_1 .

For example:

¹⁰Each simulated subject received a training distribution where the posterior probabilities were distributed according to the mixture distribution $P_A = 0.5 \times \text{Beta}(3, 1) + 0.5 \times \text{Beta}(1, 1)$. Simulated subjects in the believable condition were tested on posteriors sampled from the same distribution, those in the unbelievable condition were tested on posteriors sampled from the mixture distribution $P_B = 0.5 \times \text{Beta}(1, 3) + 0.5 \times \text{Beta}(1, 1)$. An equal number of simulated subjects received P_B as the training distribution (with P_A as the test distribution in the unbelievable condition).

Q₁: “Given the presence of a chair in a photo, what is the probability of there also being a painting, plant, printer, or any other object starting with a P in that photo?”

Q₂ (Similar): “Given the presence of a book in a photo, what is the probability there is any object starting with a P in the photo?”

Q₂ (Dissimilar): “Given the presence of a road in a photo, what is the probability there is any object starting with a P in the photo?”

appendix We biased the responses to Q₁ for half the subjects using an unpacking manipulation, which produces subadditivity of probability judgments. A subadditivity effect occurs when the perceived probability of a hypothesis is higher when the hypothesis is ‘unpacked’ into a disjunction of multiple typical sub-hypotheses^{???}. Using an example from our own work, when subjects were told that there was a “chair” in the scene, they tended to assign higher probability to the ‘unpacked’ hypothesis “painting, plant, printer, or any other object starting with a P”, than a control group who was asked about the ‘packed’ hypothesis “any object starting with a P”. The true posterior is the same across these different conditions. Critically, we found that the subadditivity group assigned higher probability to the hypothesis queried in Q₂ than the control group, holding fixed Q₂ across groups. This means that the bias induced by Q₁ was detectable in Q₂, indicating that some computations involved in answering Q₁ were re-used to answer Q₂. Importantly, we found that this bias was only detectable if the cue objects across Q₁ and Q₂ were similar (Figure ??A). For example, first being asked Q₁ about the probability of the set of “objects starting with a P” in the presence of a “book”, and afterwards being asked Q₂ about the probability of “objects starting with a P” in the presence of a “chair” produced a memory effect whereas asking the same Q₂ did not show this mem-

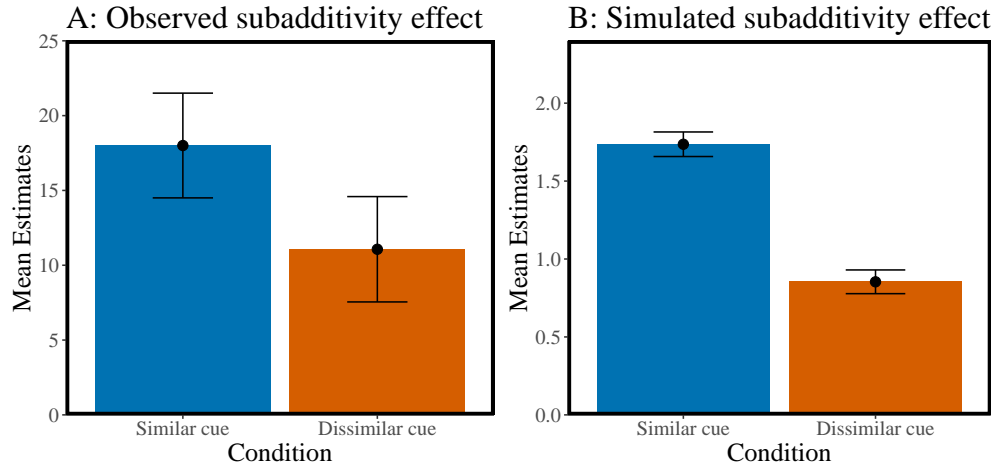


Figure 3.13: Memory effect. (A) Observed subadditivity effect in query 2 reported in[?]. Cues that were similar to a previous query showed a higher effect than cues that were less similar, indicating strategic reuse of past computation. (B) Simulated subadditivity effect. Provided that the model was trained to exhibit a subadditivity effect in a first query, this effect remained stronger for similar queries than for dissimilar queries. Error bars represent the standard error of the mean.

ory effect when subjects in Q_1 were asked about the probability of the set of “objects starting with a P” in the presence of a “road”. We argued that this was a sign of intelligent reuse of computation, since a chair is more likely to co-occur in scenes with a book than in scenes with a road.

In[?], we modeled reuse using amortizations of samples in a Monte Carlo framework. However, a basic problem facing this framework is that the Monte Carlo sampler cannot “know” about similarity (measured in terms of KL divergence) without knowing the true posterior, which of course is the entity it is trying to approximate. The Learned Inference Model provides an answer to this conundrum, by adaptively amortizing (i.e., reusing) past computations without access to the KL divergence or other omniscient similarity measures.

In the interest of simplicity, we simulate these effects in a smaller version of the original environment, rather than using the full-scale scene statistics as in our original study[?]. We simulated a data

set of scene statistics with 12 objects with 2 different “topics” that drive the multinomial probability distributions over these 12 objects². Using this setup, one can derive the joint probability of any 2 objects. The joint probability is all that is required for blackbox variational inference², so we are able to train a larger version of the Learned Inference Model (with 1 hidden layer, 10 hidden units and a radial basis function non-linearity), which takes as input each object d (a 12-dimensional one-hot vector) and outputs the 12-dimensional multinomial probability distribution $P(h|d)$ over all objects.

We then manipulated $P(h_i, d)$ for a specific cue object d and query object h_i by biasing it to be higher than its true value (analogous to the subadditivity manipulation) and trained the Learned Inference Model with the biased joint distribution for a few steps. This caused the model to partially amortize Q1, which in turn influenced its answer to Q2 (a memory-based subadditivity effect), since the same network was used to answer both. Our simulations demonstrate that the subadditive effect is significantly larger for similar compared to dissimilar cue objects (Figure ??B; $t(58) = 4.62$, $p < .001$). Our model therefore reproduces the difference in the memory effect reported by ?. Note however that the simulations are carried out in a different generative model (i.e., a simplified version of the empirical environment), and the sizes of the effects are not directly comparable.

Note that we did not attempt in this section to more directly model subadditivity, as this would require the introduction of additional mechanisms into our framework. Prior work by ? suggests how Monte Carlo sampling naturally explains subadditivity. As we address further in the General Discussion, there are a number of ways that the Monte Carlo and amortized variational inference frameworks could be integrated.

3.6 AMORTIZATION AS REGULARIZATION

We introduced amortization as a method for optimizing a function that maps queries to posterior distributions. Another view of amortization is as a method for regularizing an estimator of the posterior distribution for a single query. The intuition behind this is that one might have gained over experience some knowledge of what the relevant task parameters and the resulting posteriors generally are, and use that to regularize a noisy estimator for the posterior for a new query at run-time. At first glance, it may seem odd to think about the variational optimization procedure as producing an estimator in the statistical sense, since the posterior is a deterministic function of the query. To explain why this is in fact not odd, we need to lay some groundwork.

An inference engine that is not bound by time, space or computational constraints will reliably output the true posterior distribution, whereas a constrained inference engine will output an approximate posterior. There is no way for the constrained inference engine to know exactly how close its approximation is to the true posterior. Another way of saying this is that the constrained inference engine has *epistemic uncertainty*, even if the engine itself is completely deterministic and hence lacks any *aleatory uncertainty* (i.e., uncertainty arising from randomness).¹¹ We can thus regard the approximate posterior as an estimator of the true posterior, and ask how we might improve it through the use of inductive biases: if we have some prior knowledge about which posteriors are more likely than others, we can use this knowledge to bias the estimator and thereby offset the effects of computational imprecision.

¹¹Epistemic uncertainty due to computational imprecision has been studied systematically in the field of *probabilistic numerics*².

