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

Dissertation advisor: Professor Samuel J. Gershman Ishita Dasgupta

Algorithmic approaches to ecological rationality in humans and machines

Abstract

should state the problem, describe the methods and procedures used, and give the main results or conclusions of the research.

Contents

This is the dedication.

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer 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 consectetuer. 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.

1

Introduction

This is the Introduction – to be written.

2

Approximations 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.

Kepler is not alone: the history of science is replete with examples of "unconceived alternatives"[347], 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 $P(h|d)$ to each hypothesis $h \in \mathcal{H}$ proportional to its prior probability $P(h)$ and the likelihood of the observed data under h, $P(d|h)$:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}. \tag{2.1}$$

Many studies have found that when $\mathcal{H}$ is supplied explicitly, humans can come close to the Bayesian ideal e.g.,[146,148,104,287,263].* However, when humans must generate the set of hypotheses themselves, they cannot generate them all and instead generate only a subset, leading to judgment biases[358,74,45,205,384,127]. Some prominent biases of this kind are listed in Table 3.1.

Most previously proposed models of hypothesis generation rely on cued recall from memory based on similarity to previously observed scenarios c.f.[358]. 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.[137] 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 con-

*This 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 see[47] for discussion.

Table 2.1: 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).	[103]
Superadditivity	Perceived probability of a hypothesis is lower when the hypothesis is described as a disjunction of atypical component hypotheses (unpacked to atypical examples).	[339, 154]
Weak evidence effect	The probability of an outcome is judged to be lower when positive evidence for a weak cause is presented	[95]
Dud alternative effect	The judged probability of a focal outcome is higher when implausible alternatives are presented	[388]
Self-generation effect	The probability judgment over hypotheses that participants have generated themselves is lower as compared to the same hypotheses generated by others	[205,200]
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.	[377]
Anchoring and Adjustment	Generated hypotheses are biased by the hypothesis that is prompted at the start.	[365]

straints 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 number of samples that can be generated, giving rise to systematic biases. We explore what sampler designs can reproduce the phenomena listed in Table 3.1, and then test our theory's novel predictions in two experiments.

2.1 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 domains[47]. 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"[238], 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 3.1 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 see[149] 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.

2.1.1 Monte Carlo methods

In their simplest form, sample-based approximations also known as *Monte Carlo* approximations;[303], take the following form:

$$P(h|d) \approx \hat{P}_N(h|d) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[h_n = h], \tag{2.2}$$

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

In general, a bounded reasoner cannot directly sample from the posterior, because the normalizing constant $P(d) = \sum_h P(h,d)$ 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 exact 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 prob-

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

ability matching"[389,68,255,375]: 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[348] 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 Hamrick et al.[157]. Fourth, Markov chains generate autocorrelated samples, consistent with autocorrelation in hypothesis generation[122,377,30]. Correlation between consecutive hypotheses manifested as anchoring effects (where judgments are biased by the initial hypothesis[365]) 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[219]. Finally, work in theoretical neuroscience has shown how MCMC algorithms could be realized in generic cortical circuits[41,279,255].

We have emphasized properties that emerge in the finite sample regime because people tend to only generate a small number of hypotheses[198,305,127,384,73]. Although this may seem to be manifestly sub-optimal, it can be justified within a "resource-rational" framework[375,145,119,319]. 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, Hamrick et al.[157] showed that people generated more hypotheses when they were more uncertain. By the same token, cognitive load[346] or response time pressure[74] act as disincentives, reducing the number of generated hypotheses.

2.1.2 A specific Markov chain Monte Carlo algorithm

The space of MCMC algorithms is vast[303], 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 h' are drawn from a proposal distribution $Q(h'|h_n)$, where h_n is the hypothesis at step n. This proposal is accepted or rejected according to:

$$P(h_{n+1} = h'|h_n) = \min\left[1, \frac{P(d|h')P(h')Q(h_n|h')}{P(d|h_n)P(h_n)Q(h'|h_n)}\right]. \tag{2.3}$$

If the proposal is rejected, then the chain stays at the same hypothesis, $h_{n+1} = h_n$. Although the posterior cannot be directly evaluated, we assume it can be computed up to a normalizing constant, since $P(h|d) \propto P(d|h)P(h)$. 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, $Q(h'|h) = Q(h|h')$. This reduces the acceptance function to:

$$P(h_{n+1} = h'|h_n) = \min\left[1, \frac{P(d|h')P(h')}{P(d|h_n)P(h_n)}\right]. \tag{2.4}$$

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 networks[1] and combinatorial spaces[3,4,2]. This locality has been shown to be optimal as a foraging strategy[17,2] 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 in[21,9]; 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 hypotheses

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[3,8,4]. 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.

2.2 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.

2.2.1 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[52]. To address this problem, approximate probabilistic inference algorithms (including Monte Carlo methods) are now widely-established e.g.,[330,177,162]. 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 3.1 have been documented in clinical settings[292,83,384]; 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 $d = 1, \ldots, D$. The diseases are connected to the symptoms via a noisy-or likelihood, following[331]:

$$P(k_s = 1|h) = 1 - (1 - \varepsilon) \prod_{d=1}^{D} (1 - w_{ds})^{h_d}, \tag{2.5}$$

where $k_s = 1$ when symptom $s = 1, \ldots, S$ 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 2.2). 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.

Table 2.2: Parameters used for noisy-or model.

Symptoms	Diseases								
	lung cancer	TB	resp. flu	cold	gastro-enteritis	stomach cancer	stomach flu	food poi-soning	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
bloating	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

2.2.2 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 sampling[375]. This gives rise to *subadditivity* (see Table 3.1): 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)[368,74], despite the fact that mathematically these are equal. For example, the probability of a 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 exam-

ple 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 2.1.

Experiments in[74] 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 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 2.1.

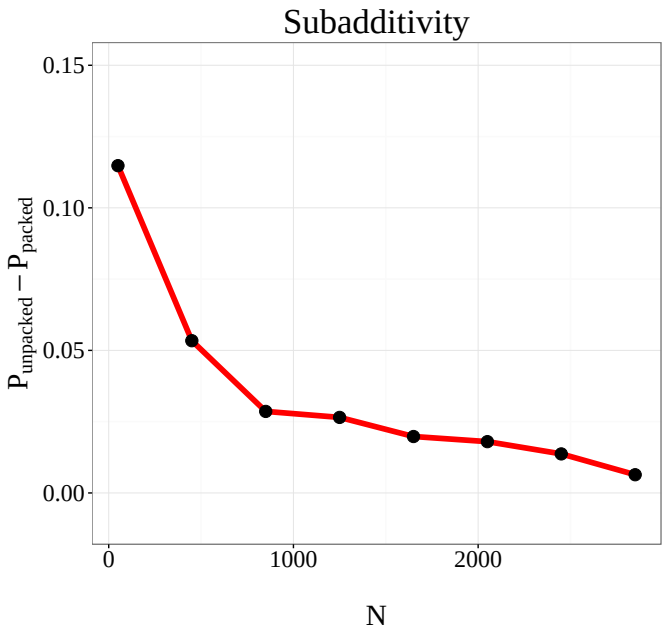


Figure 2.1: Subadditivity. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Packed: what is the probability that these symptoms were caused by the presence of a gastrointestinal disease? (b) Unpacked to typical examples: what is the probability 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.

2.2.3 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 3.1): 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)[339,154], despite the fact that mathematically they should be equal. This effect occurs when unpacking to atypical (low probability) examples and subadditvity 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

$$\sum_{h\in\mathcal{A}} \hat{P}(h|d) = \sum_{h\in\mathcal{A}} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[h_n = h] = \frac{\sum_{h\in\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h]}{\sum_{h\in\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h] + \sum_{h'\notin\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h']} \quad (2.6)$$

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

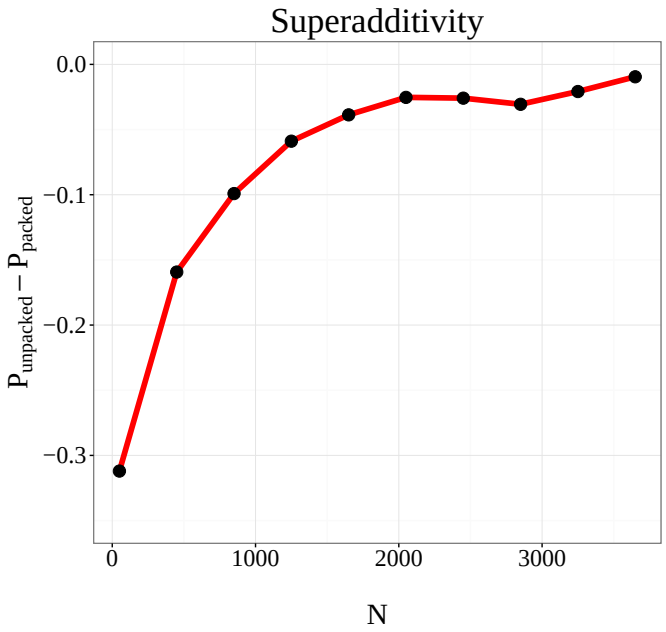


Figure 2.2: Superadditivity. MCMC-estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Packed: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Unpacked to atypical examples: What is the probability 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 represnted by the red line. This effect diminishes as the number of samples increases.

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 2.2.

Previous accounts of subadditivity e.g.,[358,258] 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 by[53] 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 superadditvity, 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.[339] explain superadditivity by suggesting that atypical examples divert attention from more typical exam-

ples and thus lower the probability estimate. But an explanation at the level of a rational 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 cause[95]. 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 2.2). 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 generation of more alternative hypotheses outside the focal space and a lower probability judgment than in the first case (Figure 2.3).

Another such bias is the *Dud alternative effect*: presenting low probability (or "dud") alternate hypotheses increases the perceived probability of the focal space of hypotheses[388]. This can be viewed

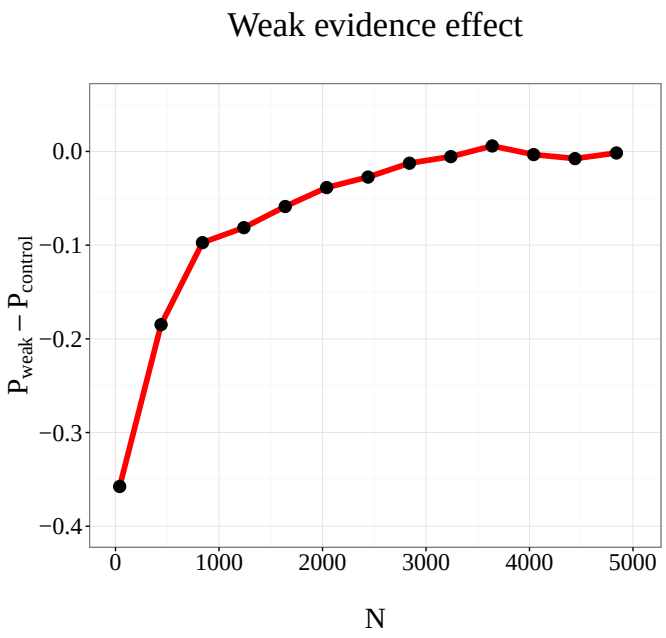


Figure 2.3: Weak evidence effect. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Control: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Evidence for a weak cause: What is the probability 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.

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 2.4 that the model predicts that the scenario with dud alternatives produces higher probability judgments than the control. Findings in[388] 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.

2.2.4 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[200,205]. 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., not accept any of the proposals and thus not generate them at all. This, in the small sample limit, results in the

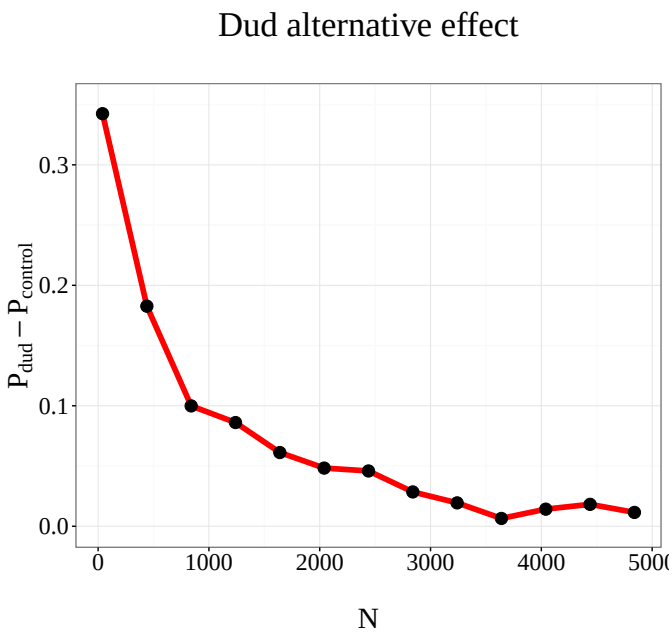


Figure 2.4: Dud alternative effect. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Control: What is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of a gastrointestinal disease)?, (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

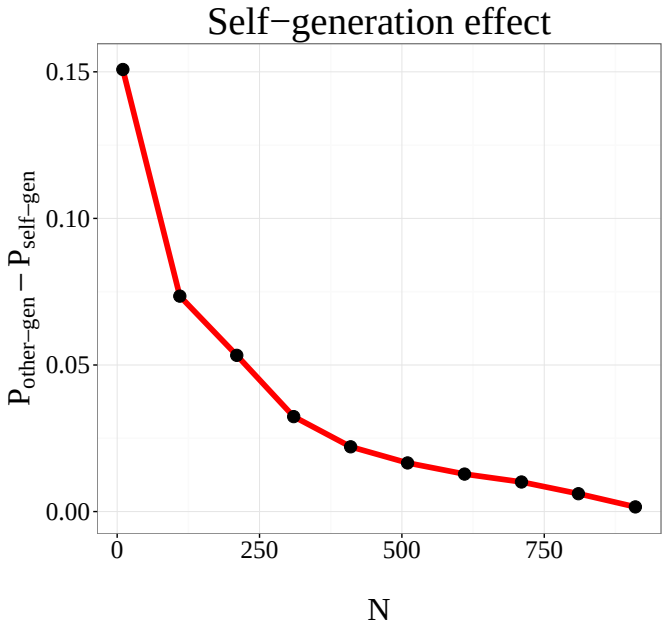


Figure 2.5: Self-generation effect. MCMC estimates for the following query: Given the symptoms *fever* and *fatigue*, (a) Self-generated: What are the two most likely respiratory diseases to have caused these symptoms? Estimate the probability that these symptoms are caused by either of these two diseases. (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

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 in [200] 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 2.5.

While this effect has previously been understood in terms of the generation of alternatives[200], 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.

2.2.5 Anchoring and adjustment

In a classic experiment, Tversky & Kahneman[365] 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 Tversky & Kahneman[365] 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.

Lieder et al.[219] 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 autocorrelation 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 anchor-

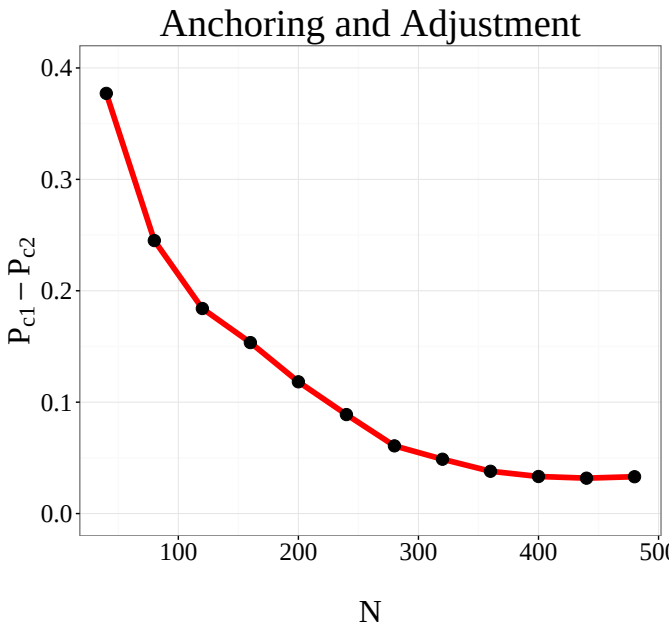


Figure 2.6: Anchoring and adjustment. The y axis represents the difference in the probabilities of respiratory flu and stomach flu given the symptoms *fever* and *fatigue* as returned by two different chains that are initialized differently. The chains are initialized in the two different clusters, at 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.

ing 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 2.6 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.

2.2.6 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*[354]. 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 samples gets one closer to the asymptotic regime, where the estimates are exact and the error tends to zero.

This error analysis was extended by[377] 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 2.7. 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 chain*, 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 in[377]—the error of the responses monotonically declines with the number of samples, and the error reduction is greater when averaging across (compared to within) individuals.

Our MCMC model can replicate this effect because it generates auto-correlated samples. The last

*We 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.[377] find this to not be the case empirically.

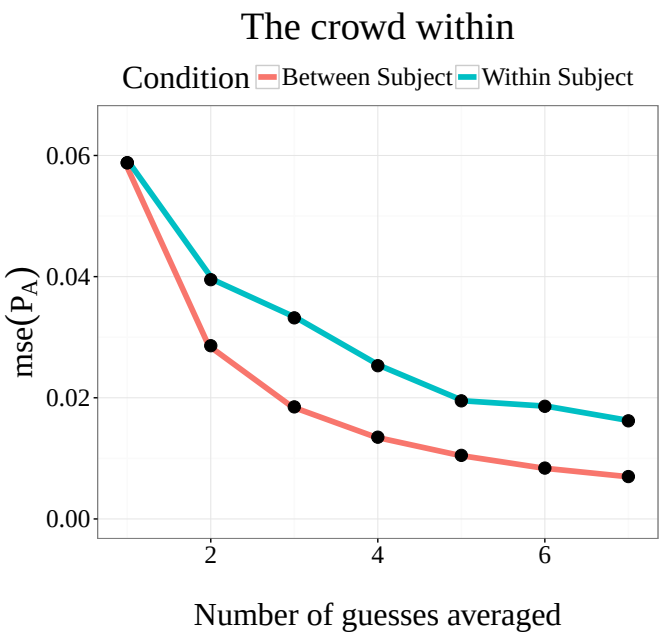


Figure 2.7: The crowd within. Errors in the MCMC estimates for the following query: Given the symptoms *nausea* and *shortness of breath*, what is the probability that these were caused by the presence of a respiratory disease? The estimates are averaged either over samples from the same individual (blue) or over samples from different individuals (red)

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.

2.2.7 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 $Q(h)$, but unlike MCMC it samples multiple hypotheses independently and in parallel. These samples are then weighted to obtain an approximation of the posterior:

$$\hat{P}_N(h|d) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[h_n = d] w_n, \tag{2.7}$$

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

$$w_n \propto \frac{P(h_n, d)}{Q(h_n)}. \tag{2.8}$$

Intuitively, the importance weight corrects for the fact that the importance sampler draws samples from the wrong distribution. Shi et al.[327] 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, Shi & Griffiths[326] demonstrated how importance sampling could be realized in a biologically plausible neural circuit see also[2].

Some of the effects we have replicated in this work could also be captured by an importance sampling algorithm with limited samples. Thomas et al. [358] 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 it they are generated without prompting (self-generated). However, Experiment 2 in [200] 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 3.1 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 tracking[374] and change detection[39]. 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.

2.3 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 [12,335]. 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 Greene [141]. Specifically, Greene [141] 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) model [28], 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 $h \in \mathcal{H}$ 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 $P(h \in H|d)$ that h 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.

2.4 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 2.3: 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 3.2, 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 3.4.

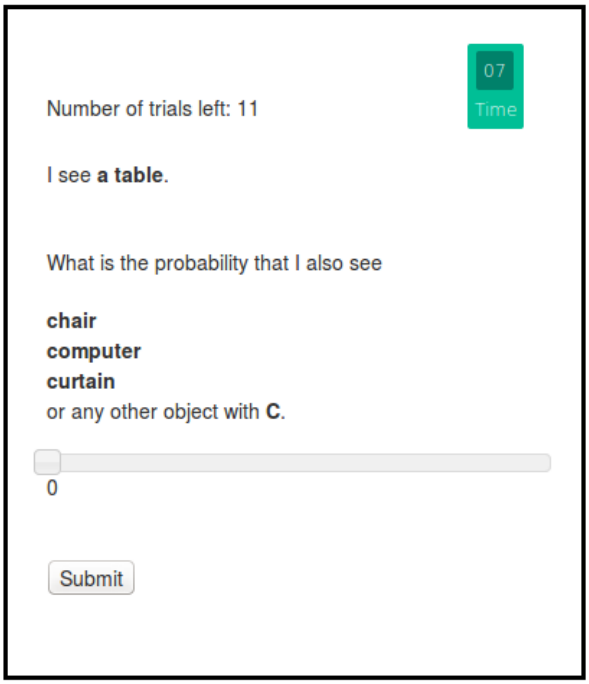


Figure 2.8: Experimental setup. Participants were asked to estimate the conditional probability using a slider bar within a 20-second time limit.

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 in[141]. This value for N as the number of samples is higher than numbers found in some previous work like[375] 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 2.9). 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 by[339], 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.