To formalize this idea in the context of amortized inference, the optimization problem in Eq. ?? can be rewritten (up to an irrelevant constant factor) as follows:

$$\phi^* = \operatorname{argmin}_{\phi \in \Phi} \left[\mathcal{D}[\mathcal{Q}(b|d) || P(b|d)] + \frac{1}{P_{\text{query}}(d)} \mathbb{E}_{P_{\text{query}}} \{ \mathcal{D}[\mathcal{Q}(b|d') || P(b|d')] | d' \neq d \} \right]. \quad (3.13)$$

This expression separates a “focal” query d (the one you are trying to answer now) from the distribution of other queries ($d' \neq d$). If the focal query is high probability, the second term counts less, and in the limit disappears, such that the optimization problem reduces to fitting the variational parameters to the focal query. When the focal query is low probability, the second term exerts a stronger influence, and in the limit the optimization problem completely ignores the focal query. We can think of the second term as a regularizer: it pulls the variational parameters towards values that work well (minimize divergence) under the query distribution, and this pull is stronger when the focal query is low probability.

The regularization perspective allows us to connect our framework to the “correction prior” theory developed by ?. According to Zhu and colleagues, the brain approximates the posterior by generating stochastic hypothesis samples, and then “corrects” this approximation by regularizing it towards a meta-Bayesian prior over posteriors see also[?]. The theoretical motivation for correction is that the posterior approximation is a random variable due to the stochastic sampling process; when only a few samples are drawn cf.[?] this produces a noisy estimate of the posterior that may deviate significantly from the true posterior. The correction procedure reduces variance in the posterior estimate by increasing bias, pulling it towards the meta-Bayesian prior over posteriors (intuitively,

towards an ‘a priori’ guess based on past experience), and therefore partially compensates for the error in the sampling process.

More formally, the stochastic hypothesis sampling procedure corresponds to a form of Monte Carlo approximation (see Eq. ??). In the simple binary setting, $\mathcal{H} = \{0, 1\}$, the Monte Carlo inference engine generates \mathcal{M} samples from $P(b|d)$ ¹². In our generic formalism, the approximate posterior is parametrized by the proportion of “successes” $\phi = K/\mathcal{M}$, where $K = \sum_m \mathbb{I}[b^m = 1]$. The approximate posterior is then given by $Q(b|d) = \phi^b(1 - \phi)^{1-b}$. This approximation will exhibit large stochastic deviations from the true posterior for small \mathcal{M} .¹³

To reduce the variance of the Monte Carlo estimator, ? proposed a meta-Bayesian inference procedure that computes the posterior over the optimal parameters ϕ^* given the “data” supplied by the random variable ϕ :

$$P(\phi^*|\phi) \propto P(\phi|\phi^*)P(\phi^*). \quad (3.14)$$

When the prior $P(\phi^*)$ is a Beta(A,B) distribution, the posterior mean estimator is given by:

$$\mathbb{E}\{\phi^*|\phi\} = w\phi + (1 - w)\frac{A}{A + B}, \quad (3.15)$$

¹²We assume for simplicity that the inference engine can directly sample from the posterior, though in most cases of practical interest the inference engine will sample from a proxy distribution. For example, in Markov chain Monte Carlo schemes, the inference engine samples from a Markov chain whose stationary distribution is the posterior²².

¹³The variance of the Monte Carlo estimator for a binomial distribution with success probability p is $p(1 - p)/\mathcal{M}$.

where $w = \frac{1}{M+A+B}$ controls the balance between the Monte Carlo estimate ϕ and the prior mean $\frac{A}{A+B}$, which acts as a regularizer. Intuitively, a larger sample size (M) or weaker prior ($A + B$) shift the balance from the prior to the Monte Carlo estimate. When $A = B$, as assumed in §, the prior mean is $1/2$. This gives rise to a form of “conservatism” in which probabilities greater than $1/2$ are underestimated, and probabilities less than $1/2$ are overestimated[?]. We remind the reader that this form of conservatism is distinct from the under-reaction that we modeled in previous sections, which is sometimes referred to as conservative probability updating[?].

§ found evidence for such a “conservative” prior using two different data sets. The first one was data collected by §, who asked subjects to estimate probabilities for a range of weather events (e.g., cold, windy or sunny), or to estimate probabilities of future events (e.g., “Germany is in the finals of the next World Cup.”). The second one was data collected by §, who assessed the variability of probability estimates for different phrases such as “improbably” or “quite likely”. The sampling and correction prior model was able to quantitatively capture the observed conservatism effect: people weighted their probability estimates towards 0.5 when providing their judgments (Figure ??A). It also led to a novel prediction that the variance of probability estimates should be a quadratic function of the true probability, with a peak at $1/2$ (see Footnote ??). This prediction was confirmed in the experimental data (Figure ??B).

We now show that we can capture the same behavioral phenomena (mean and variance effects) using the Learned Inference Model. This analysis provides an important insight: the random nature of the approximate posterior is not necessary (as in the correction prior framework), and that regu-

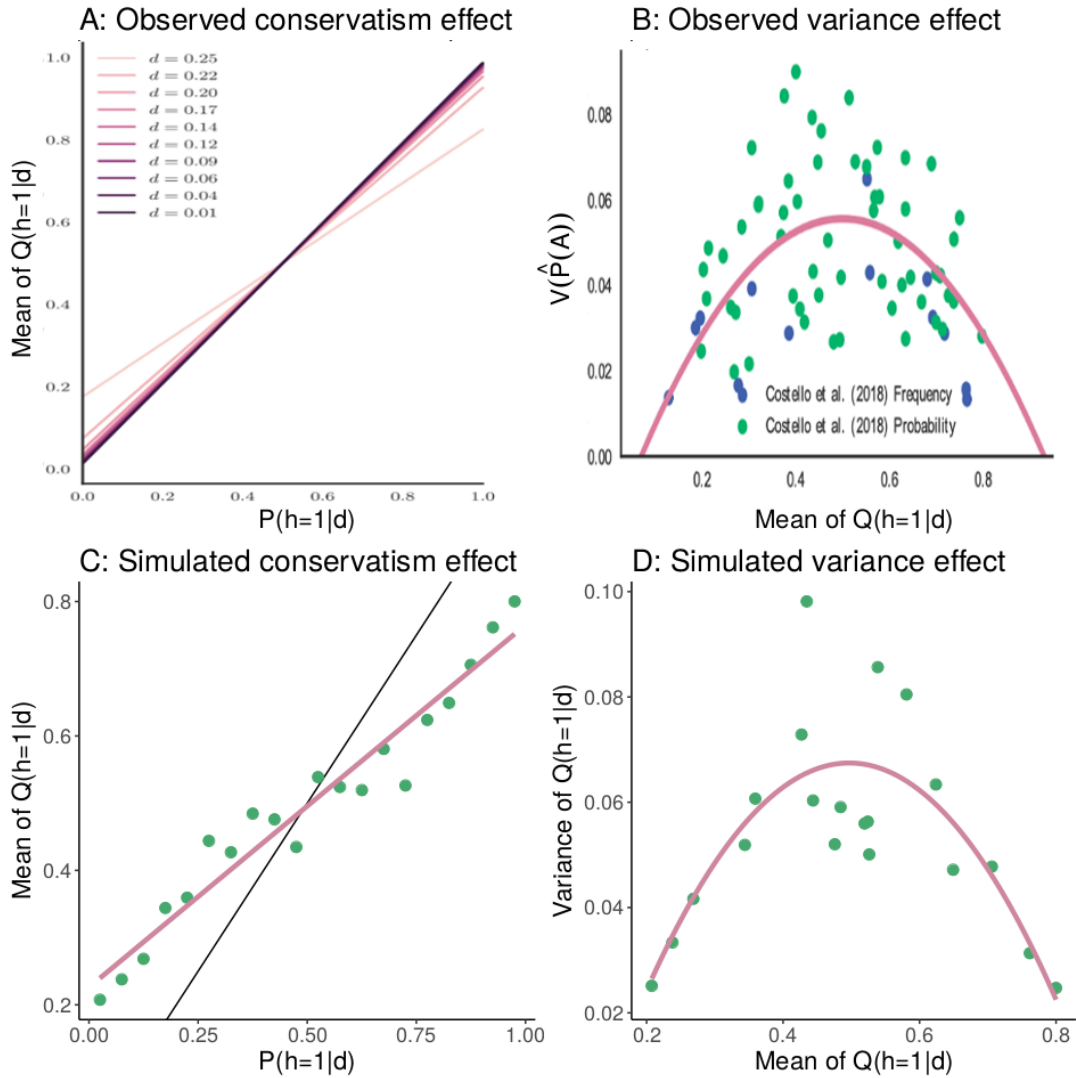


Figure 3.14: Correction prior. (A) Simulation results from the correction prior model in ³ exhibiting conservatism. Black line represents the optimal response and the colored lines show estimates from different parameterizations of the model. (B) Quadratic relation between the variance of subjective probability estimates and mean subjective probability estimates, as observed by ³. Points show data points from previous empirical studies. The line shows best fit quadratic fit to this data. (C) The Learned Inference Model replicates the conservatism effect. Points represent mean estimates from our model, the pink line represents the best fit linear regression to these points, the black line represents the optimal response. (D) The Learned Inference Model replicates the variance effect. Points represent variance of the subjective responses from our model for different mean subjective responses. The pink line represents the best fit quadratic fit to these points.

larization, which can act even on deterministic approximations (provided these approximations are capacity limited as in the Learned Inference Model), can explain the observed effects.

To simulate the experimental data, we created a query distribution that would give rise to posteriors distributed according to $\text{Beta}(0.27, 0.27)$, which Zhu and colleagues obtained by fitting their correction prior to data on probability judgments collected by ?. We then trained the Learned Inference Model on queries sampled from this distribution. When tested on a range of queries, the trained model replicated the conservatism effect in Figure ??C. Regressing the expected probabilities onto the models' responses revealed an estimated slope of 0.54, which was significantly smaller than 1 (Wald test: $t = 14.49, p < .001$). This arises from regularization towards the mean response of 0.5.² explained the quadratic relationship between the expectation and the variance as a feature of the sampling approximation. However, our results demonstrate that the effect can arise even when the approximation is deterministic, as long as it is capacity limited. The key observation is that the Learned Inference model contains degeneracies in the mapping from true to approximate posterior and these degeneracies are more apparent further from the mean. This increase in degeneracy results in lower variance at extreme probabilities. This can also be interpreted as a bias-variance trade-off² – the increased bias towards the mean response (conservatism) at extreme probabilities causes the variance of the estimator to decrease.

Regressing the models predictions onto the simulated variance, we find that a quadratic model performs better than an intercept-only model as also reported by², $F(1, 19) = 78.73, p < .001$ (Figure ??D). Solving the resulting quadratic regression for its maximum showed that this function peaked at 0.498 (i.e., close to 0.5 as predicted by the correction prior). We conclude that our Learned

Inference Model can reproduce the conservatism and quadratic variance effects reported by², but without a stochastic sampling algorithm. In the General Discussion, we return to the relationship between learning to infer and stochastic sampling.

3.7 GENERAL DISCUSSION

Although many studies suggest that the human brain is remarkably adept at carrying out Bayesian inference e.g.,^{2,22,23}, many other studies present evidence for systematic departures from Bayesian inference e.g.,^{24,25,26}. What does this mean for theories of probabilistic reasoning? Should we abandon Bayesian inference as a descriptive model? Are people using Bayesian inference in some situations and heuristics in others? These questions motivated our effort to formulate a new theory—*learning to infer*.

The starting point of our new theory is the assumption that the brain must efficiently use its limited computational resources²⁷. This assumption means that Bayes-optimality is *not* the appropriate normative standard for probabilistic reasoning. Rather, we must consider how accuracy of probabilistic reasoning trades off against the computational cost of accuracy. A learning system that is trained to approximate probabilistic inference will, when a limit on the computational cost is imposed (modeled here as a computational bottleneck), exploit regularities in the distribution of queries. These regularities allow the system to efficiently use its limited resources, but it will also produce systematic errors when answering queries that are low probability under the query distribution. We showed that these are precisely the errors made by people.

We implemented a specific version of this theory (the *Learned Inference Model*) using a neural network function approximator, where the computational bottleneck corresponds to the number of nodes in the hidden layer. Our choice of neural network function approximator was motivated by a natural complementarity between the strengths of probabilistic generative models and neural networks. Neural networks are best thought of as pattern recognition and function approximation tools, rather than as ways to represent causal knowledge about the world². In contrast, probabilistic generative models are good ways to represent knowledge about causal structure, and define what problem we are trying to solve in inferring hidden causes from data, but they do not specify good effective inference algorithms. By using neural networks to learn to infer in a probabilistic generative model, a cognitive agent can combine the strengths of these two approaches. Neural networks are used not to recognize patterns in the external world, but patterns in the agent’s own internal computations: what kinds of observed data typically indicate that a particular inference is appropriate?

The model reproduced the results of several classical and recent experiments in which people under-react to probabilistic information. We first observed patterns in under-reaction predicted by limited capacity. We then found that the model can reproduce sample size effects, in particular different reactions to the strength and weight of evidence, by more strongly reacting to sources of information that have historically been more diagnostic of the posterior. This led to the new predictions that under-reaction to the evidence should occur when the queried posteriors covary more strongly with the prior than with the likelihood (causing the function approximator to “attend” to more to the prior), whereas under-reaction to the prior should occur when the queries covary more with the likelihood than the prior. We tested this prediction in a new experiment that varied the structure of

the query distribution, confirming that people make different inferential errors depending on the query distribution, even when all probabilistic information is provided to them. We also applied the analysis of under-reaction to several other experimental factors, such as sample size, between- vs. within-subject designs, and continuous hypothesis spaces.

The Learned Inference Model also provided insights into a range of other inferential errors. For example, we showed how it could explain belief bias in probabilistic reasoning, the finding that people are closer to the Bayesian norm when given probabilities that are consistent with their real-world knowledge[?]. Belief bias arises, according to the model, because the function approximator has to make predictions about the posterior in a region of the query space that it was not trained on. Another example is the finding of sequential effects in probabilistic reasoning: a single query can bias a subsequent query, if the two posterior distributions are sufficiently similar[?]. This arises, according to the model, because learning in response to the first query alters the function approximator's parameters, thereby biasing the output for the second query.

Finally, we showed how the Learned Inference Model offers a new realization for the correction prior proposed by[?], according to which inferences are regularized towards frequently occurring posterior probabilities. Taken together, these results enrich our understanding of how people perform approximate inferences in computationally challenging tasks, which can be accomplished by learning a mapping between the observed data and the posterior. Our proposed Learned Inference Model is a powerful model of human inference that puts learning and memory at the core of probabilistic reasoning.

3.7.1 RELATED WORK

Egon Brunswik famously urged psychologists to focus on the structure of natural environments, and the corresponding structure of features that the mind relies on to perform inferences². Herbert Simon proposed the metaphor of the mind’s computations and the environment’s structure fitting together like the blades of a pair of scissors, such that psychologists would have to look at both blades to understand how the scissors cut³. This interdependence between people’s strategies and their environments has been stressed by psychologists for decades⁴, and our proposed Learned Inference Model fits well into that tradition. Essentially, what we have argued for here is that subjects do not rely on a stable and fully rational engine for probabilistic inference, but rather that they learn to infer—i.e., they optimize a computationally bounded approximate inference engine, using memory to learn from previous relevant experience. Our proposal emphasizes the importance of studying an agent’s environment, in particular the query distribution they are exposed to. For example, whereas subjects who experienced informative priors in our urn experiment ended up showing conservatism, subjects who experienced informative likelihoods showed base rate neglect. Our proposal also stresses the importance of both memory (people re-use past computations) and structure learning (people learn a mapping between observable and the posterior) to explain subjects’ probabilistic reasoning more generally.

The idea that memory plays an important role in inference has been studied by a number of authors. For example, ? developed a theory of hypothesis generation based on memory mechanisms see⁵ for an overview of this research program. Related ideas have also been explored in behavioral

economics to explain decision making anomalies². Our contribution has been to formalize these ideas within a computational rationality framework², demonstrating how a resource-limited system could adaptively acquire inferential expertise, which would in turn produce predictable inferential errors.