2.5 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

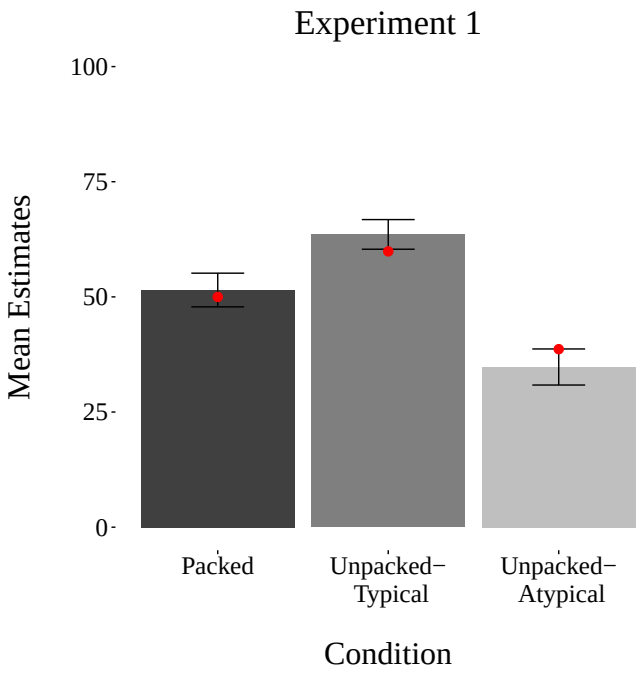


Figure 2.9: Experiment 1 results. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene.

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 3.3). Each participant responded to one unpacked-typical, one unpacked-atypical and one packed scenario in random order.

Table 2.4: 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 2.10, 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 3.3) 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.

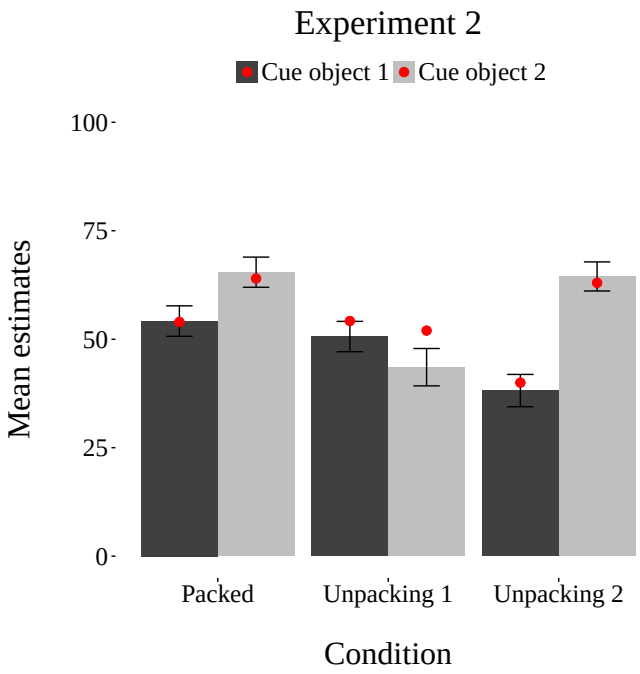


Figure 2.10: Experiment 2 results. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene. Unpacking 1 is typical for cue object 1 and atypical for cue object 2; unpacking 2 is typical for cue object 2 and atypical for cue object 1.

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.

2.6 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 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, Thomas et al.[358] 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[375]. Like other models based on Monte Carlo methods e.g.,[122,219,327], our model predicts exact Bayesian inference in the limit of large sample sizes. However, cognitively bounded agents are expected to be *computationally rational*[119]: sampling takes time and energy, and hence the optimal sampling strategy will tend to generate relatively few hypotheses[375].

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 2.5 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 sampling[327] or cued recall[358] is that the generation and the evaluation processes are largely

Table 2.5: 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	×*	✓	✓
Crowd within	×	✓	✓
Anchoring & adjustment	×	×	✓

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).[219] 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[122,377,30]. For example, the "crowd within" phenomenon[377], 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.

*While 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 in[200]

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 rates[303], there is reason to believe that human probability estimation makes uses of local proposal distributions. Evidence for locality has been found in domains analogous to that of hypothesis generation[1,342], further suggesting that humans use local proposal distributions.

Previous work demonstrating the effect of superadditivity[339] did not find subadditivity except in situations where the search was over an ill-defined fuzzy category, such that unpacked typical examples 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 in[339], 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.

2.7 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[Yildirim & Kulkarni,118]. 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[119,375,145]. 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 projectile[355], judging mass in collisions[315], and judging the balance of block towers[156]. Furthermore, research also suggests that humans sample noisy simulations of future trajectories[343,157], 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.

3

Amortization: Re-use of previous solutions

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 models[150,264]. 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., [357,56,376,123,69,314,370]. 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:

$$P(\text{toaster} \wedge \text{microwave}|\text{sink}) = P(\text{toaster}|\text{sink}, \text{microwave})P(\text{microwave}|\text{sink}). \tag{3.1}$$

Even though this is a trivial example, standard inference engines do not exploit these kinds of regularities 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*[352,117]. 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.[*] 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.

[*] Formally, this is known as *variational inference*[185], 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.

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 samples[56,376], we expect this bias to be measurable.

Amortization has a long history in machine learning; the *locus classicus* is the Helmholtz machine[65,173], which uses samples from the generative model to train a recognition model. More recent extensions and applications of this approach e.g.,[298,272,195,300] have ushered in a new era of scalable Bayesian computation in machine learning. We propose that amortization is also employed by the brain see[391] 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 Dasgupta et al.[56]. 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[56], 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[223,221,376].

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.

3.1 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.

3.1.1 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 $P(h)$:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \tag{3.2}$$

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, $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \mathcal{H}_K$ is the product space of latent objects, so if there are K latent objects and M possible objects, $|\mathcal{H}| = M^K$. If we imagine there

are $M = 1000$ 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 $\theta = \{h_1, \ldots, h_N\}$ from the posterior distribution over the hypothesis space. We can understand Monte Carlo methods as producing a recognition model $Q_\theta(h|d)$ parametrized by θ see [311] for a systematic treatment. In the idealized case, each hypothesis is sampled from $P(h|d)$. The approximation is then given by:

$$P(h|d) \approx Q_\theta(h|d) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[h_n = h], \tag{3.3}$$

where $\mathbb{I}[\cdot] = 1$ if its argument is true (and o otherwise). The accuracy of this approximation improves with N, but from a decision-theoretic perspective even small N may be serviceable [376,221,120].

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 approximation [230]. 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 h 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 scale [348]. 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 observations [158]. Finally, Markov chains also generate autocorrelated samples. This is consistent with autocorrelation in hypothesis generation [31,123,377,221]

Table 3.1: 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). $P(\mathrm{A}_{typical} \cup \mathrm{B}) < P(\mathrm{A}_{typical}) + P(\mathrm{B})$	[103, 368]
Superadditivity	Perceived probability of a hypothesis is lower when the hypothesis is described as a disjunction of atypical component hypotheses (unpacked to atypical examples). $P(\mathrm{A}_{atypical} \cup \mathrm{B}) > P(\mathrm{A}_{atypical}) + P(\mathrm{B})$	[340,154]

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 h' from a proposal distribution $\varphi(h'|h_n)$ and then accepts this proposal with probability

$$P(h_{n+1} = h'|h_n) = \min\left[1, \frac{P(d|h')P(h')\varphi(h_n|h')}{P(d|h_n)P(h_n)\varphi(h'|h_n)}\right]. \tag{3.4}$$

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, $\varphi(h'|h_n)$ is usually constructed to sample proposals from a local region around h_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 childhood[370] and creative insight in adulthood[353]. Even low-level visual phenomena like perceptual multistability can be described in these terms[123,255].

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

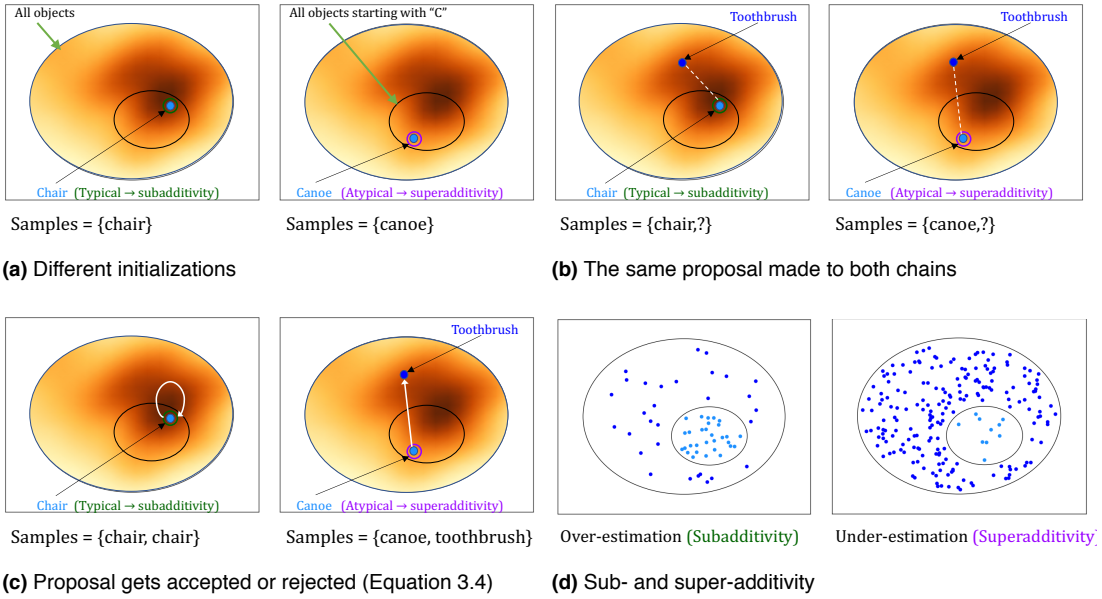


Figure 3.1: Demonstration of how MCMC sampling can give rise to sub- and super-additivity for different unpacked versions of the question : "In the presence of a table, what is the probability that there is also another object starting with C?". The color gradient indicates probability density. (a) The 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.

pled, 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 [218,223]. 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 [56], 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 (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 3.1). 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.[368] and lower when the unpacked exemplars were atypical a "superadditivity" effect; cf.[340] 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 Greene[141], to a latent Dirichlet allocation (LDA) model[28]. 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 object[56]. 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 effect[56]. 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.

3.1.2 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 exploited[266].

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 3.2 illustrates the

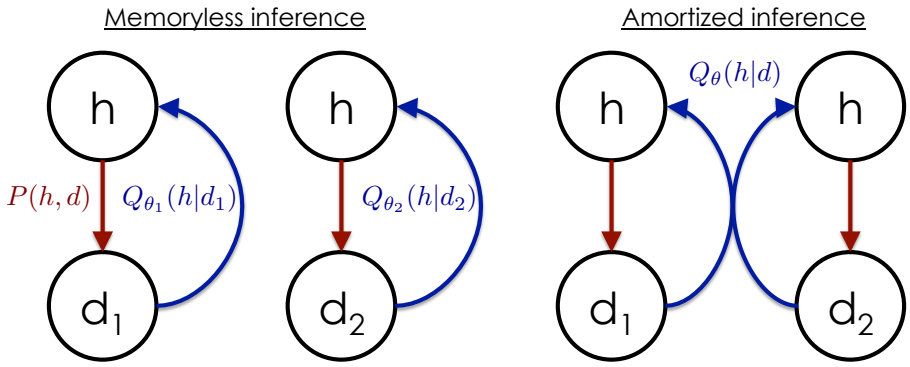


Figure 3.2: Theory schematic. (Left) Standard, memoryless framework in which a recognition model $Q_\theta(h|d)$ approximates the posterior over hypothesis h given data d. The recognition model is parametrized by θ (e.g., a set of samples in the case of Monte Carlo methods). Memoryless inference builds 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."

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 $Q_\theta(h|d)$ 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)—$Q_{\theta_1}(h|d_1)$, $Q_{\theta_2}(h|d_2)$, 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, 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.

3.1.3 Two amortization strategies

In our experiments, we ask participants to sequentially answer pairs of queries (denoted $Q1$ and $Q2$). 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.[*] 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., [352,298,272], 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 $Q2$ can be expressed as the composition of the answer to $Q1$ and another (putatively simpler) computation. In Experiment 2, $Q2$ queries a hypothesis space that is the union of the hypothesis space queried in $Q1$ and a disjoint hypothesis space. For example if $Q1$ is "What is the probability that there is an object starting with a C in the scene?", $Q2$ 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 $Q1$

[*] We focus on sampling-based amortization strategies because our earlier experiments support the idea that human probability judgment is sample-based [56]. 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).

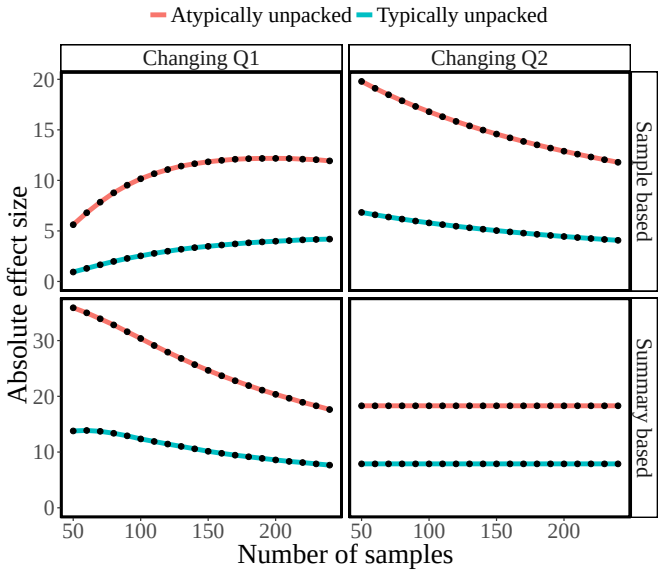


Figure 3.3: Simulation of subadditivity and superadditivity effects under sample-based (top) and summary-based (bottom) amortization strategies. In all panels, the y-axis represents the unstandardized effect size for $\mathcal{Q}2$. Left panels show the effects of changing the sample size for $\mathcal{Q}1$; right panels show the effects of changing the sample size for $\mathcal{Q}2$. When sample size for one query is changed, sample size for the other query is held fixed at 230 the sample size estimated by [56].

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 $\mathcal{Q}2$ ("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 exactly would involve their storage and is very intensive in its use of memory – in this aspect they are similar to exemplar models of categorization [261,244]. 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 predictions about how subadditivity and superadditivity change as a function of the sample size (Figure 3.3, details of these implementations can be found in the Appendix). For sample-based amortization, as the sample size for $Q1$ grows, the effect for $Q2$ asymptotically *diminishes* and eventually vanishes as the effect of biased initialization in $Q1$ washes out. However, initially increasing the sample size for $Q1$ also *amplifies* the effects for $Q2$ under a sample-based scheme, because this leads to more biased $Q1$ 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 in[56]. This effect can be counteracted by increasing the sample size for $Q2$. These are unbiased samples, since $Q2$ 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 $Q1$ will only *diminish* the effects for $Q2$, because the bias from $Q1$ 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 $Q2$ does not influence the effect size because the initialization of the chain for $Q2$ is not influenced by the samples or summary statistic from the answer to $Q1$. Under the summary-based strategy, the subadditivity and superadditivity effects for $Q2$ derive entirely from the same effects for $Q1$, which themselves are driven by the initialization see[56].

We test the different predictions of these strategies by placing people under cognitive load during either $Q1$ or $Q2$ in Experiment 2, a manipulation that is expected to reduce the number of produced samples[56,357]. In this way, we can sample different parts of the curves shown in Figure 3.3.

3.1.4 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

$$D_{\mathrm{KL}}(P||Q) = \sum_h P(h) \log \frac{P(h)}{Q(h)}. \tag{3.5}$$

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.[*] 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.

[*]Our 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.

3.2 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 elicited[56]. 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 the first query.

3.2.1 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 work[56] as the sub- and superadditivity effects might be weaker for the amortized answers to the second query.

3.2.2 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,* where the stimuli on every trial corresponded to the rows in Table 3.2. On each trial, participants first answered $Q1$ 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

*Note 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.

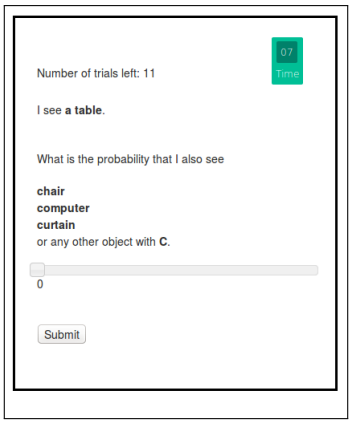


Figure 3.4: Experimental setup. Participants were asked to estimate the conditional probability using a slider bar within a 20-second time limit.

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 3.4).

The $Q1$ framing (typical-unpacked, atypical-unpacked or packed) was chosen randomly. Participants then completed the same procedure for $Q2$ (immediately after $Q1$), conditional on the same cued object. The framing for $Q2$ was always packed and $Q2$ 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.

3.2.3 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 studies[56], we found both subadditivity and superadditivity effects for $Q1$, depending on the unpacking: probability judgments were higher for unpacked-typical queries

Table 3.2: Experimental stimuli and queries for Experiment 1.

Cue	$Q1$	$Q1$-Typical	$Q1$-Atypical	$Q2$
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

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 $= 12.60 \pm 1.25$, $t(610.49) = 10.083$, $p < 0.0001$) and superadditivity effect (difference $= -15.69 \pm 1.32$, $t(615.46) = -11.89, p < 0.0001$).

Additionally, we found evidence that participants reused calculations from $Q1$ for $Q2$: even though all $Q2$ queries were presented in the same format (as packed), the estimates for that query differed depending on how $Q1$ was presented. In particular, estimates for $Q2$ were lower when $Q1$ 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 see[56].*

We calculated the difference between each participant's response for every $\mathcal{Q}2$ and the mean response for the same object averaged over all responses to $\mathcal{Q}2$ conditional on $\mathcal{Q}1$ 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 $= 4.39 \pm 1.14$, $t(606.40) = 3.83, p < 0.001$) and superadditivity effect (difference $= -7.86 \pm 1.21, t(610.41) = -6.50, p < 0.0001$).

We calculated each participant's mean response to all packed hypotheses for $\mathcal{Q}2$ 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 $\mathcal{Q}2$ 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 3.5.

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 also[56].

Next, we explored whether responses to $\mathcal{Q}1$ predicted trial-by-trial variation in responses to $\mathcal{Q}2$. Figure 3.6 shows the difference between participants' estimates for $\mathcal{Q}1$ and the true underlying probability of the query (as derived by letting our MCMC model run until convergence) plotted against the same difference for $\mathcal{Q}2$.† If participants do indeed reuse computations, then how much their estimates deviate from the underlying truth for $\mathcal{Q}1$ should be predictive of the deviance of their estimates for $\mathcal{Q}2$.

*The 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).

†Although 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.

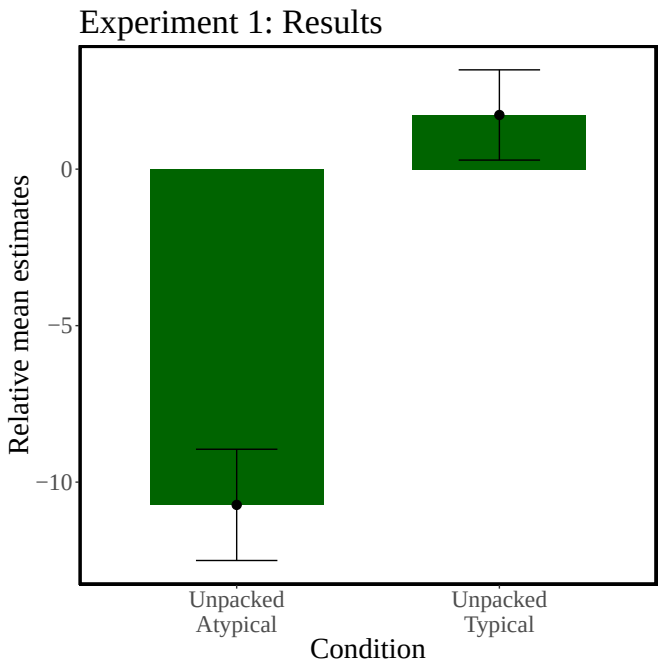


Figure 3.5: Experiment 1: Differences between $\mathcal{Q}2$ responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

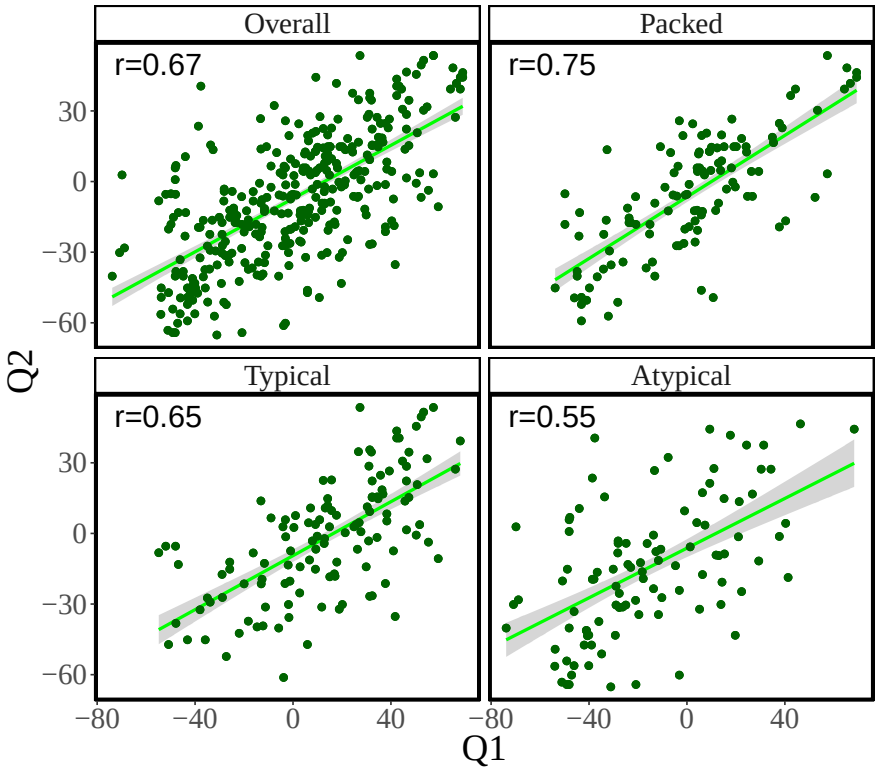


Figure 3.6: Trial-by-trial analyses of Experiment 1. Difference between $\mathcal{Q}1$ responses and true probability (as assessed by our MCMC model) plotted against the same quantity for $\mathcal{Q}2$. Lines show the least-squares fit with standard error bands.