Ours is not the first proposal to apply a neural network-based approach to explain how people reason about probabilities.² used an adaptive network model of associative learning to model how people learned to categorize hypothetical patients with particular symptom patterns as having specific diseases. Their results showed that when one disease was far more likely than another, the network model predicted base rate neglect, which they confirmed in subjects across 3 different experiments. This is similar to our prediction that the Learned Inference Model will start ignoring the prior if it has been historically less informative, for example because one disease has never appeared during learning. Using a similar paradigm,² showed that some versions of base rate neglect can be accounted for by a simple connectionist model. Both of these studies, however, provided subjects with direct category feedback, whereas our Learned Inference Model only requires access to the joint probabilities, making it more algorithmically plausible.² showed how vector space semantic models were able to predict a number of biases in human judgments, including a form of base rate neglect based on typical and non-typical descriptions of people and judgments about their occupations.

That the prior and the likelihood can be differentially weighted based on their importance has been proposed before. For example,² argued that neither the base rate nor the likelihood are ever fully ignored, but may be integrated into the final judgment differently, such that whether they are predictive of the eventual outcome would influence the weight people place on them. The idea that

people ignore aspects of probability descriptions if they are not informative is a pivotal part of ecological definitions of rationality, for example as part of the priority heuristic². In one exemplary demonstration of how ignoring unpredictable information can be beneficial,² simulated environments in which base rates changed more frequently than cue accuracies, and found that models ignoring either the base rate or the likelihood could perform as well as their fully Bayesian counterparts.

3.7.2 INTEGRATING WITH SAMPLING-BASED APPROACHES

Our theory relies heavily on a variational framework for thinking about the optimization problem that is being solved by the brain's approximate inference engine. This creates some dissonance with prevailing ideas about approximate inference in cognitive science, most of which have been grounded in a hypothesis sampling (Monte Carlo) framework see³ for a review, with small numbers of samples. Hypothesis sampling has also been studied independently in neuroscience as a biologically plausible mechanism for approximate inference e.g.,^{3,4}. In our own prior theoretical work, we have employed hypothesis sampling to explain a range of inferential errors^{3,5}. The question then arises of how (if at all) we can reconcile these two perspectives – one of a variational approximation learned over several past experiences, versus the other of a Monte Carlo approximation consisting of a handful of samples in response to the current query. We discussed in broader terms the potential role of a learned inference model in augmenting predictions from a noisy sampler as part of our section on 'Amortization as Regularization'. Here we sketch a few more concrete possibilities for how these approaches might be combined to build new, testable models of human probabilistic

inference.

Almost all practical Monte Carlo methods rely on a proxy distribution for generating samples. Markov chain Monte Carlo methods construct a Markov chain whose stationary distribution is the true posterior, often making use of a proposal distribution to generate samples that are accepted or rejected. Importance sampling methods simultaneously draw a set of samples from a proposal distribution and reweight them. Particle filtering methods apply the same idea to the case where data are observed sequentially. One natural way to combine variational inference with these methods is to use the variational approximation as a proposal distribution. This idea has been developed in the machine learning literature e.g.,^{??}, but has not been applied to human judgment.

For Markov chain Monte Carlo methods, another possibility would be for the variational approximation to supply the initialization of the chain. If enough samples are generated, the initialization should not matter, but a number of cognitive phenomena are consistent with the idea that only a small number of samples are generated, thereby producing sensitivity to the initialization. For example, probability judgments are influenced by different ways of unpacking the sub-hypotheses of a disjunctive query[?] or providing incidental information that serves as an “anchor”^{??}. In these studies, the anchor is usually provided as an explicit prompt in the experiment – learned inference strategies provide a model for what such an anchor for a new query could be in the absence of an explicit prompt, in the form of an ‘a priori’ guess based on past judgment experience.

Several recent methods in the machine literature combine the complementary advantages of sampling approximations and variational approximations leading to several new algorithms^{???} that could also be studied as models for human judgment.

The blackbox variational inference algorithm that we use (see Appendix A) does in fact involve sampling: the gradient of the evidence lower bound is approximated using a set of samples from the variational approximation. Although we are not aware of direct evidence for such an algorithm in brain or behavior, the idea that hypothesis sampling is involved in the learning process is an intriguing possibility that has begun to be studied more systematically^{???}. It resonates with work in other domains like reinforcement learning, where people seem to engage in offline simulation to drive value updating^{???}.

3.7.3 CONNECTIONS TO OTHER MODELS FOR JUDGMENT ERRORS

In addition to the sampling-based approaches that we discuss in the previous subsection, there may also be other sources of probabilistic judgment errors in humans. Some of these include misinterpretation or misunderstanding of the question being posed by the experimenter[?], inability to map the provided probabilities onto an intuitive causal model[?], or simply disbelief in the experimenter's description of the data-generating process.¹⁴ We have restricted most of our attention to studies in which subjects had to reason about data-generating processes that are explicitly described (e.g., how many balls of each color were present in an urn). Considerable evidence suggests that people's judgments and decisions differ depending on whether they have received a problem as a description or have experienced probabilities through experience^{??}. These are all likely part of the explanation

¹⁴While these models predict deviations from optimality, they do not always specify a model for the responses actually produced, when participants do not understand, internalize, or believe the data-generating process presented by the experimenter. One possibility is that they fall back upon 'a priori' notions of the data-generating process. Our Learned Inference Model provides a model for what these context-sensitive 'a priori' beliefs might be – in particular how these could be learned from past judgment experience.

for the judgment errors discussed in this paper. Below, we suggest a few ways in which predictions from our model could be integrated with, or distinguished from, predictions driven by these other mechanisms.

The Learned Inference Model in its current formulation assumes that the correct data-generating process is provided in the query, and only learns how to do inference within this data-generating process. It does not account for uncertainty about or disbelief in the data-generating process itself, and is insensitive to whether information about it is acquired through description or learned from previous experience. One could manipulate the amount of experience participants have with the data-generating process by letting them observe samples from it within the experiment, rather than only providing them with a description of the probabilities. This would manipulate the certainty participants have in the data-generating process, and pave the way towards assessing its influence on probability judgments in these domains – independent of the effects predicted by a Learned Inference Model which assumes perfect knowledge of the data-generating process.

Domain knowledge and pre-experimental experience can also contribute to uncertainty about the presented data-generating process. Most of our results are from highly controlled domains (i.e., balls in urns), that people likely do not have strong intuitions for based on past experience. Our findings in these domains are modeled with inference strategies learned within the experiment. Considerable evidence shows that people's judgments and decisions are influenced by whether the data-generating process presented matches pre-experimental intuitions about the causal structure of the real world^{??}. The Learned Inference Model in its current formulation has no notion of real-world causal structure, and therefore no intuition about it. It can learn inference strategies from

within-experiment experience in any data-generating process irrespective of whether it respects such intuitions. Expanding our results to naturalistic settings, where people might have ‘a priori’ causal intuitions from previous experience, would allow us to manipulate how ‘intuitive’ the presented data-generating process is and tease apart its role in judgment errors from the predictions of the Learned Inference Model.

Finally, we discussed in the previous section how learned inference strategies might be integrated with memoryless sampling-based approaches that approximate responses at each query independently with a small number of samples. We discussed this as a bias-variance trade-off in our section on ‘Amortization as regularization’. A prediction of this framework is that the extent of such regularization will depend on the amount of experience accrued in that domain, with more experience favoring a learned inference strategy over memoryless stochastic sampling. Empirical results suggest that experts and novices employ different decision strategies, with experts appearing to rely more on memory-based heuristics^{???}. Studying judgment errors across domains where participants vary in pre-experimental experience, or even over the course of an experiment as within-experiment experience increases, would allow us to better understand how learned and memoryless inference strategies interact and trade-off.

More broadly, our theory of learning to infer allows us to frame many of these errors in the context of resource-rationality^{??}, and explains how biases observed in the lab could be inevitable consequences of algorithms that let resource-bounded minds solve hard problems in real time. Many of the alternative mechanisms for judgment errors suggested above have also been interpreted this way^{???}. Our model uniquely addresses how such biases could derive rationally from limited capac-

ity inference strategies learned from the history of past judgment experience. We leave many questions open for further investigation, for example: how the mechanisms of learning to infer interact with other approximate inference strategies; which of these phenomena are best explained by our approach as opposed to others, and under what circumstances; and how previously proposed accounts in part might also be consequences of learned inference strategies.

3.7.4 LIMITATIONS AND FUTURE DIRECTIONS

We modeled the mapping between queries and the posterior using a multilayer neural network. This model does not assume any explicit representational structure; the mapping is optimized using blackbox variational inference, and many different mappings can be learned depending on the capacity of the neural network. While this model provides a good first-order approximation of what the brain might be doing, it remains to be seen whether the functional form we chose is the best relative to other possibilities. For example, our recent work on function learning suggests that people have a strong inductive bias for compositional functions—i.e., functions that can be built up out of simpler building blocks through algebraic operations⁷.

Another limitation of our work is that we focused on cases where the posterior is defined over a single random variable, but in the real world people frequently need to make inferences about subsets of variables (or functions of those subsets) drawn from very large sets of variables with complex joint distributions. This complexity was the motivation for our previous work on hypothesis sampling, which offers a computationally tractable solution to this problem⁷. The memory-based subadditivity effects that we modeled⁷ are an example of a phenomenon in which amortized inference

and hypothesis sampling might be unified, but we have not provided a comprehensive unification (though the previous section describes some potential avenues). For example, although our model can capture the fact that more similar query items can lead to higher subadditivity effects than less similar items, it currently does not explain how subadditivity arises to start with.

In our model, the inputs are already boiled down to only the relevant variables and therefore very low-dimensional, and the cost function only evaluates how well the network predicts posterior probabilities from these inputs. Inputs in the real world, however, are likely more noisy and high dimensional. Several related but different tasks are often multiplexed into the same network representations in the brain^{??}. Extending our theory to more noisy and uncertain real-world learning is an important and interesting challenge.

We have assumed that the computational bottleneck is fixed, defining a limited representational capacity for the function approximator that must be shared (possibly unequally) across queries. However, in particular when considering computational capacity as a cost, another possibility is that the bottleneck is flexible: representational capacity might increase (e.g., through the allocation of additional units) when greater accuracy becomes worth the cost of this greater investment, possibly by commandeering resources from other cognitive systems. This predicts that more accurate probabilistic judgment should be associated with poorer performance on other concurrent tasks that share cognitive resources, and that properly incentivizing people should improve their performance. Contrary to this hypothesis, evidence suggests that incentives have little to no effect on some inferential errors, such as base rate neglect^{???}, and this point is corroborated by evidence that inferential errors also appear in real markets with highly incentivized traders[?].

3.7.5 CONCLUSION

In his paper criticizing past research on base rate neglect,² argued that “adding up studies in which base rate neglect appears or disappears will lead us nowhere. Progress can be made only when we can design precise models that predict when base rates are used, when not, and why.” Here, we have offered such a model. Concretely, our proposal is that people *learn to infer* a posterior from observed information such as the priors, likelihoods and data. Our Learned Inference Model explains a host of effects on belief updating such as under-reaction, belief bias, and memory-dependent subadditivity. Our model renders inference approximate and computationally tractable, making it a plausible process model of human probabilistic inference.

3.8 IMPLEMENTATION DETAILS

3.8.1 BLACKBOX VARIATIONAL INFERENCE

In the main text, the variational optimization problem is stated in terms of minimizing KL divergence. This is useful for clarifying the nature of the problem, but less useful from an algorithmic perspective because the objective function is not tractable (it requires knowledge of the true posterior distribution, which is what we are trying to approximate). Nonetheless, we can obtain a tractable objective function using the following identity:

$$\log P(d) = \mathcal{L}[Q(h|d)] + \mathcal{D}_{\text{KL}}[Q(h|d) || P(h|d)], \quad (3.16)$$

where

$$\mathcal{L}[Q(h|d)] = \mathbb{E}_{Q(h)} [\log P(h, d) - \log Q(h)] \quad (3.17)$$

is the *evidence lower bound* (ELBO), also known as the *negative free energy*. The term ELBO comes from the fact that $\mathcal{L}[Q(h|d)]$ is a lower bound on the “evidence” (log marginal likelihood) $\log P(d)$. Maximizing the ELBO will produce the same variational approximation as minimizing the KL divergence. Critically, the ELBO eliminates the dependence on $P(h|d)$, only requiring access to the unnormalized posterior, the joint distribution $P(h, d)$.

In certain special cases, the ELBO can be tractably computed ², but this is not true for arbitrary joint distributions and approximations. Because the Learned Inference Model uses a flexible neural network function approximator, we adopt an approximate technique for evaluating and optimizing the ELBO known as *blackbox variational inference*³. The key idea is to approximate the gradient of the ELBO with a set of M samples:

$$\nabla \mathcal{L}[Q(h|d)] \approx \frac{1}{M} \sum_{m=1}^M \nabla \log Q(h^m|d) [\log P(h^m, d) - \log Q(h^m)], \quad (3.18)$$

where $h^m \sim Q(h|d)$. Using this approximation, the variational parameters can be optimized with stochastic gradient descent updates of the form:

$$\varphi_{t+1} \leftarrow \varphi_t + \xi_t \nabla \mathcal{L}[Q(h|d)], \quad (3.19)$$

where t indexes iterations and ϱ_t is an iteration-dependent step-size. Provided ϱ_t satisfies the Robbins-Monro stochastic approximation conditions ($\sum_{t=1}^{\infty} \varrho_t = \infty$, $\sum_{t=1}^{\infty} \varrho_t^2 < \infty$), this optimization procedure will converge to the optimal parameters with probability 1.

3.8.2 FUNCTION APPROXIMATION ARCHITECTURE

We used a three-layer neural network architecture as the function approximator for the approximate posterior. Each unit took as input a linear combination of all the units in the layer below, and then passed this linear combination through a nonlinear transfer function. The details of this architecture varied depending on the structure of the inference problem.

When the hypothesis space was binary, the output of the network was a Bernoulli parameter; thus, the network implemented a function $f : \mathcal{D} \mapsto [0, 1]$, where \mathcal{D} denotes the data space, and the variational approximation was $Q(b|d) = \text{Bernoulli}(b; f(d))$. The data space was modeled by 5 input variables: one for the prior parameter, two for the likelihood parameters, and two for the strength and weight of the evidence, and the output space consisted of a single output that represented a Bernoulli parameter. The hidden units use a radial basis function non-linearity, the mean and variance of which were also optimized, and the activation function at the topmost layer was a softmax in order to ensure the final output lay between 0 and 1. To vary the capacity of the network, we vary the number of hidden units; unless otherwise mentioned, networks contain 1 hidden unit since that provides the strongest bottleneck and best demonstrates the effects of interest. We use 2 hidden units only in the replication of the empirical evidence reviewed in ?. Some of the experiments therein are more complex (larger and more varied space of priors, likelihoods and sample

sizes) than the subsequent experiments we model, and we found that while a network with 1 hidden unit still captured the qualitative patterns of interest in the empirical results, it could not capture some of the variation and therefore looked visually less similar to the empirical data. We also use a variant of this function approximation architecture in the section on memory-modulated subadditivity, where the number of inputs increases to 12, and the output is a multinomial distribution of dimension 12. Learning a 12 dimensional multinomial is much harder than learning a binomial, so we increase the number of hidden units to 10.

When the hypothesis space was real-valued, the output was a mean and log standard deviation parametrizing a Gaussian distribution; thus, the network implemented a function $f : \mathcal{D} \mapsto \mathbb{R}^2$, and the variational approximation was $Q(h|d) = \mathcal{N}(h; f(d))$. The data space was modeled by three inputs: the prior mean, the mean of the evidence and the number of samples, the output space consisted of two outputs that represented the mean and variance of a normal distribution. The hidden units used a hyperbolic tangent activation function, and the activation function at the topmost layer made no transformation at the node representing the mean, and took an exponential at the node representing the variance to ensure that the final output was greater than zero.