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 o ($r = 0.55, p < 0.01$). Moreover, the within-participant effect size (the response difference between the unpacked conditions and the packed condition) for $\mathcal{Q}1$ was correlated with responses to $\mathcal{Q}2$ 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 $\mathcal{Q}1$ also showed correspondingly greater effects for $\mathcal{Q}2$.

3.2.4 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 $Q1$ carried over to $Q2$, and participants showing stronger effect sizes for both sub- and superadditivity for $Q1$ also showed greater effects for $Q2$.

3.3 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 3.3, manipulating cognitive load at $Q1$ and $Q2$ in order to exogenously control the number of samples see[357,56] for a similar approach.

In addition to cognitive load, we manipulate the "overlap" of $Q1$ with $Q2$, 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.

3.3.1 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.

3.3.2 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 $Q2$

Table 3.3: Experimental stimuli and queries for Experiment 2.

Cue	$Q1$	$Q1$-Typical	$Q1$-Atypical	$Q2$ Partial overlap	$Q2$ 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

queries to the list shown in Table 3.2. These $Q2$ queries have no overlap with the queried hypothesis for $Q1$ (for example, 'T or R' instead of 'C or R' in the trial shown in the first row in Table 3.2). In other words, these queries could not be decomposed such that the biased samples from $Q1$ would be reflected in the answer to $Q2$, so the sub- and super-additive effects would not be seen to carry over to $Q2$ 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 3.3.

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 $Q1$ and on half at $Q2$. 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 $\mathcal{Q}2$ should increase the amortization effect: if more samples are generated during $\mathcal{Q}1$ (which are the samples that contain the sub- or superadditivity biases) and these samples are concatenated with fewer unbiased samples during $\mathcal{Q}2$, then the combined samples will be dominated by biased samples from $\mathcal{Q}1$ and therefore show stronger effects. Vice versa, if participants are utilizing summary-based amortization, then cognitive load during $\mathcal{Q}1$ should increase the amortization effect: if less samples are generated during $\mathcal{Q}1$, 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).

3.3.3 Results

Analyzing only the queries with partial overlap (averaging across load conditions), we found that probability judgments for $\mathcal{Q}1$ 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 $\mathcal{Q}2$ (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 $\mathcal{Q}1$ 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 $= 13.64 \pm 1.57$, $t(396.95) = 8.70, p < 0.0001$) and superadditivity effect (-14.90 ± 1.56, $t(395.48) = -9.55, p < 0.0001$). Afterwards, we repeated the same analysis for responses to $\mathcal{Q}2$ (as in Experiment 1). This analysis again showed significant indicators of amortization as both the subadditivity (difference $= 5.37 \pm 1.34$, $t(398.01) = 4.02$, $p < 0.001$) and the superadditivity effect (difference $= -4.92 \pm 1.34$, $t(398.01) = -3.69, p <$

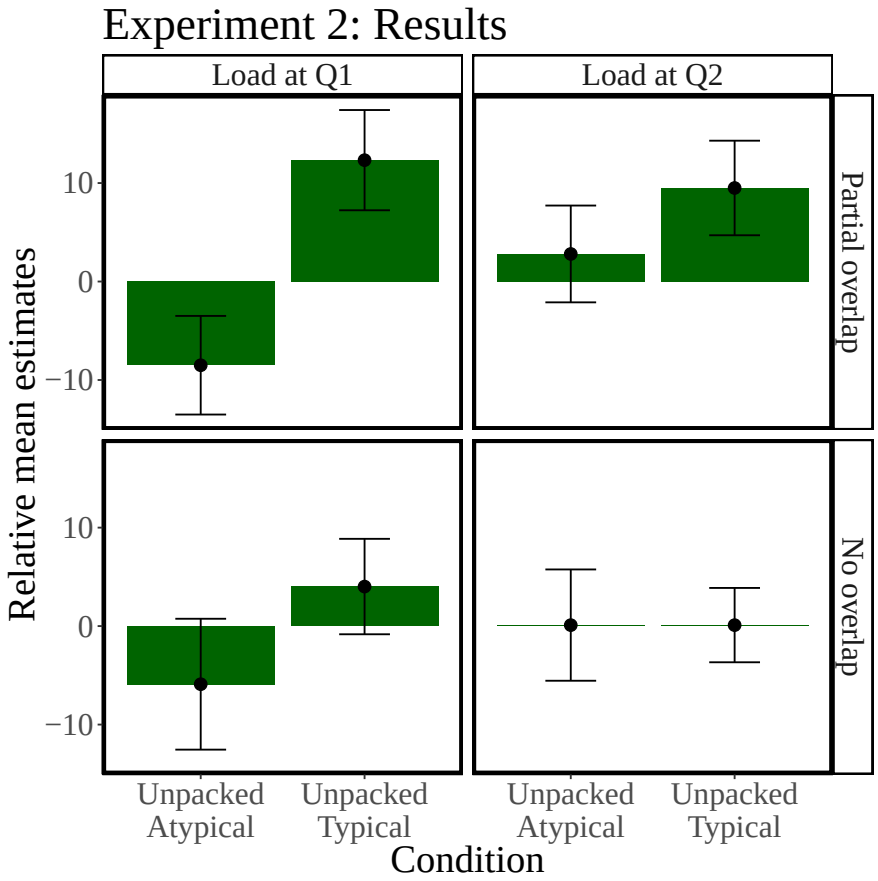


Figure 3.7: Experiment 2: Differences between $\mathcal{Q}2$ responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

0.001) were still present during $\mathcal{Q}2$.

Next, we assessed how the memory effect was modulated by cognitive load and overlap (Figure 3.7). When cognitive load occurred during $\mathcal{Q}2$ and there was no overlap, none of the conditions produced an effect significantly different from 0 (all $p > 0.5$). When cognitive load occurred during $\mathcal{Q}2$ 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 $\mathcal{Q}1$ 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 $\mathcal{Q}1$ and there was partial overlap, both conditions showed the expected subad-

ditive ($t(38) = 4.18, p < 0.05$) and superadditive ($t(46) = -1.89, p < 0.05$) effects. Moreover, calculating the average effect size of amortization for the different quadrants of Figure 3.7, the partial overlap-cognitive load at $Q1$ condition produced the highest overall effect ($d = 0.8$), followed by the partial overlap-cognitive load at $Q2$ condition ($d = 0.56$) and the no overlap-cognitive load at $Q1$ condition ($d = 0.42$). The no overlap-cognitive load at $Q2$ condition did not produce an effect greater than 0. Partial overlap trials were also more strongly correlated with responses during $Q1$ 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 $Q2$ and the mean responses to $Q2$ over queried objects provided that $Q1$ was packed. This difference was entered into a linear mixed-effects regression that contained overlap, cognitive load, and the presentation format of $Q1$ 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 $= 5.25 \pm 2.12$, $t(417.08) = 2.48\, p < 0.05$) but no superadditivity effect (difference $= -3.19 \pm 2.17$, $t(419.23) = -1.47, p = 0.17$) when cognitive load was applied during $Q2$. Importantly, both the subadditivity (difference $= 5.83 \pm 2.25$, $t(418.91) = 2.59, p < 0.05$) and the superadditivity (difference $= -6.86 \pm 2.21$, $t(419.80) = -3.102, p < 0.01$) effects were present when cognitive load was applied during $Q1$. This finding points towards a larger amortization effect in the presence of cognitive load on $Q1$, thus supporting a summary-based over a sampled-based amortization scheme.

Further, on trials with cognitive load at $Q2$, 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 $Q2$ if the computations they executed for $Q1$ are reusable in answering $Q2$. This also indicates that these results cannot be explained by simply initializing the chain for $Q2$ where the chain for $Q1$ ended, which would not have affected the required computations.

Interestingly, there was no evidence for a significant difference between participants' responses to

$\mathcal{Q}2$ under cognitive load in Experiment 2 as compared to participants' responses to $\mathcal{Q}2$ in Experiment 1 when no cognitive load during either $\mathcal{Q}1$ or $\mathcal{Q}2$ was applied ($t(314) = -1.44, p = 0.15$).

Finally, we assessed how much the difference between responses for $\mathcal{Q}1$ and the true underlying probabilities were predictive of the difference between responses for $\mathcal{Q}2$ and the true underlying probabilities (Figure 3.8). There was a strong correlation between responses to $\mathcal{Q}1$ and $\mathcal{Q}2$ 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 $\mathcal{Q}1$ and $\mathcal{Q}2$ 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 $\mathcal{Q}1$ also showed stronger effects for $\mathcal{Q}2$ 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.

3.3.4 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 3.3), we argue that the effect of cognitive load at $\mathcal{Q}1$ on $\mathcal{Q}2$ 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[120,145,221]. Further supporting this idea is our finding that performance on the secondary task was better in the partial overlap conditions, 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

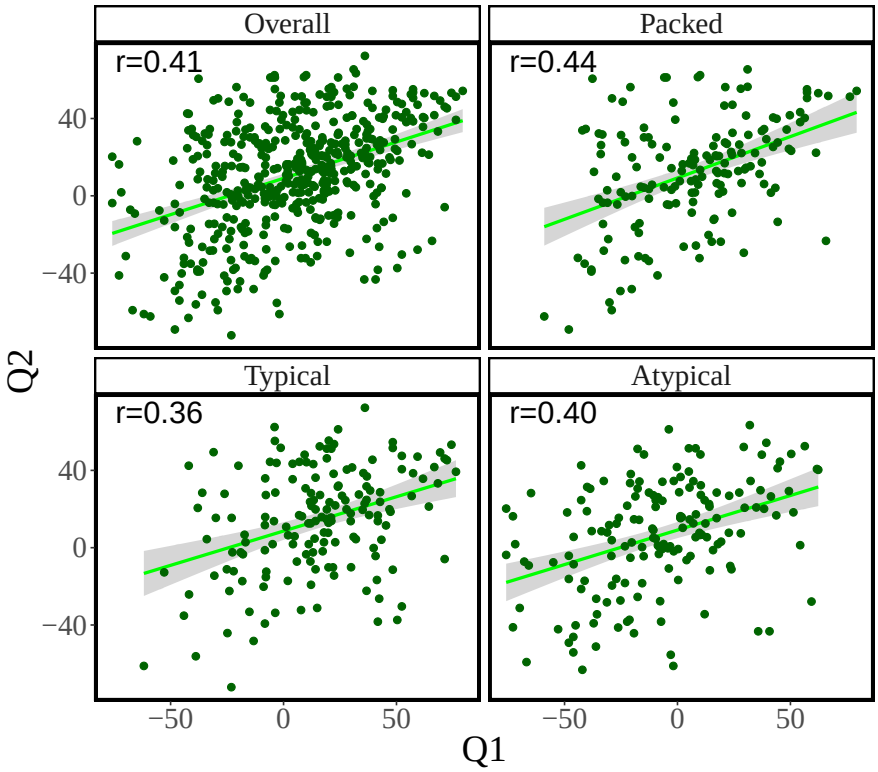


Figure 3.8: Trial-by-trial analyses of Experiment 2. Relationship between difference between $\mathcal{Q}1$ responses and true probability (as assessed by our MCMC model) and $\mathcal{Q}2$ responses and true probability. Lines show the least-squares fit with standard error bands.

hypothesis, the memory effect was stronger in the high overlap condition.

3.4 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 $Q1$ and $Q2$ 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 $Q1$ are always "reusable" in $Q2$ 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 $Q1$ not being reflected in $Q2$ 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).

3.4.1 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 time response.

3.4.2 Procedure

The procedure was similar to Experiments 1 and 2. The only difference was that participants were shown a new cue word for $Q2$, asking them to judge the probability of objects starting with the same letter as the letter from $Q1$ with no conjunction of letters (i.e., same query space, full overlap). The query for $Q2$ was always packed, as in previous experiments. The new cue words for $Q2$ were generated to either have posterior with a low (similar cues) or a high (dissimilar cues) KL divergence from the $Q1$ posterior. The range of KL divergences fell between 0 and 9; all similar cue words had conditional

Table 3.4: Experimental stimuli and queries for Experiment 3. Kullback-Leibler (KL) divergence between the posteriors for $\mathcal{Q}1$ and $\mathcal{Q}2$ are shown in parentheses.

Cue1	$\mathcal{Q}1$	$\mathcal{Q}1$-Typical	$\mathcal{Q}1$-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)

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 3.4.

3.4.3 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 $\mathcal{Q}1$ in the typically unpacked queries were higher than in the unpacked condition (subadditivity effect: $t(92) = 4.67, p < 0.001$) and that probability 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 $\mathcal{Q}2$, 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 $\mathcal{Q}1$ 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

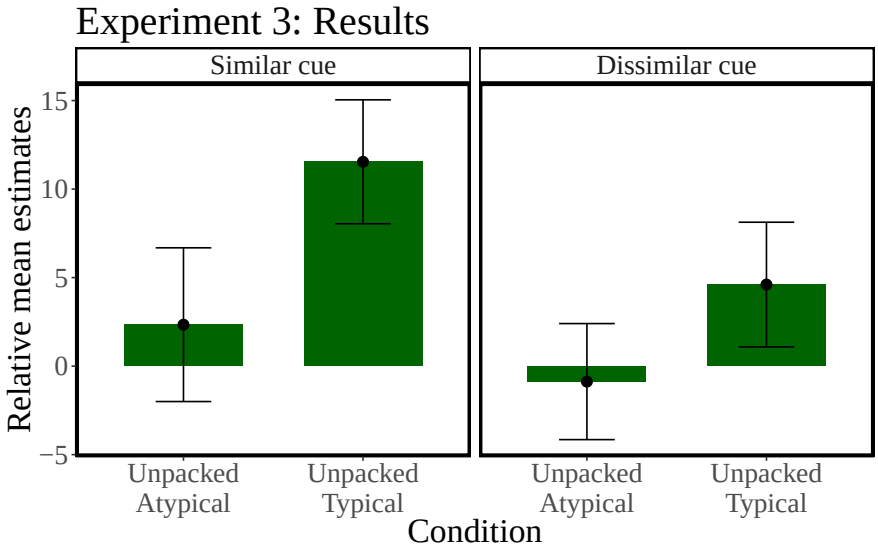


Figure 3.9: Experiment 3: Differences between $\mathcal{Q}2$ responses for each condition and an average packed baseline. A negative relative mean estimate indicates a superadditivity effect and a positive relative mean estimate a subadditivity effect. Error bars represent the standard error of the mean.

analysis for responses to $\mathcal{Q}2$ 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 superaditivity ($t(47) = -0.27$, $p = 0.79$). However, for the similar cues at $\mathcal{Q}2$, 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 $\mathcal{Q}1$ and $\mathcal{Q}2$ ($r = 0.34, p < 0.001$), 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 3.10. Participants who showed higher subadditivity or superadditivity effects for $\mathcal{Q}1$ also showed higher effects for $\mathcal{Q}2$

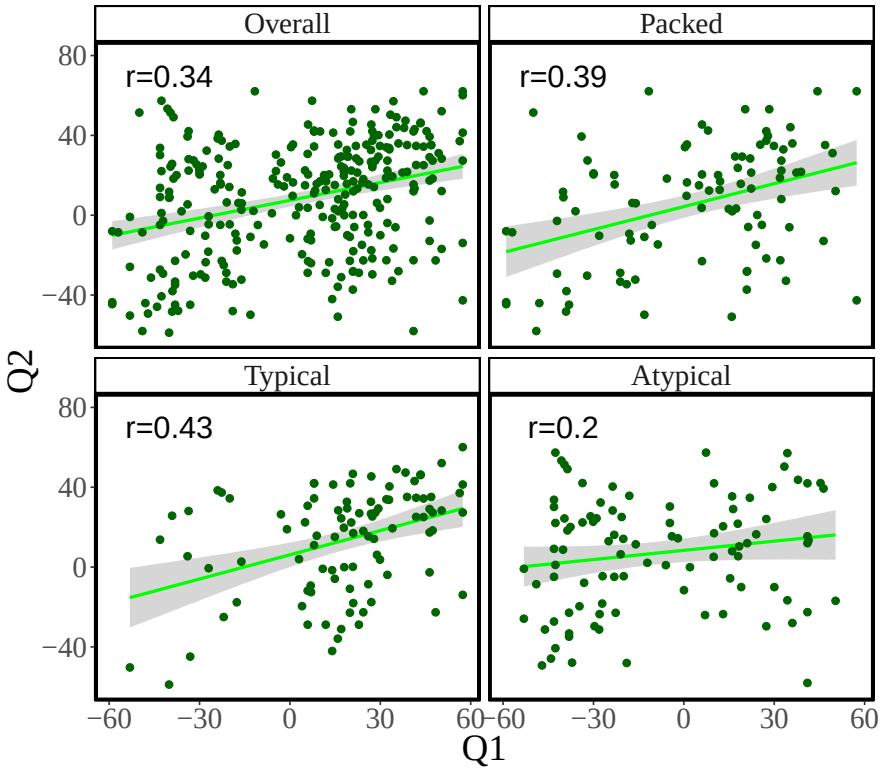


Figure 3.10: Trial-by-trial analyses of Experiment 3. Relationship between difference between $Q1$ responses and true probability (as assessed by our MCMC model) and $Q2$ responses and true probability. Lines show the least-squares fit with standard error bands.

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 ($r = 0.11, p = 0.29$).

3.4.4 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.

An unexpected finding was that while the superadditivity effect in aytpically-unpacked $Q1$ was significant, neither the memory-based superadditivity effect (in $Q2$) nor correlations across the queries

for atypically-unpacked $Q1$ were significant. This indicates that the answers to the atypically-unpacked $Q1$ are not detectably being reused in $Q2$ 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.

3.5 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 replicating[56], 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 results[117], 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[218,376,145,120]. 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.

3.5.1 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[164,160,161,159]. For example, Heit & Hayes[164] 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 decision about a noisy stimulus (whether dots are moving to the left or right of a reference) influences a subsequent continuous judgment about motion direction[180]. Stocker and colleagues[351,228] 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. Fleming & Daw[101] 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[220,125,218,376]. 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[382,383]. The

key idea, derived from a quantum probability model of cognition see also [363], 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 [382] 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, Costello & Watts [55] 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 [232]. Costello & Watts [55] 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 Costello & Watts [55].

Most closely related to the present paper is the work of Dougherty and colleagues [72,358,359,77,76], 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 Dasgupta et al. [56], 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 Dasgupta et al. [56], 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 [62,61,203]. "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 algorithms [194], thus occupying a middle ground in the time-space trade-off.

3.5.2 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., [352,298] 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 treatment 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 3.3, *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 ($\mathcal{Q}1$) to the sample set for another query ($\mathcal{Q}2$). So if N_1 samples were generated in response to $\mathcal{Q}1$, and N_2 new samples are generated in response to $\mathcal{Q}2$, in the absence of amortization, the responses to the two queries $\mathcal{Q}1$ and $\mathcal{Q}2$ would be generated as follows:

$$P(h_{\mathcal{Q}1}|d) \approx \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbb{I}[h_n = h_{\mathcal{Q}1}]$$
$$P(h_{\mathcal{Q}2}|d) \approx \frac{1}{N_2}\sum_{n=1}^{N_2}\mathbb{I}[h_n = h_{\mathcal{Q}2}]$$

Under the sample-based amortization scheme, the response to $\mathcal{Q}2$ is given by a calculation carried out over all $N_1 + N_2$ equally weighted samples.

$$P(h_{\mathcal{Q}1}|d) \approx \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbb{I}[h_n = h_{\mathcal{Q}1}]$$
$$P(h_{\mathcal{Q}2}|d) \approx \frac{1}{N_2+N_1}\left(\sum_{n=1}^{N_1}\mathbb{I}[h_n = h_{\mathcal{Q}2}] + \sum_{n=1}^{N_2}\mathbb{I}[h_n = h_{\mathcal{Q}2}]\right)$$

Under this scheme, all the computations carried out for $\mathcal{Q}1$ are available for flexible reuse in the computation for $\mathcal{Q}2$.

In *summary-based amortization*, we reuse a summary statistic computed from $\mathcal{Q}1$. This strategy is only applicable to problems where the answer to $\mathcal{Q}2$ can be expressed as the composition of the answer to $\mathcal{Q}1$, and an additional simpler computation. For example if $\mathcal{Q}1$ is "What is the probability that there is an object starting with a C in the scene?", $\mathcal{Q}2$ 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 $\mathcal{Q}1$ 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 $\mathcal{Q}2$ ("the probability of an object starting with C or R").

$$P(h_{\mathcal{Q}1}|d) \approx \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbb{I}[h_n = h_{\mathcal{Q}1}]$$
$$P(h_{\mathcal{Q}2}|d) \approx \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbb{I}[h_n = h_{\mathcal{Q}1}] + \frac{1}{N_2}\sum_{n=1}^{N_2}\mathbb{I}[h_n = (h_{\mathcal{Q}2} - h_{\mathcal{Q}1})]$$

Under this scheme, only the final product of the computation carried out for $\mathcal{Q}1$ is reused in the

calculations for $Q2$.

4

Amortization gives rise to ecologically rational heuristics

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.

4.1 Introduction

Studies of probabilistic reasoning frequently portray people as prone to errors[366,341,142,99]. The cognitive processes that produce these errors is the subject of considerable debate[245,128,313]. One influential class of models holds that rational probabilistic reasoning is too cognitively burdensome for people, who instead use a variety of heuristics[366,132,324]. Alternatively, rational process models hold that errors arise from principled approximations of rational reasoning, for example some form of hypothesis sampling[57,314,150]. These different perspectives have some common ground; certain heuristics might be considered accurate approximations[130,274,18].

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*[8,24,142,190], the finding that people under-react to prior probabilities relative to Bayes' rule. Other studies report *conservatism*[288,284], the finding that people under-react to evidence.*

Heuristic models respond to this challenge by allowing heuristics to be context-sensitive, an example of *strategy selection*[129,235]. Most models of strategy selection assume that people are able to assess the

*We 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.

usefulness of a strategy, through cost-benefit analysis[181,14,216], reinforcement learning[84,299], or based on the strategy's applicability in a particular domain[236,319]. 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?"[144] 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[247] for a similar argument. Some models explain why certain kinds of queries induce certain kinds of errors[57], but do not explain how errors can be modulated by other queries in the same context[117,58].

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*[65,196] that maps queries to posterior probabilities.* 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.† Importantly, this optimization is carried out without explicit feedback about the true posterior[251].

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

*When the recognition model is parametrized as a neural network, it is sometimes also referred to as an *inference network*[251,297,272].

†We 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.

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 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.

4.2 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:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}, \tag{4.1}$$

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 [288,82], 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 pro-

portion 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[285,284,286]. 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[20]. 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 $\pi(h|d)$:

$$\pi(h|d) = \frac{\pi(d|h)P(h)}{\sum_{h'} \pi(d|h')P(h')} \tag{4.2}$$

where biases in the posterior are driven by biases in the subjective sampling distribution $\pi(d|h)$. To accommodate the existence of under-reaction, subjects would need to assume subjective sampling distributions that were flatter (more dispersed) than the objective distributions. Edwards[82] proposed

that the subjective sampling distribution could be modeled as:

$$\pi(d|h) = \frac{[P(d|h)]^{\omega}}{\sum_{d'}[P(d'|h)]^{\omega}}. \tag{4.3}$$

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[283,385]. 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.,[151].

These weaknesses of the biased sampling distribution hypothesis motivated the alternative hypothesis that subjects are systematically under-weighting the likelihood[288], what Edwards[82] referred to as "conservatism bias." This hypothesis can be formalized using a generalized version of Bayes' rule:

$$P(h|d) \propto \frac{[P(d|h)]^{\gamma}P(h)}{\sum_{h'}[P(d|h')]^{\gamma}P(h')}, \tag{4.4}$$

where γ is a free parameter specifying the weighting of the likelihood. Note that this model is superficially similar to Edwards[82]'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 distribu-

tions 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 DuCharme[80], is a form of "extreme belief aversion" see also[20]. If subjects avoid reporting extreme beliefs, then large posterior odds will be pulled towards 0. Consistent with this hypothesis, DuCharme[80] 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.,[143,206,21].

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[190,189]. Instead of using laboratory-controlled scenarios involving bags filled with poker chips, Kahneman & Tversky[190] 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*.*

*Although base rate neglect was popularized by Kahneman and Tversky's work, it was in fact documented

Many subsequent studies have reported under-reaction to priors, though the interpretation of these studies has been the focus of vigorous debate see[201,11]. It has been observed in incentivized experiments e.g.,[142,107], in real-world markets[10], and in highly trained specialists such as clinicians[81] and psychologists[193].

In addition to establishing the empirical evidence for under-reaction to priors, Kahneman & Tversky[189] 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. 4.4 can be generalized to allow insensitivity to the prior[142]:

$$P(h|d) \propto \frac{[P(d|h)]^{\alpha_L} P(h)^{\alpha_P}}{\sum_{h'} [P(d|h')]^{\alpha_L} P(h')^{\alpha_P}}, \tag{4.5}$$

As before, $\alpha_L < 1$ implies insensitivity to the likelihood; in addition, $\alpha_P < 1$ implies insensitivity to the prior (base rate neglect). Grether[142] 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 Tenenbaum & Griffiths[356] 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"

earlier using the poker chip paradigm[288], but this observation was mostly ignored by subsequent research using that paradigm.

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 Democractic," even though the latter has a higher likelihood[367].

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. Gennaioli & Shleifer[111] 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 also[75]. 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[242,329]. Consistent with the diagnosticity hypothesis, Fischhoff & Bar-Hillel[98] showed greater under-reaction to the evidence when diagnosticity was higher see also[8,265]. However, a meta-analysis by Benjamin[20] 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 Peterson & Miller[282] and Edwards[82]. In particular, when the parameters of the model in Equation 4.5 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 completely[20]. 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 Massey & Wu[240]: 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.

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

4.3.1 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 constant[116]. For example, Monte Carlo algorithms[6] approximate the posterior using M weighted samples $\{h^1, \ldots, h^M\}$:

$$P(h|d) \approx \sum_{m=1}^{M} w^m \mathbb{I}[h^m = h], \tag{4.6}$$

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 $\tilde{P}(h)$, with weights given by $w^m = P(d|h^m)P(h^m)/\tilde{P}(h^m)$.

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[70,376,124]. Second, they can explain a wide range of inferential errors, ranging from subadditivity to the conjunction fallacy[314,57]. Third, they can be implemented in biologically plausible circuits with spiking neurons[42,269,155].

Monte Carlo algorithms can be thought of as procedures for generating an approximate posterior $Q_\varphi(h|d)$ parametrized by the set of weights and samples, $\varphi = \{w^m, h^m\}_{m=1}^M$. 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 $\varphi \in \Phi$) that gets "closest" to the true posterior,

$$\varphi^* = \operatorname*{argmin}_{\varphi \in \Phi} \mathcal{D}[Q_\varphi(h|d)||P(h|d)], \tag{4.7}$$

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 $M \to \infty$, they produce φ^*. It is however possible to design non-randomized algorithms that directly optimize φ[312] in a sample-based approximation. Such optimization is an example of *variational inference*[185], 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):

$$\mathcal{D}_{\mathrm{KL}}[Q_\varphi(h|d)||P(h|d)] = \sum_h Q_\varphi(h|d) \log \frac{Q_\varphi(h|d)}{P(h|d)}. \tag{4.8}$$

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[65,196,251,297,272]. 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; see[387]. For example, particular implementations of variational inference have been used to model hierarchical predictive coding in the brain[105,115].

4.3.2 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 engine 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[65,196,251,297,272,352,86,381,304,237].

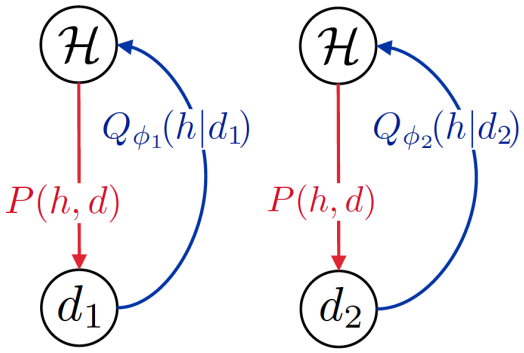
To formalize this idea, let the data variable d subsume not only the standard "observation" (e.g., symptoms in the diagnostic example) but also the information provided to the agent about the generative model $P(d, h)$ and the subset of the hypothesis space being queried (e.g., a particular diagnostic 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 4.1). 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 formalize this problem is to define it as an expectation under the query distribution $P_{\text{query}}(d)$:

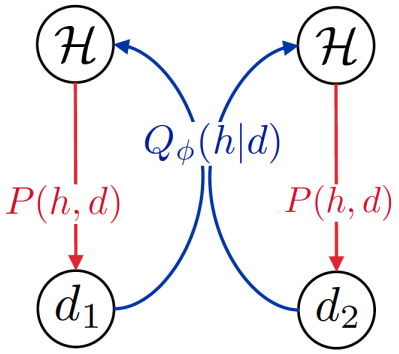
$$\varphi^* = \underset{\varphi \in \Phi}{\operatorname{argmin}} \, \mathbb{E}_{P_{\text{query}}} \left\{ \mathcal{D}[Q_\varphi(h|d) || P(h|d)] \right\} \tag{4.9}$$

Under this objective function, high probability queries will exert a stronger influence on the variational parameters (see Figure 4.2 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.

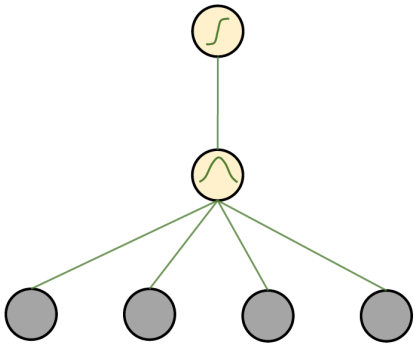
It is important to note that classical (non-amortized) approximate inference is a special case of amortized inference, and if there are no constraints on the amortization architecture, then the optimal ar-



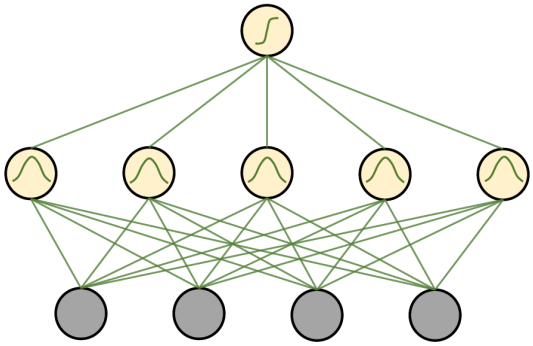
(a) Memoryless inference



(b) Amortized inference

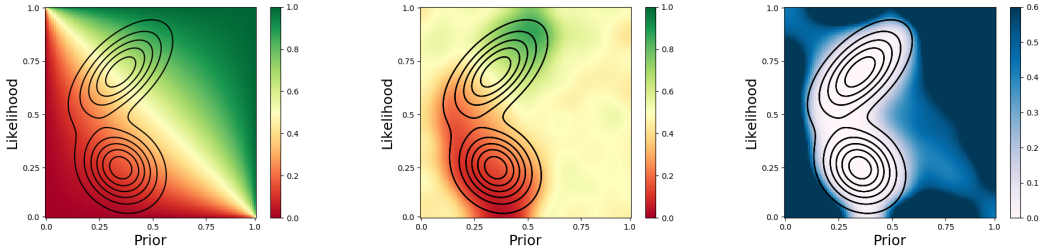


(c) Learned Inference Model with 1 hidden unit



(d) Learned Inference Model with 5 hidden units

Figure 4.1: 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_φ that is unique for each d. (B) Amortized inference allows some variational parameters to be shared across queries, optimizing them such that Q_φ 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).



(a) True posterior P **(b)** Approximate posterior Q **(c)** KL$[Q||P]$

Figure 4.2: 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 $h \sim$ Bernoulli(p_0) and $d|h \sim$ 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.

chitecture 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[93,4], which will produce the kinds of inferential errors we wish to explain.

4.3.3 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 4.1 C-D, further details can be found in Appendix A). The inputs are all the relevant information about the query subsumed by the data variable d, and the outputs uniquely determine an approximate distribution $Q_\varphi(h|d)$ over all hypotheses h. For example, if we want to model the posterior distribution $P(h|d)$ 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 $Q_\varphi(h|d)$ for all queries d (i.e., the approximation is amortized; Figure 4.1 A-B). The network encounters a series of queries d and outputs a guess for $Q_\varphi(h|d)$. This guess is improved in response to each new d, with updates to the network parameters φ. This leads to query dependence (Figure 4.2) in the learned parameters φ, and therefore in the approximation $Q_\varphi(h|d)$. The updates to φ are made using an algorithm that performs that performs the optimization in Equation 4.9 using knowledge only of the joint distribution as a learning signal [290] , 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 based process models for probability judgment for example: [75,328,58,349,166] 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 Orhan & Ma [270], 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)*. Such convergence has played an important role in theorizing about other forms of cognitive bottlenecks e.g., [93,4]. 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 dis-

*In 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.

cuss later, the model can be naturally integrated with Monte Carlo sampling accounts of approximate inference.

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.

4.4 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.

Benjamin [20] 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 Benjamin [20], 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, [20] follows [142] in allowing separate parameters for sensitivity to the likelihood and the prior (Eq. 4.5). For analytical convenience, this model can be reformulated as linear in log-odds:

$$\log \frac{P(h = R|d)}{P(h = B|d)} = \alpha_P \log \frac{P(h = R)}{P(h = B)} + \alpha_L \log \frac{P(d|h = R)}{P(d|h = B)} + \varepsilon, \tag{4.10}$$

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). [20] 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 θ) 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 (θ) 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 experimenters*. We found that the Learned Inference Model could

*Crucially 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.

successfully reproduce the 3 main findings from the Benjamin [20] meta-analysis (Figure ??).

We also applied the model to experiments in which the prior distribution was non-uniform (deviated substantially from 0.5). Figure 4.4 shows data aggregated by Benjamin [20] 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 Benjamin [20] 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 see [240] 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. Degeneracy 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 4.5A, left). Estimated sensitivity to the likelihood ($\hat{\alpha}_L$) across all diagnosticities and sample sizes was very close to the Bayesian optimal of 1 (Figure 4.5A middle and right). We also found that higher capacity mostly abolished base rate neglect (Figure 4.4C).

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

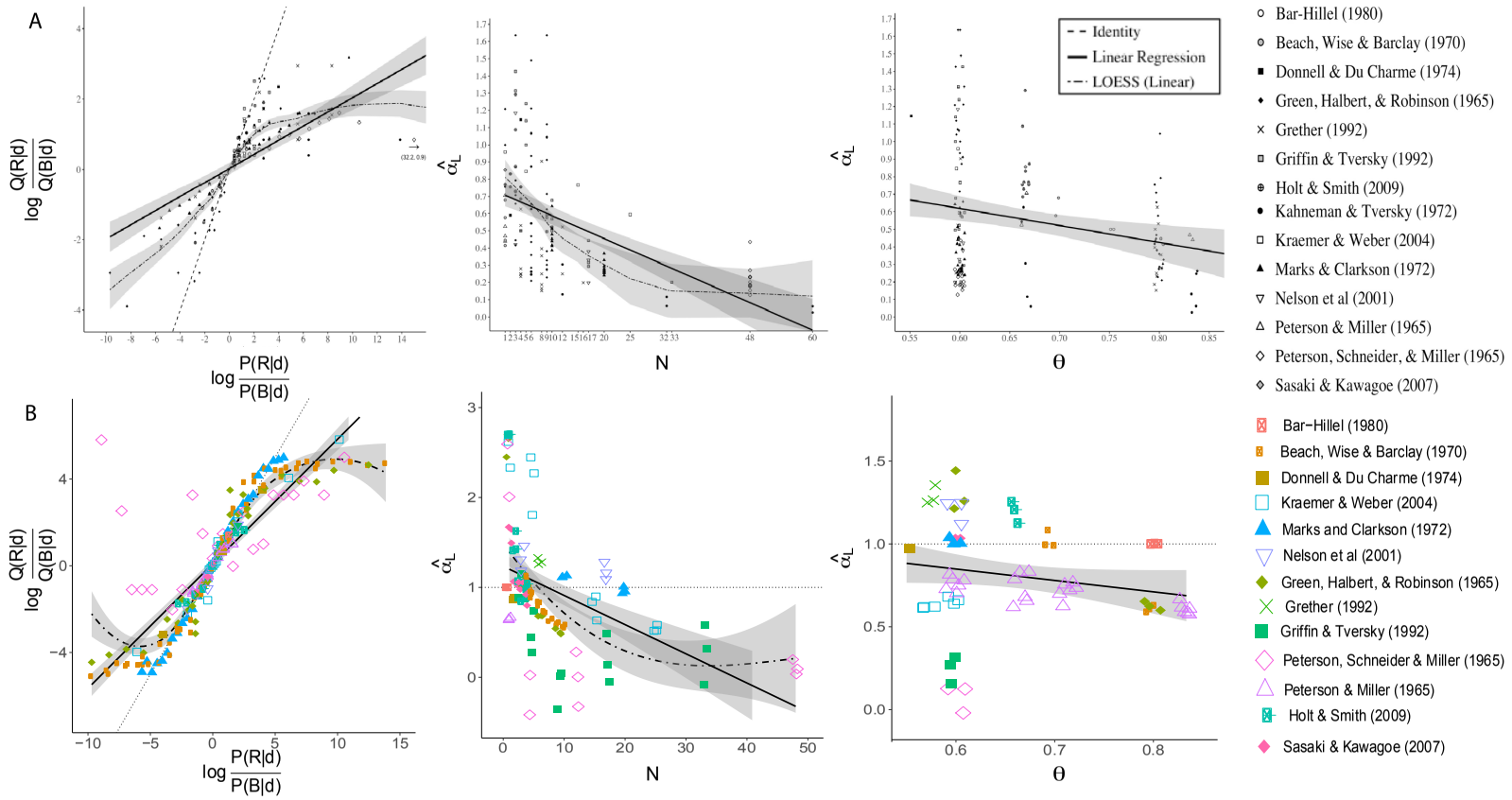


Figure 4.3: Simulation of inferential errors in binary symmetric problems with uniform priors. $P(h|d)$ represents true posterior probabilities, $Q(h|d)$ represents subjective posterior probabilities. (A) Data aggregated by[20]. (B) Learned Inference Model simulations. Left: subjective posterior log-odds 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 θ. The shaded curves show the linear and nonlinear (LOESS) regression functions with 95% confidence bands

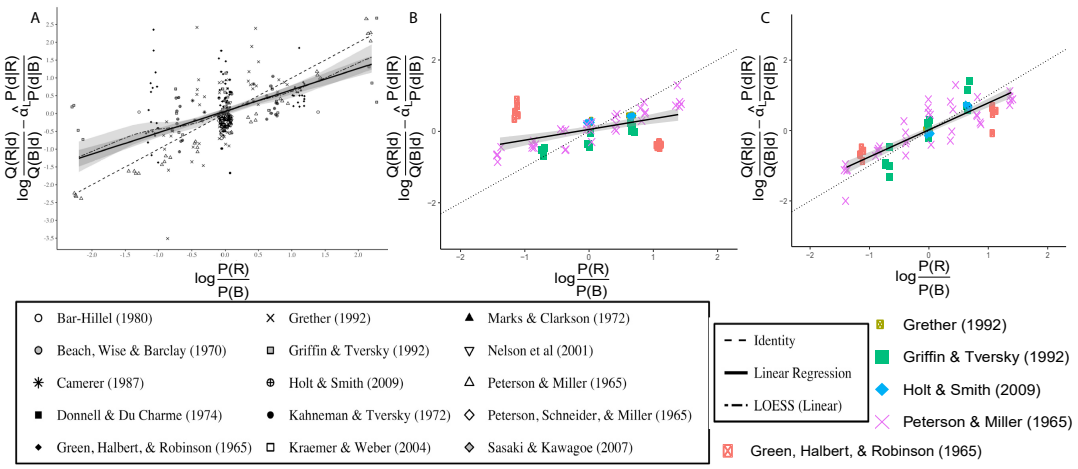


Figure 4.4: Simulation of inferential errors in binary symmetric problems with non-uniform priors. $P(h|d)$ represents true posterior probabilities, $Q(h|d)$ represents subjective posterior probabilities. Plots show prior log-odds on the x axis, and the subjective prior log-odds calculated as the subjective posterior log-odds adjusted for subjective response to the likelihood (as modulated by $\hat{\alpha}_L$). (A) Data aggregated by[20]. (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.

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 4.5B (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 4.5B, middle and right, where the estimated sensitivity $\hat{\alpha}_L$ is closer to zero.