3.9 RULING OUT ALTERNATIVE MODELS IN THE CONTINUOUS DOMAIN

Here we discuss the predictions of a hierarchical Bayesian model that learns about the underlying global variances from experience. We refer to it henceforth as the L-HBM, for learned hierarchical Bayesian model. We find that it cannot reproduce the observed effect of differentially strong reactions to data between the high and the low dispersion condition.

The L-HBM assumes the true generative model described in the section ‘Extension to a continuous domain’. The output y_{kn} for trial n in a block k is drawn from $\mathcal{N}(m_k, s)$. These m_k values are distributed over blocks as $\mathcal{N}(m_o, v)$.

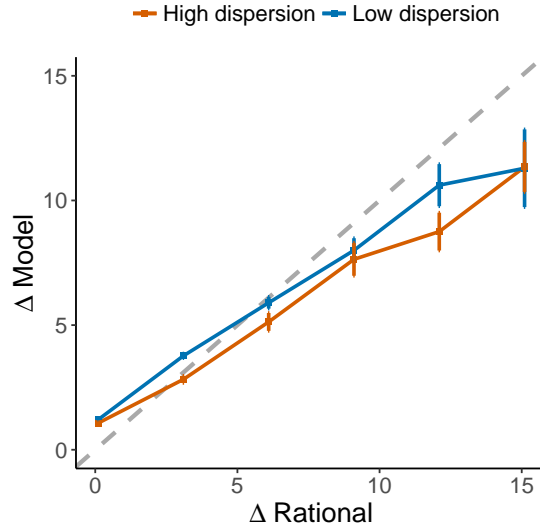


Figure 3.15: Performance of the L-HBM. Simulation results of a hierarchical Bayesian model that infers the underlying parameters in the the experiment reported by². The Y-axis shows the L-HBM’s updates from prior to posterior (Δ Data) and the X-axis shows the update of a rational (hierarchical) model (Δ Rational; a HBM that knows the true parameters for the underlying generative process). Error bars represent the standard error of the mean. Gray line represents $y = x$.

The true values of these parameters are as follows: $s = 25$, $m_o = 40$ for all participants. In the

high dispersion condition $v = 144$ and in the low dispersion condition $v = 36$. The HBM discussed in the main text receives these correct values for the parameters. The L-HBM discussed here has to infer these values. The prior distributions we assume for m_o , s , and v in the L-HBM are $\mathcal{N}(40, 10)$, half-Cauchy(0, 10), and half-Cauchy(0, 10), respectively. It then receives the observations y_{kn} and can form a joint posterior distribution over m_o , s , and v . With these it can then form a posterior predictive distribution for m_k in that block, which we use as the predicted output on each trial.

We compared the resulting updates of this L-HBM to the updates from the HBM in the main text that knows the true parameters of the underlying generative distributions (see Fig. ??). For both the high and the low dispersion conditions, the updates closely follow the diagonal line of $y = x$. This indicates that inferring m_o , s , and v (in addition to m_k) does not result in significant differences in the updates in an ideal observer. Crucially, the L-HBM does not replicate the main qualitative effect of a significant difference in updates between the high and the low dispersion condition, for the same rational update. This means that—unlike our Learned Inference Model—a hierarchical Bayesian model cannot reproduce the qualitative effects observed in the experiment.

4

Amortizing procedures: memory-based meta-learning

This is the Introduction – to be written.

5

The perils of ecological rationality in machines

This is the Introduction – to be written.

References

- [] Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference*, volume 122 (pp. 558): American Psychological Association.
- [] Abbott, J. T., Hamrick, J. B., & Griffiths, T. L. (2013). Approximating Bayesian inference with a sparse distributed memory system. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1686–1691).
- [] Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35(5), 303.
- [] Alon, N., Reichman, D., Shinkar, I., Wagner, T., Musslick, S., Cohen, J. D., Griffiths, T. L., Dey, B., & Ozcimder, K. (2017). A graph-theoretic approach to multitasking. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2097–2106): Curran Associates Inc.
- [] Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2), 5–43.

- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233.
- Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49, 307–343.
- Barbey, A. K. & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30(3), 241–254.
- Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, 24(04), 602–607.
- Beach, L. R. & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3, 439–449.
- Beach, L. R., Wise, J. A., & Barclay, S. (1970). Sample proportions and subjective probability revisions. *Organizational Behavior and Human Performance*, 5(2), 183–190.
- Belousov, B., Neumann, G., Rothkopf, C. A., & Peters, J. R. (2016). Catching heuristics are optimal control policies. In *Advances in Neural Information Processing Systems* (pp. 1426–1434).
- Benjamin, D. J. (2018). *Errors in probabilistic reasoning and judgment biases*. Technical report, National Bureau of Economic Research.

- Benjamin, D. J., Rabin, M., & Raymond, C. (2016). A model of nonbelief in the law of large numbers. *Journal of the European Economic Association*, 14, 515–544.
- Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124(1), 1.
- Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology*, (pp. 85–94).
- Birnbaum, M. H. & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, 45(4), 792.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003a). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014a). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, 74, 35–65.
- Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014b). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, 74, 35–65.

- Bordalo, P., Gennaioli, N., & Shleifer, A. (2017). *Memory, attention, and choice*. Technical report, National Bureau of Economic Research.
- Bramley, N., Rothe, A., Tenenbaum, J., Xu, F., & Gureckis, T. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3), 301–338.
- Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113(2), 409.
- Brown, S. D. & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58, 49–67.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193.
- Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011a). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7, e1002211.

- Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011b). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7, e1002211.
- Carroll, C. D. & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition*, 139, 130–153.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291.
- Cohen, A. L., Sidlowski, S., & Staub, A. (2017). Beliefs and Bayesian reasoning. *Psychonomic Bulletin & Review*, 24(3), 972–978.
- Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3), 393–405.
- Costello, F. & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3), 463–80.
- Costello, F., Watts, P., & Fisher, C. (2018). Surprising rationality in probability judgment: Assessing two competing models. *Cognition*, 170, 280–297.
- Costello, F. J. & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, 100, 1–16.

- [] Dasgupta, I., Schulz, E., & Gershman, S. J. (2017a). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25.
- [] Dasgupta, I., Schulz, E., & Gershman, S. J. (2017b). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25.
- [] Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, 178, 67–81.
- [] Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans’ choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215.
- [] Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711.
- [] Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 4(6), 396–400.
- [] Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7, 889–904.
- [] De Freitas, N., Højén-Sørensen, P., Jordan, M. I., & Russell, S. (2001). Variational mcmc. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 120–127).: Morgan Kaufmann Publishers Inc.

- Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013a). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, 126(2), 280–300.
- Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013b). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126, 285–300.
- Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013c). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126, 285–300.
- Dhimi, M. K. & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of behavioral decision making*, 14(2), 141–168.
- Dougherty, M., Gettys, C. F., & Ogden, E. E. (1999a). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, 106(1), 180–209.
- Dougherty, M., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, 70, 135–148.
- Dougherty, M. & Hunter, J. (2003a). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31, 968–982.
- Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999b). MINERVA-DM: A Memory Processes Model for Judgments of Likelihood. *Psychological Review*, 106, 180–209.

- Dougherty, M. R. & Hunter, J. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31(6), 968–982.
- Dougherty, M. R. & Hunter, J. E. (2003c). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113(3), 263–282.
- DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, 85, 66–74.
- Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. *Judgment under uncertainty: Heuristics and biases*, (pp. 249–267).
- Edwards, W. (1968). Conservatism in human information processing. *Formal Representation of Human Judgment*, 17, 51.
- Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). Medical problem solving an analysis of clinical reasoning.
- Erev, I. & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112, 912–931.
- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3), 519.
- Eslami, S. A., Tarlow, D., Kohli, P., & Winn, J. (2014). Just-in-time learning for fast and flexible inference. In *Advances in Neural Information Processing Systems* (pp. 154–162).