4.4.1 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 by[143], who suggested a specific decomposition of the posterior log-odds into the *strength*

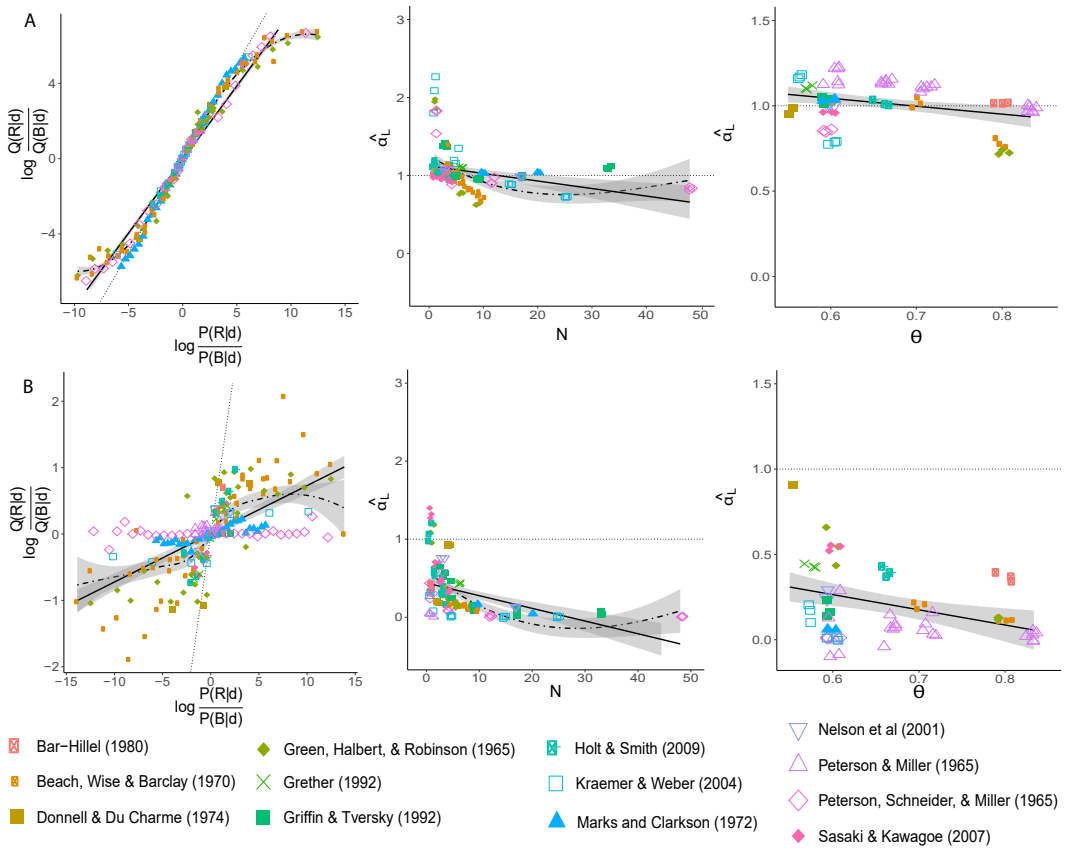


Figure 4.5: Simulations of inferential errors with high capacity and a biased query distribution. $P(h|d)$ represents true posterior probabilities, $Q(h|d)$ represents subjective posterior probabilities. (A) Simulation of high-capacity (8 hidden units) Learned Inference Model. (B) Simulation of low-capacity (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.

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 4.1: Stimuli used in [143].

(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 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 4.1), 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 $h = A$) or that it favors tails (denoted $h = B$). The prior probabilities of both hypotheses were equal. The symmetric binomial probability was fixed at $\theta = 3/5$, 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:

$$\log \frac{P(h = A|d)}{P(h = B|d)} = N \left(\frac{N_a - N_b}{N} \right) \log \left(\frac{\theta}{1 - \theta} \right), \tag{4.11}$$

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 Grether[142], Griffin & Tversky[143] 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 $\alpha_0 = \log \log \left(\frac{\theta}{1-\theta} \right)$, and allowed for random response error ε, arriving at the following regression model:

$$\log \left(\log \frac{P(h = A|d)}{P(h = B|d)} \right) = \alpha_0 + \alpha_W \log(N) + \alpha_S \log \left(\frac{N_a - N_b}{N} \right) + \varepsilon. \tag{4.12}$$

The Bayes-optimal parametrization is $\alpha_W = \alpha_S = 1$. However, Griffin & Tversky[143] 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 the set of stimuli used in the original experiment (Table 4.1), 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, ans 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 4.1.

Consistent with the experimental results, we found that the model was sub-optimally sensitive to both sources of information (Figure 4.6), with both $\hat{\alpha}_S$ and $\hat{\alpha}_W$ being less than 1. We also found that

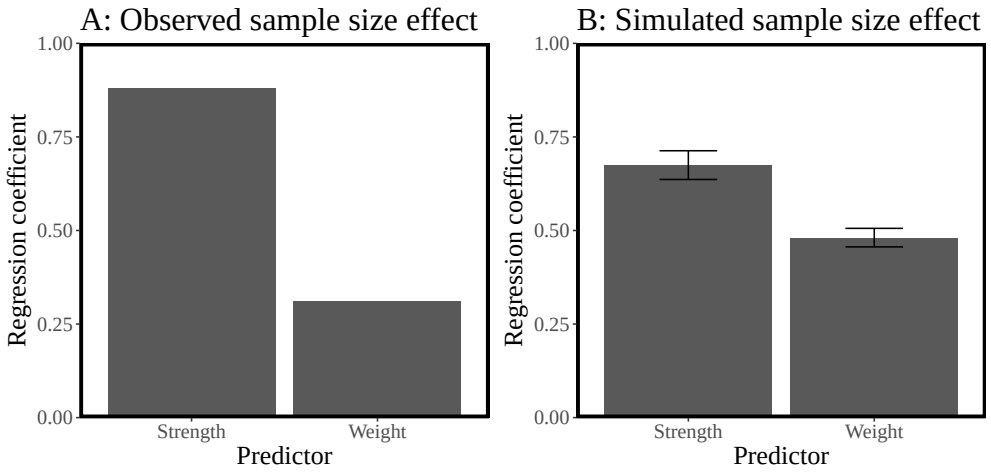


Figure 4.6: Strength and weight in probabilistic judgment. (A) Regression coefficients reported in [143]. (B) Regression coefficients estimated from simulations of the Learned Inference Model. Error bars represent the standard error of the mean.

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 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 4.7A). 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 4.7B), 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.

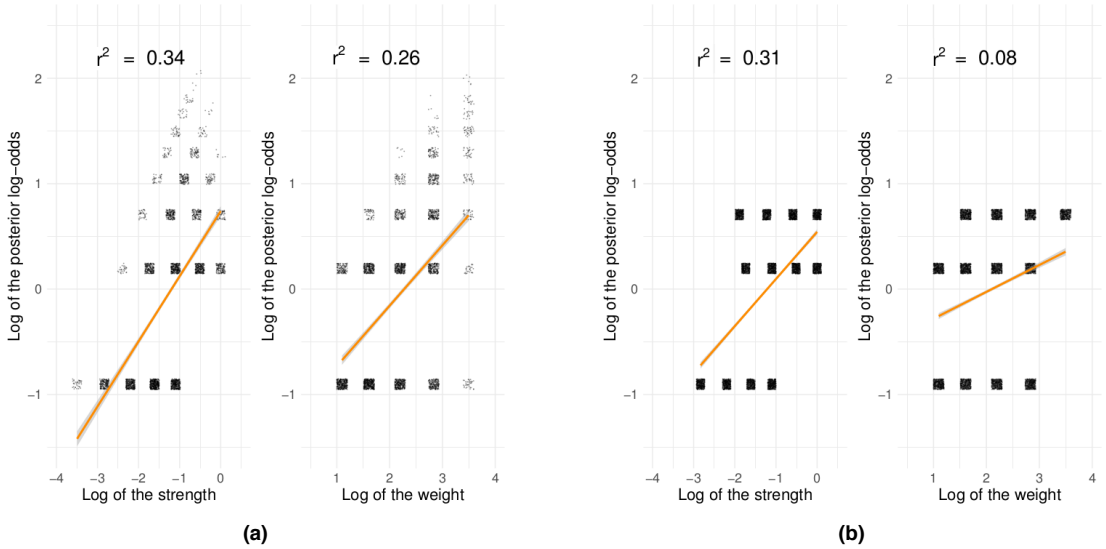


Figure 4.7: Variance explained by strength and weight independently. These plots show regressions between the log of the strength or weight of the evidence against the log of the posterior log-odds. (A) For samples drawn from the true generative process, the strength explains more variance in the posterior. (B) For the stimuli used in [143], 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.

4.4.2 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.

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 4.8). 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 4.8A); the other group went through 8 blocks of informative likelihoods and uninformative priors (Figure 4.8B). 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.

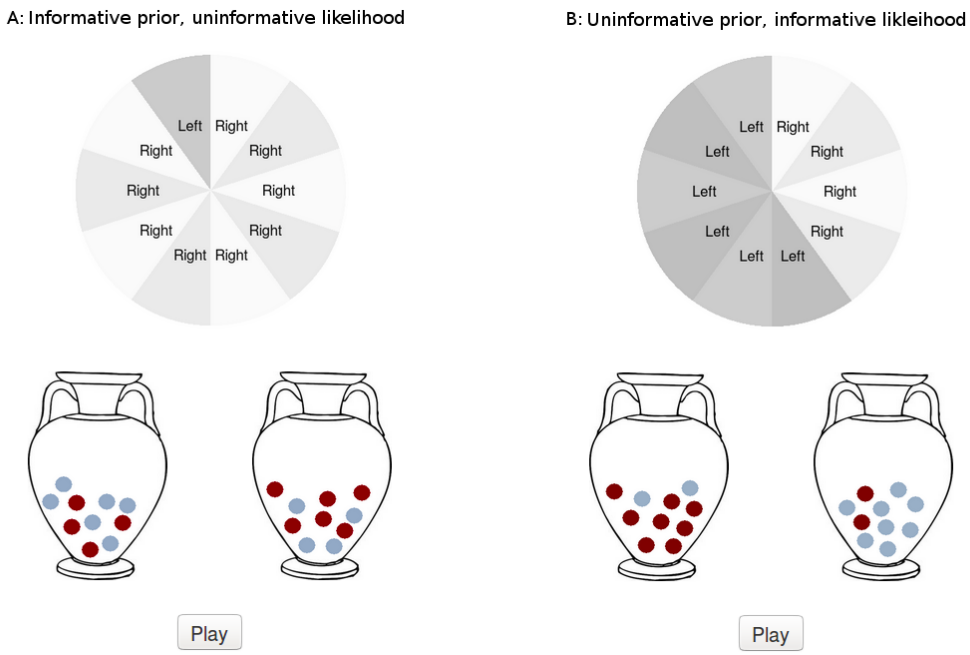


Figure 4.8: Screen shots of urn experiment. (A) In the condition with informative priors and uninformative likelihoods, the wheel of fortune had urn probabilities of $0.7, 0.8$, or 0.9. The proportions of blue balls in the urns was 0.5 or 0.6. (B) In the condition with uninformative priors and informative 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.

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. 4.10. Thus, we entered the log-odds of the prior, the log-odds of the likelihood, the condition (coded as '0' for the informative prior condition, and '1' 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 4.9A). 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 4.9B). 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. $\hat{\alpha}_L = 0.92, p < .001$).

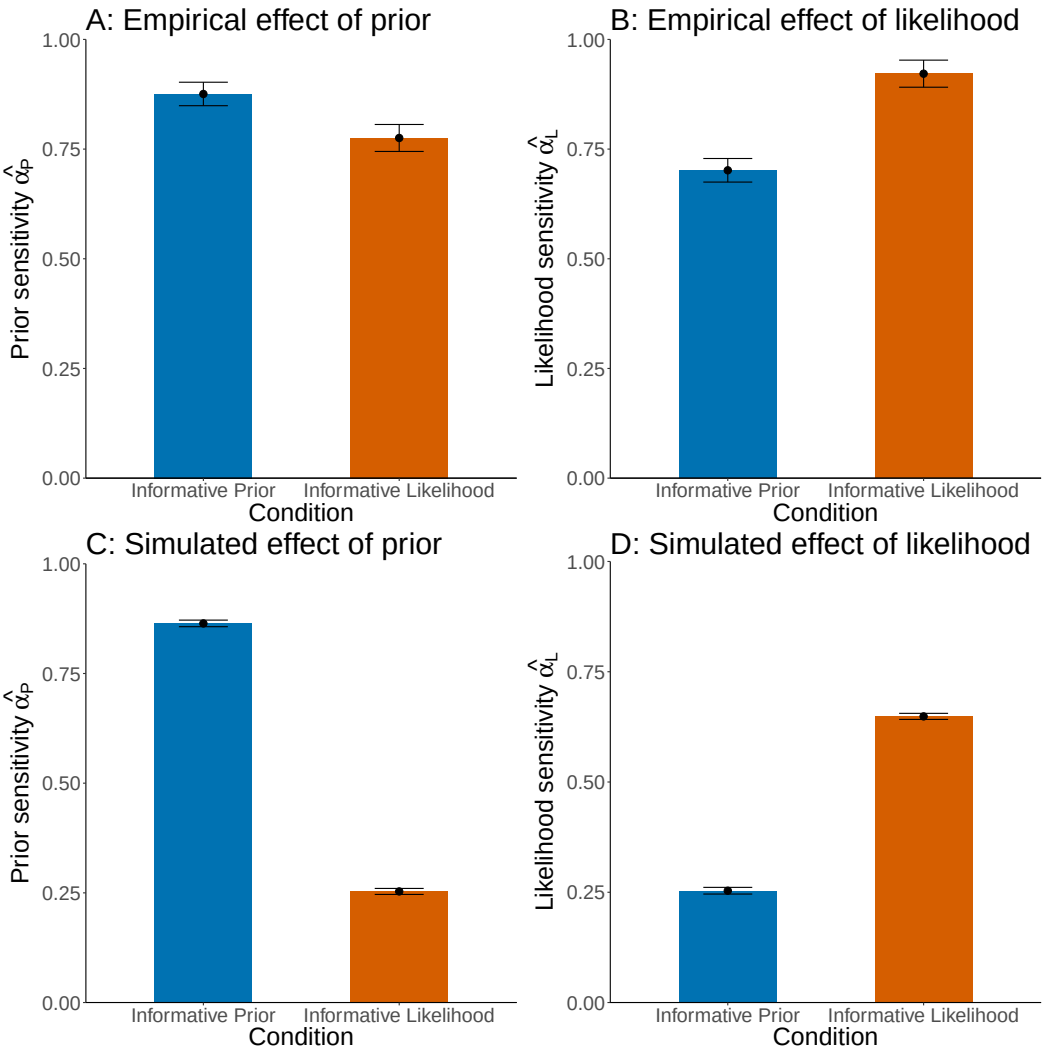


Figure 4.9: Results of urn experiment. The y-axis shows estimates for the regression coefficients α_L and α_P (see Equation 4.10), and the x-axis represents the experimental condition. (A) Subjects weighted the prior more in the informative prior than in the informative likelihood condition. (B) Subjects 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 4.9C), 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 4.9D). 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.

4.4.3 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 Koehler [201], that base rates are only ignored when they are manipulated between rather than within subjects.

Several lines of evidence support Koehler's argument. Fischhoff et al. [100] found greater sensitivity to base rates using a within-subject design, and similar results have been reported by Birnbaum & Mellers [25] and Schwarz et al. [322], though see Dawes et al. [63] for evidence that base rate neglect occurs even using within-subject designs. Ajzen [3] pointed out an asymmetry in the experiments of Kahneman & Tversky [190], 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, Ajzen [3] 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 neglect [364,8,229]. 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.

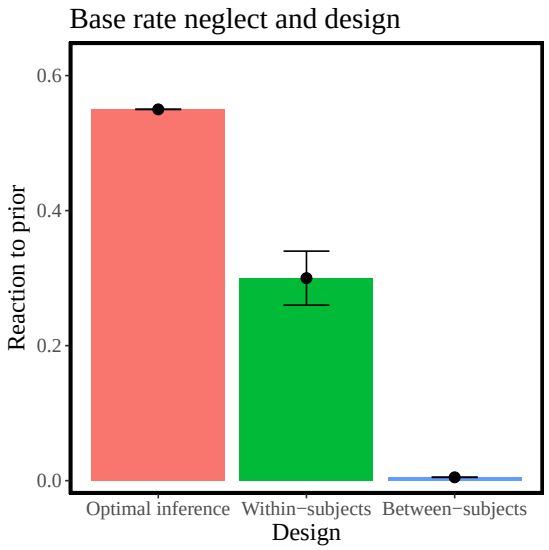


Figure 4.10: Base rate neglect within and between subjects. The y-axis shows the reaction to the prior as measured in predictions from the Learned Inference Model, the x-axis shows the different conditions. Reaction to the prior here is measured by the difference between the responses given to test queries 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 mentioned above, Fischhoff et al. [100] 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 4.10). In fact, the model exhibits total neglect of base rates in the between-subjects design, consistent with previous findings reported by [229], though not all experiments show such extreme results.* The Learned Inference Model naturally explains the difference between experimental designs 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 [8,98,265] 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 in [20] find greater under-reaction with increasing diagnosticity (Figure ??). We note that these studies predominantly used within-subject designs,† 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 parameters [240]. 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 contradic-

*The 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, observing these queries simply concentrates the query distribution in this space and reduces covariance between the prior and the posterior.

†Two exceptions to this pattern are [317] and [15].

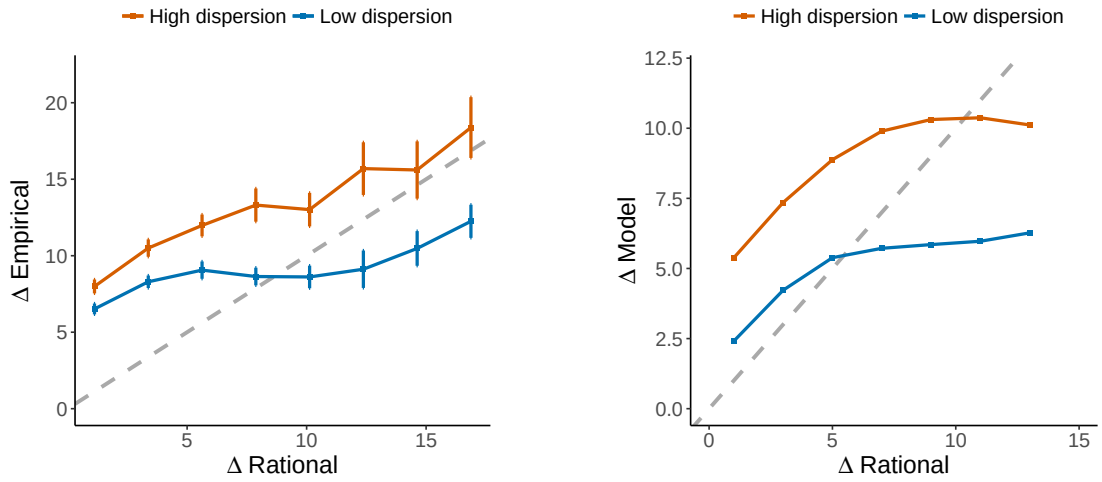
tory findings, by taking into account the experienced query distribution of each subject.

4.4.4 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 Gershman [113]. 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 $\mathcal{N}(m_0, v)$, with m_0 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 $v = 144$.

[113] 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 4.11A 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.* Subjects' updates are positively correlated with the model's predicted updates for both the high dispersion ($r(99) = 0.57, p < .001$) 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

*This rational hierarchical model is assumed to know the true parameter values for s, v and m_0. 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.



(a) Data from humans

(b) Simulation results

Figure 4.11: Inferential errors in a continuous domain. (A) Reanalysis of data from the payoff prediction task collected by[113]. (B) Simulations of the Learned Inference Model. Each panel shows subjective updates from prior to posterior (ΔData) on the y-axis and the update of a rational (hierarchical) model (ΔRational) on the x-axis. Error bars represent the standard error of the mean. Gray lines represent $y = x$.

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 return 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 4.11B, 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 4.5). 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.

4.5 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.

4.5.1 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., [89,259,262,178]. 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 [91,50]. Here we focus on the study reported by Cohen et al. [50], 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 4.12A), and were much less correlated in the unbelievable condition (Figure 4.12B).