- [] Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3), 295–306.
- [] Evans, J. S. B., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in bayesian inference. *Memory & Cognition*, 30(2), 179–190.
- [] Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, 14(1), 129–146.
- [] Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, 119(3), 459–467.
- [] Fischhoff, B. & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition*, 12, 402–410.
- [] Fischhoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90, 239–260.
- [] Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational behavior and human performance*, 23(3), 339–359.
- [] Fleming, S. M. & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91.

- Fox, C. R. & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science*, 44(7), 879–895.
- Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998–998.
- Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4, e1000211.
- Ganguly, A. R., Kagel, J. H., & Moser, D. V. (2000). Do asset market prices reflect traders’ judgment biases? *Journal of Risk and Uncertainty*, 20(3), 219–245.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1–58.
- Gennaioli, N. & Shleifer, A. (2010). What comes to mind. *The Quarterly Journal of Economics*, 125, 1399–1433.
- Gershman, S. J. (2017). On the blessing of abstraction. *The Quarterly Journal of Experimental Psychology*, 70(3), 361–365.
- Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*.
- Gershman, S. J. & Beck, J. M. (2017). Complex probabilistic inference.

- Gershman, S. J. & Goodman, N. (2014a). Amortized inference in probabilistic reasoning.

In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 517–522).: Cognitive Science Society.
- Gershman, S. J. & Goodman, N. D. (2014b). Amortized Inference in Probabilistic Reasoning.

Proceedings of the 36th Annual Conference of the Cognitive Science Society, 1, 517–522.
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015a). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015b). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.
- Gershman, S. J., Markman, A. B., & Otto, R. A. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143(1), 182–194.
- Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012a). Multistability and Perceptual Inference. *Neural Computation*, 24(1), 1–24.
- Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012b). Multistability and perceptual inference. *Neural Computation*, 24, 1–24.

- Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012c). Multistability and perceptual inference. *Neural Computation*, 24, 1–24.
- Gershman, S. J. & Wilson, R. (2010). The neural costs of optimal control. In *Advances in neural information processing systems* (pp. 712–720).
- Gershman, S. J., Zhou, J., & Kommer, C. (2017). Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, 29(12), 2103–2113.
- Gettys, C. F. & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, 24, 93–110.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103, 592–596.
- Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, 3(1), 20–29.
- Gigerenzer, G. & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143.
- Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62, 451–482.
- Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.

- Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227.
- Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science: A Multidisciplinary Journal*, 32(1), 108–154.
- Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in psychology*, 4.
- Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95(3), 537–557.
- Griffin, D. & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Griffiths, T. L. (2015). Revealing ontological commitments by magic. *Cognition*, 136, 43–48.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7, 217–229.
- Griffiths, T. L. & Tenenbaum, J. B. (2006a). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.
- Griffiths, T. L. & Tenenbaum, J. B. (2006b). Optimal predictions in everyday cognition. *Psychological Science*, 17(9), 767–773.

- Griffiths, T. L. & Tenenbaum, J. B. (2011). Predicting the future as bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, 140, 725–743.
- Griffiths, T. L., Vul, E., & Sanborn, a. N. (2012a). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, 21(4), 263–268.
- Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012b). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21, 263–268.
- Grinnell, M., Keeley, S. M., & Doherty, M. E. (1971). Bayesian predictions of faculty judgments of graduate school success. *Organizational Behavior and Human Performance*, 6(3), 379–387.
- Gu, S. S., Ghahramani, Z., & Turner, R. E. (2015). Neural adaptive sequential monte carlo. In *Advances in Neural Information Processing Systems* (pp. 2629–2637).
- Hadjichristidis, C., Stibel, J., Sloman, S., Over, D., & Stevenson, R. (1999). Opening pandora’s box: Selective unpacking and superadditivity. In *Proceedings of the European Society for the Study of Cognitive Systems 16th Annual Workshop*.
- Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90, 649–660.

- Hamrick, J. B., Battaglia, P., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (pp. 1545–1550).
- Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015a). Think again? the amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the Thirty-seventh Annual Conference of the Cognitive Science Society*.
- Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015b). Think again? the amount of mental simulation tracks uncertainty in the outcome. In *CogSci: Citeseer*.
- Hawkins, G. E., Hayes, B. K., & Heit, E. (2016). A dynamic model of reasoning and memory. *Journal of Experimental Psychology: General*, 145(2), 155–180.
- Hayes, B. K., Fritz, K., & Heit, E. (2013). The relationship between memory and inductive reasoning: Does it develop? *Developmental Psychology*, 49(5), 848–860.
- Hayes, B. K. & Heit, E. (2013). How similar are recognition memory and inductive reasoning? *Memory & Cognition*, 41(5), 781–795.
- Heckerman, D. (1990). A Tractable Inference Algorithm for Diagnosing Multiple Diseases in The QMR model.
- Heit, E. & Hayes, B. K. (2011). Predicting reasoning from memory. *Journal of Experimental Psychology: General*, 140(1), 76–101.

- Hennig, P., Osborne, M. A., & Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179), 20150142.
- Hertwig, R. & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523.
- Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and description: Exploring two paths to knowledge. *Current Directions in Psychological Science*, 27(2), 123–128.
- Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, 138(2), 211–237.
- Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431.
- Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The “wake-sleep” algorithm for unsupervised neural networks. *Science*, 268, 1158–1160.
- Jaakkola, T. S. & Jordan, M. I. (1999). Variational Probabilistic Inference and the QMR-DT Network. *Journal of Artificial Intelligence Research*, 10, 291–322.
- Janis, I. & Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology*, 33(1), 73–77.

- Jazayeri, M. & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138), 912–915.
- Johnson, E. J. & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, 31(4), 395–414.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233.
- Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment*, 15, 292–307.
- Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences*, 113(45), 12868–12873.
- Kingma, D. P. & Welling, M. (2013a). Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*.
- Kingma, D. P. & Welling, M. (2013b). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

- Klein, G. (1999). *Sources of Power: How People Make Decisions*. MIT press.
- Knill, D. C. & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
- Koehler, D. (1994). Hypothesis Generation And Confidence in Judgment. *Learning, Memory*.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1), 1–17.
- Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 28(9), 1321–1333.
- Körding, K. P. & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7), 319–326.
- Koriati, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6, 107–118.
- Kraemer, C. & Weber, M. (2004). How do people take into account weight, strength and quality of segregated vs. aggregated data? experimental evidence. *Journal of Risk and Uncertainty*, 29, 113–142.
- Krynski, T. R. & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430.

- [] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.
- [] Li, Y., Turner, R. E., & Liu, Q. (2017). Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*.
- [] Lieder, F. & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124, 762–794.
- [] Lieder, F. & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, (pp. 1–85).
- [] Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in Neural Information Processing Systems* (pp. 2690–2798).
- [] Lieder, F., Griffiths, T. L., & Goodman, N. D. (2013). Burn-in , bias , and the rationality of anchoring. *Advances in Neural Information Processing Systems* 25, 25, 1–9.
- [] Lieder, F., Griffiths, T. L., & Hsu, M. (2018a). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1), 1.
- [] Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018b). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25(1), 322–349.

- [] Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018c). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, 25(2), 775–784.
- [] Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*.
- [] Luu, L. & Stocker, A. A. (2016). Choice-dependent perceptual biases.
- [] Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4), 287–298.
- [] MacKay, D. J. (2003). Information theory, inference and learning algorithms.
- [] Marchiori, D., Di Guida, S., & Erev, I. (2015). Noisy retrieval models of over-and undersensitivity to rare events. *Decision*, 2, 82–106.
- [] Marewski, J. N. & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5, 39–59.
- [] Marewski, J. N. & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, 118, 393–437.
- [] Marino, J., Yue, Y., & Mandt, S. (2018). Learning to infer.
- [] Marr, D. & Poggio, T. (1976). From understanding computation to understanding neural circuitry.

- Massey, C. & Wu, G. (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 51(6), 932–947.
- McClelland, J. L. & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105, 724–760.
- Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207.
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269–275.
- Mercier, H. & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
- Mnih, A. & Gregor, K. (2014). Neural variational inference and learning in belief networks. In *International Conference on Machine Learning* (pp. 1791–1799).
- Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, 7, e32548.
- Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108, 12491–12496.