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 [57,221], 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

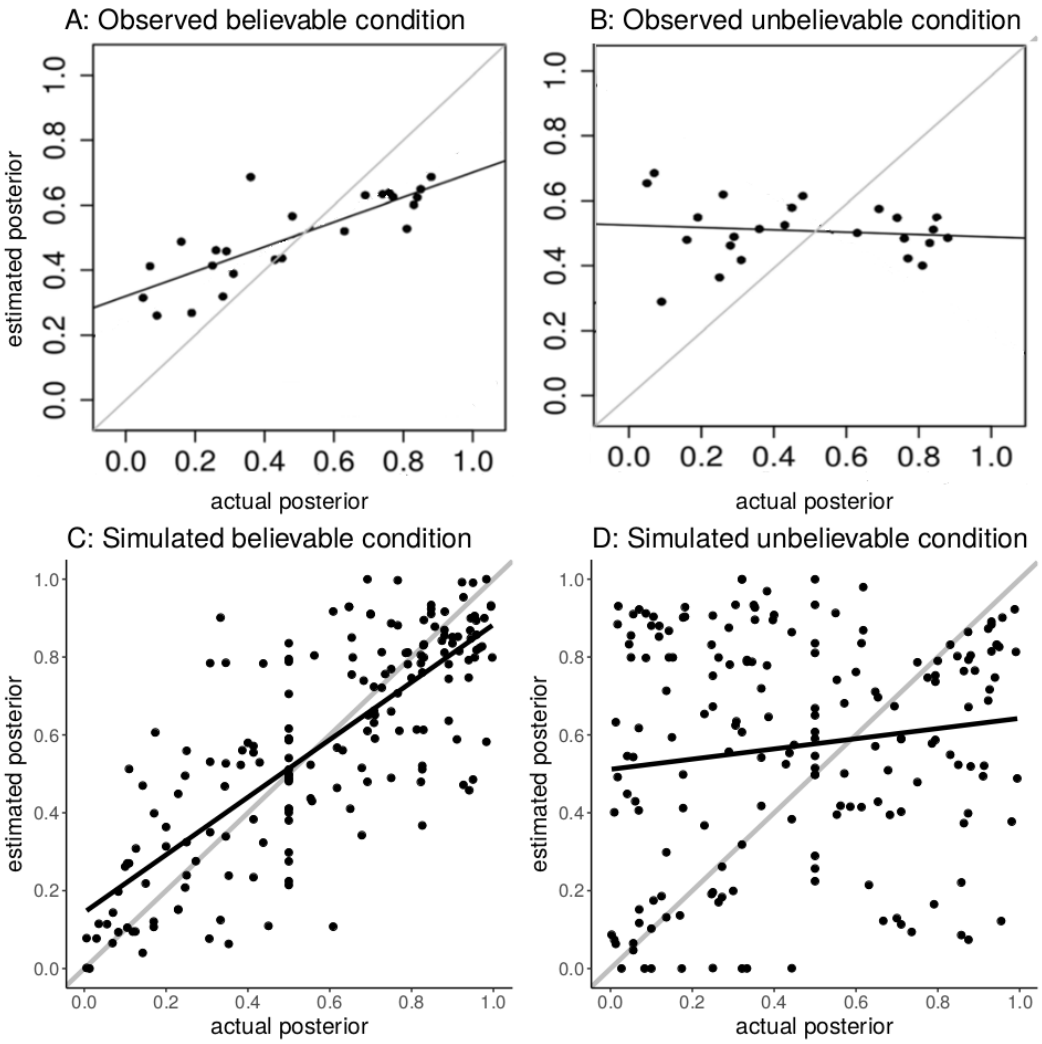


Figure 4.12: Belief bias. Top: experimental data. Bottom: simulations of the Learned Inference Model. (A) Empirical results for the believable condition[50]. (B) Empirical results for the unbelievable condition. (C) Simulated results for the believable condition. (D) Simulated results for the unbelievable 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.

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) or a different distribution P_B (unbelievable condition)*. The model produces responses that are highly correlated with the true posterior probability in the believable condition (Figure 4.12C, $r = 0.78$, $p < .001$), but this correlation is much lower in the unbelievable condition (Figure 4.12D, $r = 0.14$, $p = .06$, comparative test: $z = 2.64, p = .004$). Our model therefore reproduces the belief bias effect reported by[50].

4.5.2 Memory effects

In our own previous work[58], we observed signatures of amortized inference in subjects' probability estimates. One such signature was that their answers to a question (Q2) were predictably biased by their answers to a previous question (Q1). This bias was stronger in cases were 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 Q1 was "similar" to the one in Q2, measured by the KL divergence between the two posteriors over objects conditional on the cue object. In the other condition, the cue object in Q2 was dissimilar from the one in Q1.

For example:

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

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

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?"

appendixWe 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[103,369,57]. 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 4.13A). 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 memory effect when subjects in Q1 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 Dasgupta et al.[58], 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

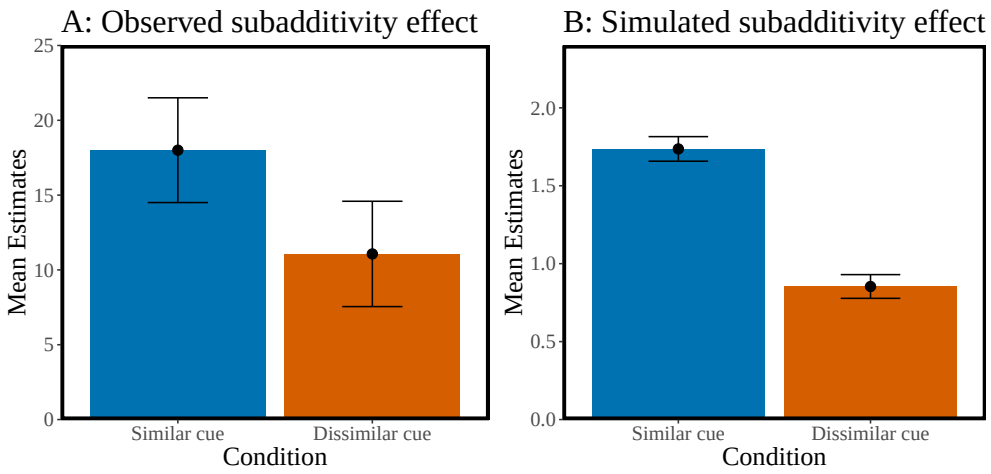


Figure 4.13: Memory effect. (A) Observed subadditivity effect in query 2 reported in[58]. Cues that were similar to a previous query showed a higher effect than cues that were less similar, indicating strategic reuse of past computation. (B) Simulated subadditivity effect. Provided that the model was 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.

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[58]. We simulated a data set of scene statistics with 12 objects with 2 different "topics" that drive the multinomial probability distributions over these 12 objects[29]. Using this setup, one can derive the joint probability of any 2 objects. The joint probability is all that is required for blackbox variational inference[290], 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 4.13B; $t(58) = 4.62, p < .001$). Our model therefore reproduces the difference in the memory effect reported by Dasgupta et al.[58]. 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 Dasgupta et al.[57] 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.

4.6 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).* 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.

To formalize this idea in the context of amortized inference, the optimization problem in Eq. 4.9 can be rewritten (up to an irrelevant constant factor) as follows:

$$\varphi^* = \underset{\varphi \in \Phi}{\operatorname{argmin}} \left[\mathcal{D}[Q_\varphi(h|d)||P(h|d)] + \frac{1}{P_{\text{query}}(d)} \mathbb{E}_{P_{\text{query}}} \left\{ \mathcal{D}[Q_\varphi(h|d')||P(h|d')] | d' \neq d \right\} \right]. \quad (4.13)$$

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 Zhu et al. [395]. 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 [291]. 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. [376,57] this produces a noisy estimate of the posterior that may deviate significantly from the true posterior. The correction procedure reduces variance in the posterior

*Epistemic uncertainty due to computational imprecision has been studied systematically in the field of *probabilistic numerics* [165].

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. 4.6). In the simple binary setting, $\mathcal{H} = \{0, 1\}$, the Monte Carlo inference engine generates M samples from $P(h|d)$*. In our generic formalism, the approximate posterior is parametrized by the proportion of "successes" $\varphi = K/M$, where $K = \sum_m \mathbb{I}[h^m = 1]$. The approximate posterior is then given by $Q_\varphi(h|d) = \varphi^h(1-\varphi)^{1-h}$. This approximation will exhibit large stochastic deviations from the true posterior for small M.†

To reduce the variance of the Monte Carlo estimator, Zhu et al. [395] proposed a meta-Bayesian inference procedure that computes the posterior over the optimal parameters φ^* given the "data" supplied by the random variable φ:

$$P(\varphi^*|\varphi) \propto P(\varphi|\varphi^*)P(\varphi^*). \tag{4.14}$$

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

$$\mathbb{E}\{\varphi^*|\varphi\} = w\varphi + (1-w)\frac{A}{A+B}, \tag{4.15}$$

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 Zhu et al. [395], the prior mean is $1/2$. This gives rise to a form of "conservatism" in which probabilities greater than $1/2$

*We 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 posterior [57,124].

†The variance of the Monte Carlo estimator for a binomial distribution with success probability p is $p(1-p)/M^2$.

are underestimated, and probabilities less than $1/2$ are overestimated[85,170]. 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[82].

[395] found evidence for such a "conservative" prior using two different data sets. The first one was data collected by[54], 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[349], 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 4.14A). 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 $1/2$. This prediction was confirmed in the experimental data (Figure 4.14B).

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 regularization, 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 Stewart et al.[349]. 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 4.14C. 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.[395] explained the quadratic relationship between the expectation and the variance as a feature of

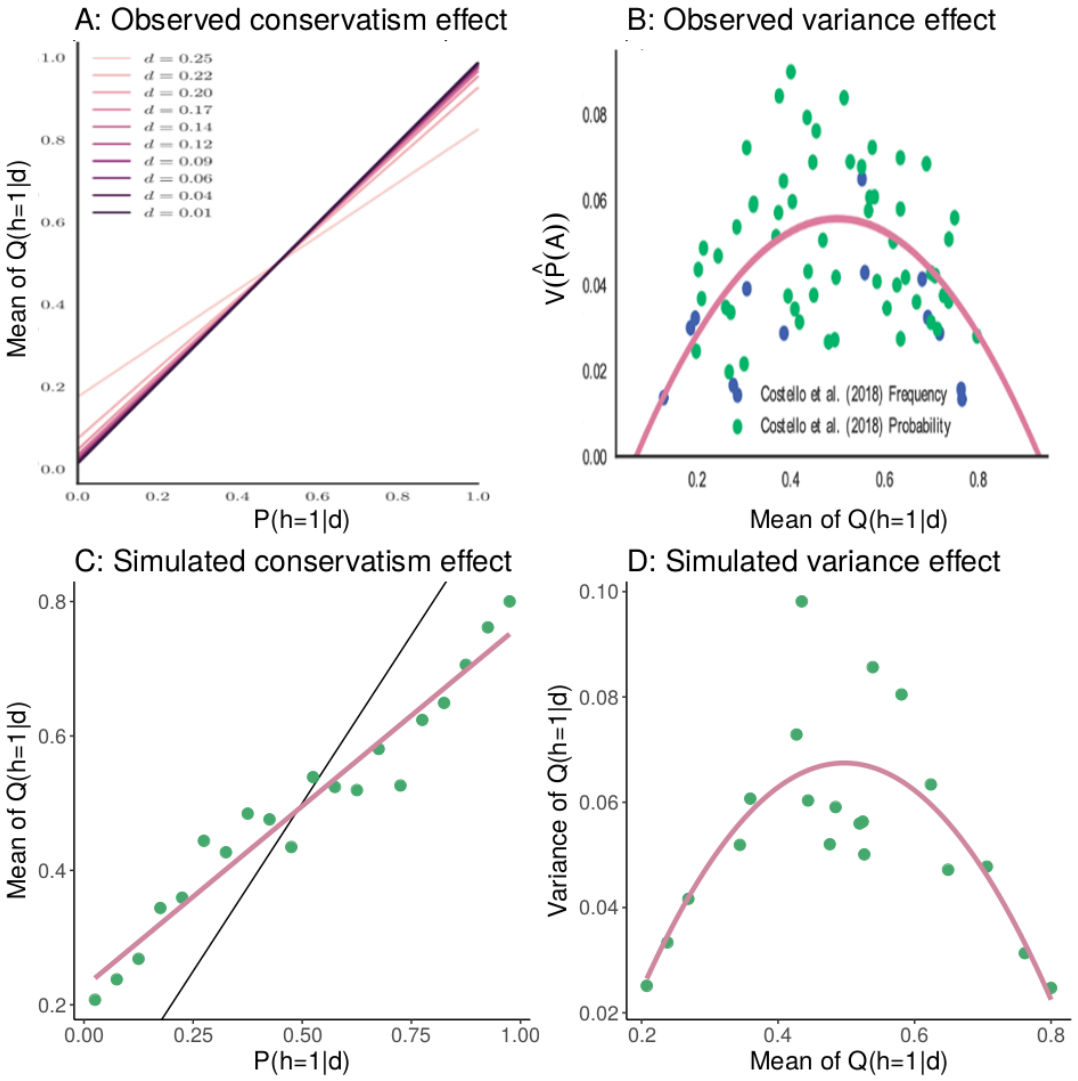


Figure 4.14: Correction prior. (A) Simulation results from the correction prior model in[395] exhibiting conservatism. Black line represents the optimal response and the colored lines show estimates from different parameterizations of the model. (B) Quadratic relation between the variance of subjective probability estimates and mean subjective probability estimates, as observed by[395]. 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.

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-off[110] – 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 by[395], $F(1, 19) = 78.73, p < .001$ (Figure 4.14D). 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 by[395], but without a stochastic sampling algorithm. In the General Discussion, we return to the relationship between learning to infer and stochastic sampling.

4.7 General Discussion

Although many studies suggest that the human brain is remarkably adept at carrying out Bayesian inference e.g.,[147,204,199,264], many other studies present evidence for systematic departures from Bayesian inference e.g.,[143,20,189,190,142]. 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 resources[120,217]. 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 world[210]. 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 underreact 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 underreaction 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[50]. 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[58]. 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[395], 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 we 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.

4.7.1 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 inferences[40]. 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 cut[333]. This interdependence between people's strategies and their environments has been stressed by psychologists for decades[361], 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, Thomas et al. [360] developed a theory of hypothesis generation based on memory mechanisms see[359] for an overview of this research program. Related ideas have also been explored in behavioral economics to explain decision making anomalies[33]. Our contribution has been to formalize these ideas within a computational rationality framework[120], 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.[135] 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,[325] 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.[22] 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,[201] 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 heuristic[38]. In one exemplary demonstration of how ignoring unpredictive information can be beneficial,[362] 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.

4.7.2 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 see[314] 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.,[42,155]. In our own prior theoretical work, we have employed hypothesis sampling to explain a range of inferential errors[57,58]. 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., [66,153], 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[57] or providing incidental information that serves as an "anchor" [221,222]. 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 [215,257,307] 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[37,36,308]. It resonates with work in other domains like reinforcement learning, where people seem to engage in offline simulation to drive value updating[121,126,254].

4.7.3 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[372], inability to map the provided probabilities onto an intuitive causal model[207], or simply disbelief in the experimenter's description of the data-generating process.* 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[166,167]. These are all likely part of the explanation 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-

*While 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.

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 [207,3]. 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 [131,71,296]. 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[217,120], 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[220,221,274]. Our model uniquely addresses how such biases could derive rationally from limited capacity 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.

4.7.4 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 black-box 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 operations[320].

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 problem[57]. The memory-based subadditivity effects that we modeled[58] 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[4,93]. 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[142,107,288], and this point is corroborated by evidence that inferential errors also appear in real markets with highly incentivized traders[10].

4.7.5 Conclusion

In his paper criticizing past research on base rate neglect,[128] 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.

4.8 Implementation details

4.8.1 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:

$$\log P(d) = \mathcal{L}[Q_\varphi(h|d)] + \mathcal{D}_{\mathrm{KL}}[Q_\varphi(h|d)||P(h|d)], \tag{4.16}$$

where

$$\mathcal{L}[Q_\varphi(h|d)] = \mathbb{E}_{Q_\varphi(h)}\left[\log P(h,d) - \log Q_\varphi(h)\right] \tag{4.17}$$

is the *evidence lower bound* (ELBO), also known as the *negative free energy*. The term ELBO comes from the fact that $\mathcal{L}[Q_\varphi(h|d)]$ 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 see[185], 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*[290]. The key idea is to approximate the gradient of the ELBO with a set of M samples:

$$\nabla_\varphi \mathcal{L}[Q_\varphi(h|d)] \approx \frac{1}{M} \sum_{m=1}^{M} \nabla_\varphi \log Q_\varphi(h^m|d) \left[\log P(h^m, d) - \log Q_\varphi(h^m)\right], \tag{4.18}$$

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

$$\varphi_{t+1} \leftarrow \varphi_t + \rho_t \nabla_\varphi \mathcal{L}[Q_\varphi(h|d)], \tag{4.19}$$

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

4.8.2 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 $f_\varphi : \mathcal{D} \mapsto [0, 1]$, where $\mathcal{D}$ denotes the data space, and the vari-

ational approximation was $Q_\varphi(h|d) = \text{Bernoulli}(h; f_\varphi(d))$. 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 Benjamin [20]. 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 $f_\varphi : \mathcal{D} \mapsto \mathbb{R}^2$, and the variational approximation was $Q_\varphi(h|d) = \mathcal{N}(h; f_\varphi(d))$. 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.

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 $\mathcal{N}(m_0, v)$.

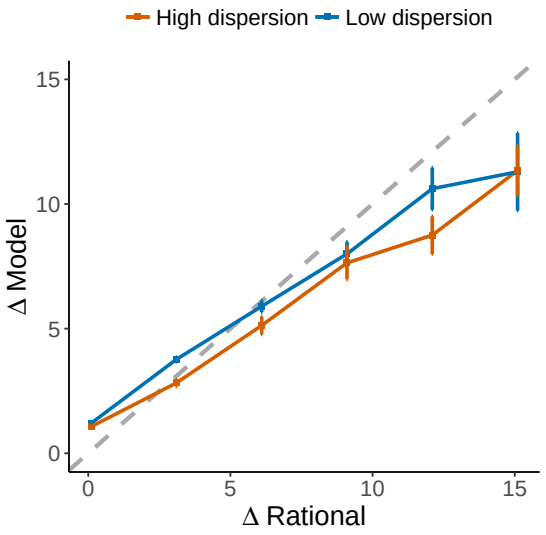


Figure 4.15: Performance of the L-HBM. Simulation results of a hierarchical Bayesian model that infers the underlying parameters in the the experiment reported by [113]. The Y-axis shows the L-HBM's updates from prior to posterior (ΔData) and the X-axis shows the update of a rational (hierarchical) model (ΔRational; 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: $s = 25, m_0 = 40$ 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_0, 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_0, 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. 4.15). For both the high and the low dispersion conditions, the updates closely follow the diagonal line of $y = x$. This indicates that inferring m_0, 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.

5

Heuristics in machines: A natural language case study

As modern deep networks become more complex, and get closer to human-like capabilities in certain domains, the question arises of how the representations and decision rules they learn compare to the ones in humans. In this work, we study representations of sentences in one such artificial system for natural language processing. We first present a diagnostic test dataset to examine the degree of abstract composable structure represented. Analyzing performance on these diagnostic tests indicates a lack of systematicity in the representations and decision rules, and reveals a set of heuristic strategies. We then investigate the effect of the training distribution on learning these heuristic strategies, and study changes in these representations with various augmentations to the training set. Our results reveal parallels to the analogous representations in people. We find that these systems can learn abstract

rules and generalize them to new contexts under certain circumstances – similar to human zero-shot reasoning. However, we also note some shortcomings in this generalization behavior – similar to human judgment errors like belief bias. Studying these parallels suggests new ways to understand psychological phenomena in humans as well as informs best strategies for building artificial intelligence with human-like language understanding.

5.1 Introduction

Recent years have seen a vast improvement in the capabilities of artificial intelligence systems, driven primarily by developments in deep neural networks see [213] for a review. These have allowed artifical system to reach human-level performance at video games [253], object recognition [310], and voice generation [268], as well as produced impressive performance in several other domains. However, some serious concerns haunt deep learning approaches and their promise as a general solution to artificial intelligence. Many of these concerns surround the lack of structure in the representations and decision criteria these systems learn [233,211]. This problem has been implicated in deep learning's data inefficiency and inability to learn abstract structure from few examples, its difficulty in utilizing hierarchical structure and fostering transfer between tasks and domains, as well as the challenge of integrating established prior information into deep learning systems. It also presents serious concerns about the interpretability of its representations and decision criteria, making them less dependable and risky for deployment in sensitive or highly variable domains.

All of this points to a crucial problem: how can we better understand the representations learned by these systems? Existing studies e.g., [192,214,392,393] primarily use approaches inspired by neuroscience methods developed to understand the brain, for example the statistical analysis of unit activations, and ablation studies where specific units are disconnected or deactivated. These methods promise interesting bottom-up insights into the inner workings of these systems. Cognitive science provides another set of tools to approach this problem from the top down [301,187,243], by decomposing cognitive processes into their computational components, building models that incorporate these components, and test-

ing these by making predictions about behavior on carefully selected test problems that distinguish different hypotheses.

The cognitive science approach has yielded huge benefits in understanding higher-level cognition in humans, a prime example of which is the human ability to learn, understand, and produce language[49,225]. This domain exemplifies a hallmark of human intelligence: the ability, in the words of von Humboldt, to "make infinite use of finite means." Specifically, human cognitive abilities have been characterized as *systematic*[102,209] – this indicates an algebraic capacity to produce new combinations from known components. For example, when a person learns a word in a specific context as part of a particular sentence, they can immediately use this new word in an infinity of other sentences in which this word has never previously been encountered. Systematicity therefore allows humans an impressive capacity to *generalize*, transferring knowledge from one context to others. This ability requires the representations underlying this newly learned word for example, to be abstract (not tied to specific contexts) and compositional (possible to combine with other words and sentences). The failure of neural network models to achieve such systematicity has been a recurring (and controversial) theme in cognitive science[102,211]. This concern has previously been studied specifically in the domain of natural language[208,112,17], demonstrating the lack of abstract compositional reasoning in certain networks. These analyses are often carried out on toy systems, and while they demonstrate conclusively the lack of systematicity, they largely neglect a deeper analysis of what the systems do learn.