- Naesseth, C. A., Linderman, S. W., Ranganath, R., & Blei, D. M. (2017). Variational sequential monte carlo. *arXiv preprint arXiv:1705.11140*.
- Neil Bearden, J. & Wallsten, T. S. (2004). Minerva-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making*, 17(5), 349–363.
- Newstead, S. E., Pollard, P., Evans, J. S. B., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, 45(3), 257–284.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.
- Oakhill, J., Johnson-Laird, P., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, 31(2), 117–140.
- Oaksford, M. & Chater, N. (2007a). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- Oaksford, M. & Chater, N. (2007b). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes*, 42, 343–363.
- Oliva, A. & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), 520–527.

- [] Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92, 530–543.
- [] Orhan, A. E. & Ma, W. J. (2017). Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, 8(1), 138.
- [] Paige, B. & Wood, F. (2016). Inference networks for sequential Monte Carlo in graphical models. In *International Conference on Machine Learning* (pp. 3040–3049).
- [] Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, 102, 127–144.
- [] Pecevski, D., Buesing, L., & Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, 7, e1002294.
- [] Peterson, C. & Miller, A. (1964). Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology*, 68, 363.
- [] Peterson, C. R., DuCharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology*, 76, 236–243.
- [] Peterson, C. R. & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70(1), 117.

- Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, 69, 522–527.
- Peterson, C. R. & Ulehla, Z. (1964). Uncertainty, inference difficulty, and probability learning. *Journal of Experimental Psychology*, 67, 523–530.
- Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19, 285–293.
- Phillips, L. D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72, 346.
- Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics* (pp. 814–822).
- Rasmussen, C. & Ghahramani, Z. (2003). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 15* (pp. 505–512).: MIT Press.
- Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine discounting unspecified possibilities. *Medical Decision Making*, 15(3), 227–230.
- Reyna, V. F. & Lloyd, F. J. (2006). Physician decision making and cardiac risk: effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied*, 12(3), 179.

- Rezende, D. & Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* (pp. 1530–1538).
- Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning* (pp. 1278–1286).
- Rieskamp, J. & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135, 207–236.
- Ritchie, D., Thomas, A., Hanrahan, P., & Goodman, N. (2016). Neurally-guided procedural models: Amortized inference for procedural graphics programs using neural networks. In *Advances in Neural Information Processing Systems* (pp. 622–630).
- Robert, C. & Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.
- Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo*, 4(10.1201).
- Ross, B. H. & Murphy, G. L. (1996). Category-based predictions: influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 736–753.
- Ruiz, F. J. R. & Titsia, M. K. (2019). A contrastive divergence for combining variational inference and mcmc. *arXiv preprint arXiv:1905.04062*.

- Rule, J., Schulz, E., Piantadosi, S. T., & Tenenbaum, J. B. (2018). Learning list concepts through program induction. *BioRxiv*, (pp. 321505).
- Saeedi, A., Kulkarni, T. D., Mansinghka, V., & Gershman, S. J. (2017a). Variational particle approximations. *Journal of Machine Learning Research*, 18, 1–29.
- Saeedi, A., Kulkarni, T. D., Mansinghka, V. K., & Gershman, S. J. (2017b). Variational particle approximations. *The Journal of Machine Learning Research*, 18, 2328–2356.
- Samuels, R., Stich, S., & Bishop, M. (2012). Ending the rationality wars. *Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010*, 2, 191.
- Sanborn, A. N. & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20, 883–893.
- Sanborn, A. N. & Griffiths, T. L. (2009). A Bayesian Framework for Modeling Intuitive Dynamics. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1145–1150).
- Sasaki, S. & Kawagoe, T. (2007). Belief updating in individual and social learning: A field experiment on the internet.
- Schulz, E., Speekenbrink, M., & Meder, B. (2016). Simple Trees in Complex Forests: Growing Take The Best by Approximate Bayesian Computation. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2531–2536). Austin, TX: Cognitive Science Society.

- Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, 99, 44–79.
- Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, 61(2), 195.
- Shah, A. K. & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134, 207–222.
- Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science*, 3(2), 143–162.
- Shi, L. & Griffiths, T. L. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In *Advances in Neural Information Processing Systems* (pp. 1669–1677).
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010a). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, 17, 443–464.
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010b). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, 17(4), 443–464.
- Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.

- Shwe, M. A. & Cooper, G. (1991). An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Computers and biomedical research, an international journal*, 24(5), 453–475.
- Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H., & Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods of information in Medicine*, 30(04), 241–255.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Simoncelli, E. P. & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1), 1193–1216.
- Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004a). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3), 573–582.
- Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004b). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3), 573–582.
- Slovic, P. & Lichtenstein, S. (1971). Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6(6), 649–744.

- Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, 128(1), 64–75.
- Smith, K. A. & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1), 185–199.
- Sprenger, A. M., Dougherty, M., Atkins, S. M., Franco-Watkins, A. M., Thomas, R., Lange, N., & Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, 2, 129.
- Stanford, P. K. (2010). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford University Press.
- Stewart, N., Chater, N., & Brown, G. D. (2006a). Decision by sampling. *Cognitive Psychology*, 53, 1–26.
- Stewart, N., Chater, N., & Brown, G. D. (2006b). Decision by sampling. *Cognitive Psychology*, 53(1), 1–26.
- Stocker, A. A. & Simoncelli, E. P. (2008). A bayesian model of conditioned perception. In *Advances in Neural Information Processing Systems* (pp. 1409–1416).
- Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. In *Advances in Neural Information Processing Systems* (pp. 3048–3056).

- Suchow, J. W., Bourgin, D. D., & Griffiths, T. L. (2017). Evolution in mind: Evolutionary dynamics, cognitive processes, and Bayesian inference. *Trends in Cognitive Sciences*, 21, 522–530.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332(6033), 1054–9.
- Tenenbaum, J. B. & Griffiths, T. (2001). The rational basis of representatives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.
- Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*.
- Thomas, R. P., Dougherty, M., Sprenger, A. M., & Harbison, J. I. (2008a). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155–185.
- Thomas, R. P., Dougherty, M. R., & Buttaccio, D. R. (2014). Memory constraints on hypothesis generation and decision making. *Current Directions in Psychological Science*, 23(4), 264–270.
- Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008b). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155–185.

- Todd, P. M. & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3), 167–171.
- Todd, P. M. & Goodie, A. S. (2002). Testing the ecological rationality of base rate neglect. In *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior* (pp. 215–223).: From animals to animats (MIT Press/Bradford Books, 2002).
- Trueblood, J. S. & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, 35(8), 1518–1552.
- Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.
- Tversky, A. & Kahneman, D. (1974a). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, 185(4157), 1124–1131.
- Tversky, A. & Kahneman, D. (1974b). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Tversky, A. & Koehler, D. J. (1994a). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.

- [] Tversky, A. & Koehler, D. J. (1994b). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4), 547.
- [] Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, 27(4), 455–480.
- [] Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from bayes’s theorem and the additivity principle. *Memory & cognition*, 30(2), 171–178.
- [] Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in Neural Information Processing Systems* (pp. 1955–1963).
- [] Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014a). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637.
- [] Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014b). One and done? optimal decisions from very few samples. *Cognitive Science*, 38, 599–637.
- [] Vul, E. & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19, 645–647.
- [] Wang, T., Wu, Y., Moore, D., & Russell, S. J. (2018). Meta-Learning MCMC Proposals. In *Advances in Neural Information Processing Systems* (pp. 4150–4160).

- Wang, Z. & Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science*, 5(4), 689–710.
- Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences*, 111(26), 9431–9436.
- Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1151–1164.
- Wheeler, G. & Beach, L. R. (1968). Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, 3(1), 36–46.
- Whittington, J. C. & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*.
- Windschitl, P. D. & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 198.
- Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, 6, e1000871.
- Yildirim, I. & Kulkarni, T. D. Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations.

- [] Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
- [] Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2018). Bayesian inference causes incoherence in human probability judgments.