In this paper, we carry out an analysis of the representations learned by a state-of-the-art model for a difficult natural language processing task. We discover that its representations are not systematic; instead, the model uses various heuristic strategies. We then investigate how these heuristics might arise. Analyses of the training distribution reveal that it is very biased, containing many unintended structural regularities that can be exploited by these much simpler heuristics. These simple rules are therefore easily acquired by the neural network, since they explain a substantial amount of variance without having to invoke a more complex systematic representations. We then carry out various augmentations to the training set to better understand if the system can learn abstract composable repre-

sentations, given the right training distribution. We find parallels between our findings and studies of human representations in terms of how systematic they are under certain circumstances, as well as in terms of when and where this systematicity breaks down. We discuss how such analyses can be fruitful to both cognitive science and machine learning.

5.2 Background

In this section we review some background on the kinds of representations we will be studying (vector space embeddings of sentences). We also review the three key factors in how such embeddings are generated: the task that they are optimized for, the architecture of the model used to perform that task, and the training distribution on which performance is optimized.*

5.2.1 Vector space embeddings

Vector space models represent items as vectors in some metric space. These have a long history in cognitive science as models of semantic representations[350,16,281]. In particular, in the domain of language, vector space models of words (also known as word embeddings) that are learned using distributional information (statistics of text corpora) have been shown to encode syntactic as well as semantic structure, and have been used in psychological models for syntactic category acquisition[293], inductive vocabulary learning[212], analogical reasoning[309], categorization[184], and high-level associative judgments[22]. Modern machine learning has allowed the mining of very large datasets to produce vector space embeddings that are now commonly used as the word representations in artificial intelligence systems for natural language processing[280,248].

Understanding language requires understanding not only words, but also their relations within a sentence. These relations are abstract and composable, allowing language to be combinatorially productive – with a finite set of words, one can systematically produce an infinite set of sentences simply

*The details and implementation of the optimization algorithm also contribute see[306] for an overview, but as long as the optimization reaches convergence this has relatively little effect, and we leave this out of our current discussion.

by creating new and longer combinations of these known words. The number of sentences in a language therefore far exceed the number of words. For this reason, generating similar vector embeddings for sentences has proven challenging. Recent papers have developed several supervised as well as unsupervised approaches to learning vector space representations of sentences using recurrent neural networks (RNNs) that are able to represent the order of words in a sentence[197,171,51]. These are intended to capture sentence-level semantic content, and have been shown to perform reasonably well on transfer tasks (sentence-level semantic tasks on which the embeddings were not specifically trained). In particular, the performance of these sentence models exceeds the performance of representations that treat sentences as bags of words (BOW models) – these patently lack any order information about the words, therefore ignoring the abstract and composable relational structure at the sentence level. However, it is unclear exactly what relational information between words is actually represented in such RNN sentence models. In this work, we start to shed light on this question.

5.2.2 Natural Language Inference

The sentence embeddings we analyze are trained on the natural language inference (NLI) task. The goal is to classify pairs of sentences (a premise and a hypothesis) into 'entailment', 'contradiction', or 'neutral', depending on the semantic relation between the two sentences. This is a popular domain for studying artificial representations since it has a lot of relatively interpretable underlying structure[134,243,260]. For example, it is a simple domain in which abstract and composable relational structure is required – word-level information is not generally sufficient to perform well on this task. The premise sentence "Anne is more cheerful than Bob" contradicts the hypothesis sentence "Anne is less cheerful than Bob", but entails the hypothesis sentence "Bob is less cheerful than Anne". Here, both the hypothesis sentences have the exact same words, and would be indistinguishable if we were just comparing the words in them. More generally, X is more Y than Z entails that Z is less Y than X, for any X, Y and Z. In this case, the specific words used almost don't even matter, and the bulk of the information is in the relations between the words in the sentence. Encoding abstract rules like this allows

us to systematically carry out natural language inference on combinatorially many different sentences, with different Xs, Ys, and Zs.

The human ability to carry out abstract reasoning of this sort is a richly studied topic. Some of these abilities however are so obvious, that they are often simply taken for granted without formal study. For example, it is reasonable to assume that any adult human (in the absence of time pressure or cognitive load) can fairly easily process that if X is more Y than Z, then in general Z is less Y than X irrespective of the specific meanings of X, Y and Z. In this paper, we investigate to what extent certain machine-learned sentence embeddings can represent and use such abstract rules in natural language inference.

Despite the generally acknowledged power of human abstract reasoning, a number of studies indicate that humans are not perfect: semantic content (for example the specific meanings of the X, Y and Zs above) has been shown to interfere with systematic inferences in an effect often termed 'belief bias'[35,183]. This effect is especially noticeable in children[92], as well as adults under time pressure or cognitive load[88]. In the last part of this paper, we discuss similarities between humans and machines in how they fail certain tests of systematicity.

5.2.3 Models for sentence embeddings

The sentence embeddings we study in this paper are from a highly successful NLI system, InferSent[51]. Each premise and hypothesis sentence are input to a sentence encoder as a sequence of pre-trained 300-dimensional GloVe word embeddings[280]. These word embeddings already contain a lot of information about the semantic and syntactic roles of the words (see section on Vector space embeddings for details), and therefore a large part of the lexical information is already represented. Therefore the bulk of the work InferSent has to do is to learn and represent how these words relate to one another in a sentence to provide meanings. The sentence encoder takes in this variable length input and, after passing it through various recurrent and convolutional layers see[51] for details, provides a 4096-dimensional vector as output. This output vector serves as a sentence embedding. To make the final inference,

these sentence embeddings for the premise and hypothesis are fed to a simple classifier described in Figure 5.1 that labels each pair as entailment, neutral or contradiction. The network is trained end-to-end with supervised learning, using a large labelled dataset for NLI (see next section for details on this dataset). The learned embeddings were shown to perform well on other sentence-level tasks (such as sentiment analysis, semantic textual similarity and other natural language inference datasets) by re-using the sentence encoder and training only the classifier for the specific task at hand. This indicates that the sentence encoder does capture some semantic content in the embeddings.

For our tasks, we replicate the procedure in Conneau et al.[51] to obtain sentence embeddings. These are henceforth referred to as the InferSent sentence embeddings. Our trained InferSent model gives us 84.73% accuracy on validation and 84.84% accuracy on the test dataset, which is comparable to the performance of the classifier reported in Conneau et al.[51]. For comparison, we also train a bag-of-words (BOW) baseline model that averages the pre-trained GloVe word embeddings for all the words in the sentence to form a sentence embedding. These embeddings cannot represent abstract relational structure, since the architecture of the model used to generate them (a simple average of the word embeddings) cannot express word order. We then train a simply classifier on these embeddings to perform natural language inference. This model achieves 53.99% accuracy on the SNLI test set comparable to the BOW performance reported in[51].

Neural networks can act as universal function approximators[332,175], and given sufficient capacity, they can represent any arbitrarily complex set of relations between the words in the sentence. The InferSent model has a very large capacity due to a large number of layers and hidden units see[51], so a lot of abstract compositional structure is in theory within the representational capacity of these sentence embeddings. In this paper, we analyze how much systematic structure is actually learned and utilized for the NLI task at hand.

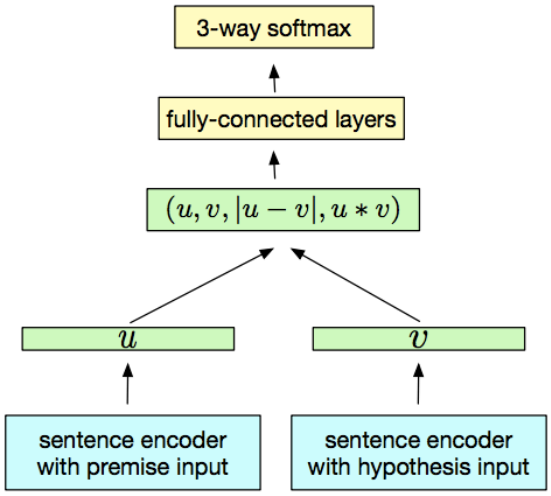


Figure 5.1: InferSent architecture[51].

5.2.4 Training datasets

To understand sentence embeddings like the ones learned by InferSent, it is imperative to not only consider the model specifications for the system that produces them (in this case the specific end-to-end architecture of the network in InferSent), but also the learning signals it receives from the training set. For many deep learning based methods, very little information about the structure of the task is baked into the architecture of the models – the only structure about language that it is endowed with before training are the biases that come with using a recurrent neural network as the architecture. This specifies that sentences have variable-length, sequential structure. These embedding models are therefore fairly 'tabula rasa', and most of what they represent about the structure of the task (in this case natural language inference) is learned from training data. As elaborated in the previous section, some abstract compositional structure is within the representational capacity of the InferSent sentence embeddings – but whether or not the right structure is actually learned and represented depends largely on the training data. The significance of the training set on the representations learned by flexible deep learning methods is often not adequately considered. One contribution of this work is to highlight and analyze this issue.

InferSent was trained on the Stanford Natural Language Inference (SNLI) dataset[34], a popular labelled dataset for natual language inference. SNLI consists of 550k premise-hypothesis sentence pairs, and is balanced (consists of equal number of pairs with entailment, contradiction and neutral relationships). The dataset was generated with a crowd-sourcing framework. Workers were presented with a scene description from a corpus of image captions that act as the premise, and asked to supply hypothesis sentences that have each of the three possible NLI relations (entailment, neutral, and contradiction) to the given premise. The freedom to produce entirely novel hypotheses leads to a rich set of sentences; however, it also leads to some artifacts that can strongly bias the representations learned by a 'tabula rasa' system. We discuss these in later sections.

5.3 A test dataset of minimal cases: The Comparisons dataset

Our goal is to understand the representations and decision criteria learned by InferSent, in particular how much systematic relational information they encode and utilize – do they represent abstract rules for the ways words combine to give meaning to sentences? In the machine learning literature on natural language processing, any performance above the bag-of-words (BOW) baseline (that only receives the words in the sentence with no order information) is often seen as proof of the encoding and utilization of relational information. However, this is an unwarranted conclusion—the BOW baseline usually receives only averaged word vectors for the sentence, and therefore also loses some of the lexical information. It often does not actually reach the best possible performance with only the words. Performance above this baseline therefore does not license the conclusion that relational information is being encoded and used at all.

Here, we pursue an alternative approach, inspired by traditions in cognitive psychology and psycholinguistics of building diagnotic test sets to investigate the underlying representations and decision rules. The goal is to generate a set of sentence pairs such that encoding the relations between words (in addition to the words themselves) is *required* to correctly classify them into the three NLI classes. Diagnostic test datasets such as these, that posit a hard baseline for performance without relational

information, provide a more foolproof way to test whether such information is being used.

We considered pairs of sentences such that the NLI relation between the sentences can be changed without changing any of the words in the sentence, only their order. We generated our test dataset using comparisons as these are easy to fit into the NLI framework, and yield many simple examples of sentence pairs that require more than word-level data to understand. For example, the premise sentence "the woman is more cheerful than the man" contradicts one hypothesis sentence, "the woman is less cheerful than the man", but entails another hypothesis sentence, "the man is less cheerful than the woman". Since both hypothesis sentences have the exact same words, they would be indistinguishable if we were just comparing their bag-of-words representations. Therefore, a model based only on the words, and not considering the relations between them, would at most get one of the two classifications right. This caps the bag-of-words performance at 50%, and some relational rules must be learned to perform above this baseline.

Generation of several such sentence pairs can be easily automated. We considered three sub-types, described below and summarized in Tables 5.1 and 5.2.

5.3.1 Same type

Premise-Hypothesis pairs differ only in the order of the words.

```
Premise: The woman is more cheerful than the man
Hypothesis: The man is more cheerful than the woman
CONTRADICTION
Premise: The woman is more cheerful than the man
Hypothesis: The woman is more cheerful than the man
ENTAILMENT
```

5.3.2 More-Less type

Premise-Hypothesis pairs differ by whether they contain the words 'more' or 'less'.

```
Premise: The woman is more cheerful than the man

Hypothesis: The woman is less cheerful than the man

CONTRADICTION

Premise: The woman is more cheerful than the man

Hypothesis: The man is less cheerful than the woman

ENTAILMENT
```

5.3.3 Not type

Premise-Hypothesis pairs differ by whether they contain the word 'not'.

```
Premise: The woman is more cheerful than the man

Hypothesis: The woman is not more cheerful than the man

CONTRADICTION

Premise: The woman is more cheerful than the man

Hypothesis: The man is not more cheerful than the woman

ENTAILMENT
```

Type	Entailment hypothesis	Contradiction hypothesis
Same	X is more Y than Z	Z is more Y than X
More-Less	Z is less Y than X	X is less Y than Z
Not	Z is not more Y than X	X is not more Y than Z

Table 5.1: Rules in Comparisons dataset for Premise: X is more Y than Z

To facilitate comparison with the SNLI dataset, we ensured that the vocabulary distribution of our Comparisons dataset is similar to the original SNLI training dataset.* This ensured that we are only

*Only a few words differed by more than 1% from their occurrence rate in SNLI, such as *not, a, than, the,*

Type	Number of sentence pairs
Comparisons (same)	14670
Comparisons (more-less)	14670
Comparisons (not)	14670

Table 5.2: Comparisons dataset summary.

manipulating the relational structure of the test set, and poor performance cannot be attributed to not having experienced the specific words before.

5.4 Testing the sentence embeddings

We tested the two classifiers based on two different sentence embeddings (the InferSent sentence embeddings, and the BOW sentence embeddings) on the constructed test set (the Comparisons dataset, Table 5.2). Both of these classifiers were trained for the same task (Natural Language Inference), on the same training dataset (SNLI), and differed only in the model used to generate them. The InferSent embeddings had access to word order, while the BOW embeddings did not (see Section 'Models for sentence embeddings' for details). The overall performance of each of the two classifiers on the Comparisons dataset are given in Table 5.3, and analyzed in greater detail in the following sections.

Type	BOW	InferSent
same	50.0	50.37
more/less	30.24	50.35
not	48.98	45.24

Table 5.3: Performance on the Comparisons dataset.

is, less, more. This was inevitable given the general structure of the comparison sentence pairs we use. All of these words however did still occur in the SNLI training corpus, and were not new to the model at test time.

5.4.1 Performance of Bag of Words

We found that the BOW embeddings make classifications that are exactly symmetric across the two true labels (entailment and contradiction) in each task (rows in Figure 5.2). This is expected since the sentence pairs with one label are just permuted versions of the sentence pairs with the other label. Therefore BOW cannot distinguish them, and necessarily classifies both of them the same way. This also ensures that the performance is capped at 50%. Asymmetry between the classifications of the two categories can occur only when relational information is encoded in the sentence embedding.

Considering the aggregate performance of BOW in Table 5.3, we found that performance, particularly on the 'more/less' type subset of the test dataset (30.24%), was significantly below 50%. This highlights the trouble with using BOW embeddings as a baseline for the encoding and use of relational information. Up to 50% performance is achievable on this dataset without using any relational information; therefore performance above the BOW baseline of 30.24% does not necessarily imply the use of relational information.

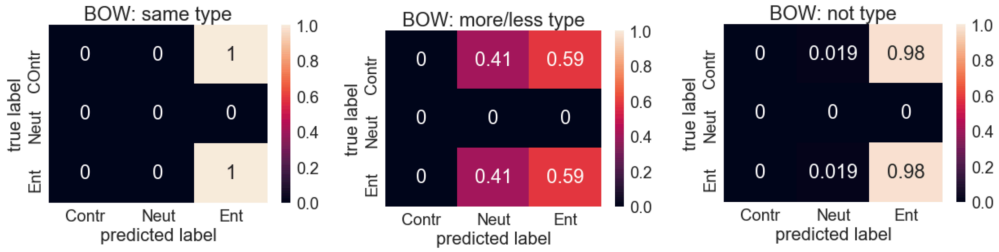


Figure 5.2: BOW embedding confusion matrices, with normalized rows.

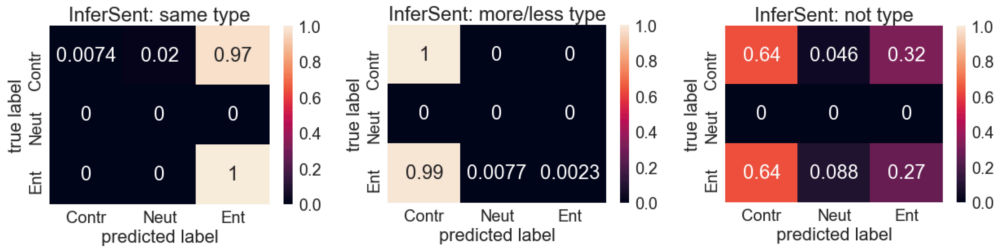


Figure 5.3: InferSent embedding confusion matrices, with normalized rows.

5.4.2 Performance of InferSent

The performance of the InferSent embeddings was slightly asymmetric (Figure 5.3), indicating that it was able to distinguish sentences slightly, based on relational information. Yet overall the InferSent embeddings were extremely poor at this task (Table 5.3), achieving performances slightly above 50% for two of the three sub-types of sentence pairs in the Comparisons dataset, and even less than 50% in a third sub-type. This indicates that InferSent embeddings do not correctly encode and utilize the kinds of abstract relational rules we tested with the Comparisons dataset.

However, InferSent's performance on another test dataset (the SNLI test dataset) is as high as 84% – so it is clearly encoding some relevant information about natural language inference. Further, a quick glance at Figure 5.3 indicates that InferSent does not respond randomly to the queries in our Comparisons dataset, but rather in some structured (though incorrect) way. Rather than simply conclude that InferSent embeddings are not systematic and leaving things at that, we can study patterns in the incorrect classifications made to better understand the underlying representations and decisions rules. Since our test dataset is highly structured, it allows a controlled way to generate and test hypotheses about the heuristic representations and decision rules InferSent implements.

Apart from isolating and characterizing these heuristics, it is also instructive to consider how InferSent might come to encode them in the first place. To answer this, we look to the study of heuristic strategies in humans. The theory of ecological rationality[361,334] posits that a system can exploit structural regularities in its learning environment using heuristics that achieve close to optimal performance in that specific environment. These might be much simpler than the most general strategy that performs well in all environments. Heuristics that leverage these structural regularities are therefore termed 'ecologically valid' in that environment. This suggests that we can better understand how heuristic strategies might arise in InferSent by examining if they are ecologically valid in its 'learning environment' (i.e., the training set). In the following sections, we delve into the heuristic strategies that explain performance on our Comparisons datset, as well as how InferSent might have come to encode them by testing their ecological validity in the SNLI training dataset.

Overlap Heuristic

We note in Figure 5.3 that almost all the sentence pairs in the same-type comparisons were classified as entailments, despite half of them being true contradictions. A distinguishing feature of the same-type comparisons is that the premise and hypothesis sentences have full word overlap (they both contain exactly the same words). This observation allows us to hypothesize an *overlap heuristic*: high overlap in words between premise and hypothesis biases InferSent against classifying the pair as a contradiction.

While we have seen some evidence that this heuristic is indeed at play (based on the performance on the same-type comparisons), the question remains as to why it encodes this rule. With our knowledge of language, we know this simple rule to reflect on incorrect understanding of natural language inference. However, all the knowledge about the NLI task that InferSent encodes is from its training dataset. If the dataset has underlying structural regularities that can be exploited by simple heuristic strategies, then a tabula rasa model for NLI such as InferSent that is trained on this dataset will learn to encode it.

We carried out an analysis of the SNLI dataset to determine if the overlap heuristic is ecologically valid in it. First, we observed anecdotally that indeed several contradictory sentence pairs have no overlap in words. For example, a contradictory sentence pair in SNLI is:

```
Premise: Several people are trying to climb a ladder in a tree.
Hypothesis: People are watching a ball game.
CONTRADICTION
```

To quantitatively verify this observation, we ranked all the sentence pairs in SNLI by overlap rate: $\frac{\text{\# of overlap words}}{\text{total \# of words}}$ (in non-increasing order). We then considered the top X sentences with highest overlap for different Xs. As shown in Table 5.4, when considering the full dataset, the distribution is balanced (the percentage of entailments, contradictions and neutral sentences are equal). However, we found that as the word overlap in the sentences increases, the percentage of contradictions drops. When considering only the top 1000 sentence pairs for overlap, we found that 91.5% of them have entailment or neutral labels, with only the remaining 8.5% having a contradiction label.

Top	Entailment	Neutral	Contradiction
All	33.4	33.3	33.3
10000	39.5	35.7	24.8
1000	50.8	40.7	8.5

Table 5.4: Percentage of entailments split by overlap rate of words in SNLI.

It is therefore natural that InferSent encodes the simple overlap heuristic as a predictor of contradiction. This explains not only the failure of InferSent to generalize its good performance on SNLI to the same-type comparisons in our test dataset, but also matches the specific failure mode we observe in its responses.

Antonyms Heuristic

We note in Figure 5.3 the opposite trend for the more/less-type comparisons, where almost all the sentence pairs were classified as contradictions, despite half of them being true entailments. A distinguishing feature of the more/less-type comparisons is that the the premise and hypothesis always differ by one word – if the premise contains the word 'more' ('less') then the hypothesis always contain the word 'less' ('more'). This observation allows us to hypothesize an *antonyms heuristic*: sentences differing in the presence of words that have opposing meanings (antonyms) tend to be classified by InferSent as contradictions, irrespective of the other words or their order in the sentence.

Similarly to the previous section, we investigated the training dataset to elucidate if this heuristic is ecologically valid in InferSent's training set. Anecdotally, we saw that the contradicting hypotheses provided by crowd workers to generate SNLI do follow this pattern. For example, a contradictory sentence pair in SNLI is:

```
Premise: A man in a white t-shirt takes a picture in the middle of the street
with two public buses in the background.
Hypothesis: A man is wearing a black t-shirt.
```

CONTRADICTION

To verify this observation quantitatively, we analyzed the statistics of antonym usage in SNLI. To test whether a sentence pair (A,B) contains antonyms, we went through each word in sentence A, and considered all synonyms of that word, and considered all antonyms of those synonyms. Finally, we checked if sentence B contained any of those antonyms. These synonyms and antonyms were found using the NLTK WordNet software[23]. We then considered two different statistics. First, we calculated P(Contradiction | Antonym), which is the probability that a sentence pair is a contradiction given that its premise and hypothesis contain an antonym pair. This measures how well the presence of antonyms predicts a contradiction label in the training set. Second, we calculated P(Antonym | Contradiction), which is the probability that a contradictory sentence pair contains antonyms. This measures how well a contradiction label predicts antonyms. Both statistics were compared with the equivalent statistic for entailment, to provide a baseline for comparison. Table 5.5 shows that the presence of antonyms strongly predicts a contradiction label in the SNLI dataset (61.2% compared to chance at 33.3%). We also found that a contradiction label predicts the presence of an antonym pair (12.2%) more strongly than entailment did (3.5%). This indicates that the antonyms heuristic can explain significant variance for the contradiction label in the training set.

	P(Antonym | X)	P(X | Antonym)
X = Contradiction	12.2%	61.2%
X = Entailment	3.5%	18.0%

Table 5.5: Percentage of entailments split by antonym word pair in the SNLI dataset.

Since most of our Comparisons dataset contained a large amount of overlap between premise and hypothesis, the rules InferSent applies when responding to these test questions might be biased towards those learned in similar high-overlap settings during training. We checked the statistics of antonymy in the the high overlap subset of SNLI (top 10,000 highest overlap) to provide a closer comparison (Table 5.6). Here, contradiction predicts the presence of an antonym pair (43.7%) more strongly than in

the whole dataset (12.2%). The difference between P(Antonym | Contradiction) and P(Antonym | Entailment) is also more pronounced in this high overlap subset. The presence of an antonym pair no longer predicts contradictions at a high rate (28.9 %), but this is possibly due to the very low base rate of contradictions in the high overlap subset of SNLI, as compared to entailments.

	P(Antonym | X)	P(X | Antonym)
X = Contradiction	43.5%	28.9%
X = Entailment	8.7%	34.3%

Table 5.6: Percentage of entailments split by antonyms in high overlap SNLI subset.

These results suggest again, that the underlying statistics of the SNLI dataset allow models, including InferSent, to perform well with simple lexical heuristics that ignore the order of words and their relations.

Negation Heuristic

We see in Figure 5.3 that the not-type comparisons are preferentially classified as contradictions. A distinguishing feature of the not-type comparisons is that the premise and the hypothesis differ by the presence of the negation 'not'. This observation allows us to hypothesize a *negation heuristic* where sentence pairs that differ in the presence of negations are preferentially classified as contradictions.

Following procedures analogous to previous sections, we first noted anecdotally, that this heuristic seems to have validity in the contradicting hypotheses in SNLI. For example, a contradictory sentence pair in SNLI is:

```
Premise: Men turn to the camera to smile on the middle of three long tables
in a refectory.
Hypothesis: The man is not smiling.
CONTRADICTION
```

We verified this observation quantitatively by looking at the statistics for negation in SNLI. We collected all sentence pairs that contain "negating N-grams": no, not, n't (by considering "n't", we included words such as "don't" or "doesn't"). We then carried out analyses similar to the previous section, where we checked (1) the predictive power of negations on contradictions (P(Contradiction | Negation)), and (2) the predictive power of contradiction on negations, P(Negation | Contradiction), and compare both of these to statistics for entailment as a baseline. We found (Table 5.7) that the presence of a negation strongly predicts contradiction in the SNLI dataset (58.4% compared to chance at 33.3%). We also found that while both numbers are very low, a contradiction predicts the presence of a negation (3.3%) slightly more strongly than entailment does (1.1%). We also carried out the same analysis for a high-overlap subset (top 10,000 highest overlap) of SNLI to maximize similarity with our comparisons dataset and saw similar results (Table 5.8). In fact, the presence of negation predicts a contradiction, P(Negation | Contradiction) = 60.0%, at rates comparable to that in the full dataset, P(Negation | Contradiction) = 58.4%, despite the much lower base rates of contradiction in this subset of the data. This indicates strong ecological validity for this heuristic in the high overlap subset of the SNLI dataset.

	P(Negation | X)	P(X | Negation)
X = Contradiction	3.3	58.4
X = Entailment	1.1	20.0

Table 5.7: Percentage of entailments split by negation in SNLI dataset.

	P(Negation | X)	P(X | Negation)
X = Contradiction	1.3	60.0
X = Entailment	0.1	7.5

Table 5.8: Percentage of entailments split by negation in high overlap SNLI subset.

Summary of heuristics

We found evidence for three heuristics that explain the bulk of the patterns seen in the performance of InferSent on our Comparisons dataset, all of which are ecologically valid in the SNLI dataset. First, we identified the *overlap heuristic* where a large overlap in words between two sentences leads InferSent to not classify them as contradictions. Second, we identified the *antonyms heuristic* and the *negation heuristic*, where the premise and hypothesis differ in the presence of an antonym or a negation, which leads InferSent to classify them as contradictions.

These illustrate a disproportionate dependence on lexical (rather than relational) meaning in the representations and decision rules used by InferSent. While these heuristics serve well in certain domains, for example in SNLI, they don't amount to a more general encoding of entailment and contradiction between sentence pairs, as evidenced by InferSent's poor performance on our Comparisons dataset.

The analysis so far has highlighted word-level heuristics that InferSent might be using. Yet the confusion matrix results (Figure 5.3) show a slight asymmetry, indicating at least minor multi-word effects. This suggests that InferSent might be using some (potentially also heuristic) encodings for word order. However, a systematic analysis of the effect of word order, and how much variance such heuristics might explain, is challenging due to the combinatorial explosion in the number of possibilities. We leave a thorough investigation of this to future work.

5.5 Augmenting the learning environment

The foregoing results suggest that such ecological validity of simple heuristics in the SNLI training data (InferSent's learning environment) could explain why InferSent acquires them over a more abstract, systematic representation of the relations between words in a sentence. This leaves open the question of whether architectures such as InferSent are capable of learning the abstract relational rules needed to succeed at our task given a different training set where simple heuristics no longer explain so much

of the variance. RNN architectures like the one in InferSent can in theory represent the relational structure required to encode the abstract rules of the sort in Table 5.1 (see Section 'Models for sentence embeddings' for details). But how might we get them to learn and use them? In this section, we explore this question by training the InferSent model on part of the Comparisons dataset, and testing on a held-out subset of it. This serves to test whether simple training on examples of the rules in Table 5.1, will enable InferSent to encode some abstract relational rules.

The total training subset of our Comparisons dataset consists of 40k sentence pairs (7% the size of the 550k pair SNLI training set). Validation and test sets consist of 2000 sentence pairs each. There are no overlapping sentence pairs between any of these sets, therefore simply memorizing the training set will not allow good test performance. Good test performance requires the encoding and utilization of an abstract relational rule.

Epoch	Performance (%)		
	Train(Comp)	Test(Comp)	Test(SNLI)
0	47.81	45.36	84.84
13	99.91	99.8	56.37

Table 5.9: Results of fine-tuning InferSent on the Comparisons dataset.

We started with the original InferSent embeddings already trained on the SNLI dataset, and then fine-tuned it on our new Comparisons dataset using the same protocols used in[51] to train InferSent. Results are shown in Table 5.9. We found that using this method, performance on the SNLI data task degrades over the course of fine-tuning on the new Comparisons dataset from 84.84% to 56.37%. This points to over-fitting to the Comparisons data, at the cost of representing information necessary for SNLI. We found however, that performance on the Comparisons test set is much higher (99.8 %) than when trained only on SNLI (47.81%). Note that this test set consists of sentence pairs InferSent has never seen before. We thus find that the model architecture for InferSent, given the right training data, can encode some form of abstract relational structure that allows it to learn rules of the form in Table 5.1 and apply them to new sentence pairs – in particular sentence pairs with Xs, Ys and Zs that it has

never seen in that combination before.

Epoch	Performance (%)		
	Train(Combined)	Test(Comp)	Test(SNLI)
0	33.33	33.33	33.33
12	90.99	100.00	84.96

Table 5.10: Results of retraining Infersent on both SNLI and the Comparisons dataset.

We then checked whether InferSent can represent this relational structure without losing the information necessary for SNLI. We started with an untrained network, and then trained on an augmented version of the original training data. Here, examples from the SNLI training set were randomly interleaved with examples from our Comparisons training dataset, otherwise using the same training protocols reported in[51]. The test results are reported in Table 5.10. We found that the accuracy obtained this way on the SNLI test set (84.96 %) is comparable to the model trained only on SNLI (84.84 %). Moreover, test accuracy on the Comparisons dataset is close to perfect (99.55 %) and is much higher than the model trained only on SNLI (47.81 %). This establishes that in this case the model has enough capacity to achieve high performance on specially designed edge-cases like the Comparisons dataset, without loss of performance on the more general SNLI dataset.

This result also verifies that the heuristics we find in the original InferSent are an ecologically rational response to a training environment that licenses these 'shortcut' strategies, and not because of shortcomings in representational or learning abilities of the model itself. This points to the benefits of understanding the learning environment in greater detail, and potentially including specially designed data to guard against incorrect heuristics that don't generalize. Research on the generation of adversarial examples targets this intuition. The idea is to have a separate 'adversarial' model that generates edge-case training examples optimized to try and fool the main model into giving the wrong answer[136,394]. It does so by generating examples that violate the heuristics the main model has learned from training thus far. Subsequently, the training environment for the model is augmented to include these edge cases making the current heuristics no longer ecologically valid. The main model therefore

updates its representations and decision rules accordingly and the process is continued. Our work provides some insight into how we can leverage a top-down understanding of the structure of language and systematic stimulus design, to generate such edge-case training data and potentially improve the representations learned by machine learning systems.

A key hurdle for the scalability for such augmentation as a solution to improving artificial representations of language however is that there are an infinite number of possible stimuli, with brand new combinations of words that may never have been encountered before. No finite amount of augmentation will allow a system to represent and process this infinite space of natural language sentences unless it can also *generalize* its knowledge gained from the examples observed thus far to new examples. In this section we saw that InferSent can generalize rules like those in Table 5.1 to never previously observed combinations of X, Y and Z to perform well on the test set of the Comparisons dataset. In the following sections we further discuss the generalization capacities of the representations learned by InferSent, and focus in particular on their differences and similarities to human generalization.

5.6 Generalization

An important and well-studied aspect of human-like representations is that rules learned with one set of tokens can be systematically generalized to other tokens[102,209]. In the section on 'zero-shot reasoning' we study if our machine-learned representations can perform such generalization to tokens that have never previously been observed. More often however, the tokens to which we want to generalize learned rules have previously been observed, but simply in a different context. The historical contexts of tokens can determine some of their properties – like syntactic category, and semantic content – which in turn inform how humans generalize rules to them, sometimes deviating from entirely systematic generalization. In our section on 'context-tying', we examine how the historical context of tokens influences systematic generalization in our machine-learned representations, and how these effects compare to those in humans.

Throughout this section, we will only consider sentence pairs that are similar in structure to ones in

our Comparisons dataset, and will no longer consider performance on SNLI. We will predominantly be studying the model that has been trained jointly with our Comparisons dataset in addition to SNLI (referred henceforth to as the augmented-InferSent model).

5.6.1 Zero-shot reasoning

Zero-shot reasoning is the ability to solve tasks involving a term that has never been seen before. This (often also called zero-shot learning) has commonly been used as a test for systematicity [211] – a human can carry out inferences like "Anne is more boffy than Bob" entails that "Bob is less boffy than Anne" without ever having encountered the word "boffy" before.

But this ability requires the representation learned to be abstract, and not be tied to the Xs, Ys, and Z's seen in training. Instead it has to encode encode an abstract relational rule where "X is more Y than Z" entails "Z is more Y than X" for all possible X, Y and Z, irrespective of their specific values. If the representation are tied to the observed values of Xs, Ys and Zs and cannot generalize to new values for these, each possible X, Y and Z has to have occurred in the training dataset. However, these can be arbitrarily complex (e.g., "The old woman with a flower in her hair *is more deliriously happy than* the tall young man wearing the blue bowler hat" implies that "The tall young man wearing the blue bowler hat *is less deliriously happy than* the old woman with a flower in her hair"). Ensuring that every possible such X, Y and Z have been seen in the training data is impossible, and this kind of generalization is key to human-like language understanding.

In this section we consider the performance of the augmented-InferSent model. We already know that this model performs well on both SNLI, and generalizes to new combinations of X, Y, and Z in our Comparisons dataset (see Table 5.10), where each X, Y and Z have previously been seen. In this section, we analyze its ability to generalize to 3 different kinds of Xs and Zs that have never been encountered during training.

- Held out nouns: Nouns (from the GloVe dataset) that never occur in the training data (neither SNLI nor our Comparisons dataset).

- Made up "words": Directly using a 300 dimensional vector randomly sampled from an uncorrelated Gaussian distribution, as a stand-in for a real GloVe vector.
- Long noun phrases: The Xs and Zs used in training as part of the Comparisons dataset were of the type "the man". Here we generate longer noun phrases of the form "the grumpy man in front of us" consisting of randomly sampled adjectives, nouns and prepositional phrases.

For each sub-type in the Comparisons dataset (same, more-less and not types), we generated a test set of 1,000 sequences by substituting Xs and Zs of the above kinds. The Ys were sampled in the same way as in the Comparison dataset (random adjectives that appear in SNLI). We then tested on these sentences, and reported the average accuracy. Note that not only had these specific sentences (combinations of X, Y and Z) never been seen during training, even the individual Xs and Zs had not been seen. We found that InferSent generalizes to all three new kinds of Xs and Zs quite well (Table 5.11). The held-out nouns are the most similar to the Xs and Zs seen during training since they are also exactly one word, and are nouns sampled from GloVe. It is notable that generalization performance with these is comparable to that with the very different kinds of Xs and Zs such as the made-up words, or longer noun phrases, indicating a fairly abstract representation of relational rules that are not tied to the specific value of X and Z.

Test set	InferSent (%)	augmented-InferSent (%)
Held-out nouns	47.9	82.0
Made up words	48.0	83.2
Long noun phrases	49.1	84.9

Table 5.11: Zero-shot reasoning: Performance on previously unobserved Xs and Zs.

This indicates that the representation learned by augmented-InferSent is partially abstract and composable, allowing some systematic generalization to a variety of Xs and Zs that have never been seen before. In the next section we further probe contextuality of generalization and how that interacts with the training set / learning environment, making comparisons to human generalization.

5.6.2 Context Tying

We saw in the previous section that augmented-InferSent has some of the the central human-like capacity of zero-shot reasoning. This indicates some systematicity in its representations. However, even humans do not always succeed at fully systematic generalization. In this section we investigate these exceptions and qualifications to the widest interpretation of systematic generalization, focusing on the role of context in generalization. We do this in two ways: using type violations and biased exposure.

Type violations

One extreme of learning a purely abstract rule like in Table 5.1 is to be completely insensitive to any properties of the Xs, Ys and Zs, and generalize this rule to all possible tokens. However, this very strong generalization may not always match human intuitions. For example the sentence pair

```
Premise: The punctual is more cheerful than the man
Hypothesis: The punctual is not more cheerful than the man
```

does not seem to have a right answer. The rule applies easily only to Xs, Ys and Zs that are of the right type – in this case the right syntactic category.

While syntactic structure is not directly provided to the embedding model, some notion of syntactic category will be implicit. Information about the syntactic category of a word can be gleaned from its contexts, i.e. the other words around it[48,293,344], and in some cases can be decoded from word embeddings directly[280].

We investigated generalization of rules in augmented-Infersent to test items which, unlike in the previous section, had been previously seen, but had only occurred in a different syntactic role (i.e., a different context). We generated a test set of ungrammatical sentences using Xs and Zs that are random non-nouns, in our case random adjectives from SNLI. Crucially, these words had been seen before, but never in the position/context that X and Z occupy in the Comparisons dataset, since appearing in those positions violates syntax. We then evaluated the performance of the augmented-InferSent model in the same way as in the previous section on zero-shot generalization. We found that accu-

racy on such sentence pairs is low, giving poor performance (Table 5.12). This indicates that the rules learned, though at least partially abstract as indicated by generalization to held-out nouns, come with restrictions on the type of (known) items they will apply to. This follows closely how humans generalize – that learned rules don't generalize indiscriminately to all tokens, but rather only within some fixed categories. These categories in turn, like syntactic categories, can be gleaned from the contexts in which these tokens usually appear. In the next section, we examine the role of semantic content in the context of tokens, and how that influences generalization.

Test set	InferSent (%)	augmented-InferSent (%)
Held-out nouns	47.9	82.0
Non-noun words	47.9	49.3

Table 5.12: Type violations: Performance with tokens from the wrong syntactic category, versus with held-out tokens from the right syntactic category

Biased exposure

In this section, we manipulate the context of various tokens, without violating the syntactic rules, to study its effect on generalization. In all the augmentations we have used so far, some token X is equally likely to occur in the context of a same-type sentence pair as it is in the context of a more/less-type sentence pair. Similarly, X is as likely to occur in the context where it is 'more cheerful than the man' as it is to occur in the context where it is 'less cheerful than the man'. Therefore, apart from the restrictions of syntactically correct placement, there is no additional structure around which contexts which tokens occur in – they are all randomly distributed. However, in the real world, tokens are not uniformly sampled into contexts even within constraints of syntax; a word is much more likely to be sampled repeatedly in certain contexts than others. This is because the appearance of tokens in naturally occurring sentences is not determined solely by their syntactic role, but also by their semantic role. For example, one is more much likely to encounter the sentence "broccoli is more nutritious than candy" than the sentence "candy is more nutritious than broccoli", since one is true of the real

world, and the other is not. Nonetheless, the premise "candy is more nutritious than broccoli" still logically entails the hypothesis "broccoli is less nutritious than candy". Statistics of how often certain implications and inferences are made in the learning environment (that will be reflected in semantic beliefs about the real-world) can interfere with such logical inferences in humans in both deductive[89] and probabilistic[91] reasoning. This is often termed 'belief bias'.

In this section, we test if the representations we are studying exhibit belief bias. We manipulate the uniformity in the co-occurrence of tokens with contexts (subject to syntactic constraints), and examine if a newly augmented InferSent model can generalize a token it has seen in one context, to cases where it appears in a different context. We compare this to a zero-shot control condition, where the test token has never been seen before.

To this end, we first generated variants of our Comparisons dataset where tokens are no longer uniformly sampled into contexts. We considered only two sub-types of the comparison types summarized in Table 5.1: the *same-type* (C_2) and the *more/less-type* (C_1). These consist the two contexts C_1 and C_2 in which tokens can appear. Noun phrases were generated using the same procedure used for the long noun phrases in the section on zero-shot reasoning—phrases (tokens) of the form "the grumpy man in front of us". These tokens were then randomly divided into T^0-type and T^*-type (460 each). Therefore there is no structural difference between the the T^0 and T^* tokens, only the context in which they are seen will differ across conditions.

We built four sets of sentence pairs that vary in their context-token combination: C_2T^0 consisted of combinations of T^0 tokens in a C_2 context, so on and so forth for C_2T^*, C_1T^0, and C_1T^*. Each such context-token combination set was independently divided into train and test sets (each of size 5000). The sentence pairs in each of the four test sets had never been seen before in any of the four training sets.

We augmented the original InferSent embeddings with different combinations of samples from the four different train sets.* We then compared their performance on all four of the test sets to examine

*In this experiment we only make comparisons between the performances of differently augmented models,

how different context-token combinations seen during training influenced test generalization. The three different embeddings that result are as follows:

- Zero-shot control condition: Only the T^0 tokens were seen in training; no T^* token were seen at all. Therefore testing with tokens from T^* is analogous to zero-shot reasoning. The training set consisted of the full training sets from C_1T^0 and C_2T^0.
- Experimental conditions: Both T^0 and T^* tokens were seen in training, therefore testing with tokens from T^* is not analogous to zero-shot reasoning. However, the contexts in which T^0 and T^* tokens appear during training differed. There are two different embeddings we trained of this kind.
  - Exposed–C_1T^* : This embedding saw T^0 tokens in both C_1 and C_2 contexts (as with the control condition), and additionally also saw T^* tokens – but only in the C_1 context. In order to balance the number of training examples from each context between conditions, the training set consisted of the full training sets from C_2T^0 and half (randomly selected) of the training set from each of the C_1T^0 and C_1T^* context-token combination sets.
  - Exposed–C_2T^* : This embedding saw T^0 tokens in both C_1 and C_2 contexts, but saw T^* tokens only in the C_2 context. The training set was balanced across contexts here as well.

Test set	Performance (%)		
	Zero-shot	Exposed–C_1T^*	Exposed–C_2T^*
$C_1T^0 + C_2T^0$	97.44	97.02	98.0
C_1T^*	95.72	99.7	61.16
C_2T^*	95.78	67.71	99.96

Table 5.13: Biased exposure: Results from InferSent embeddings augmented with different training sets that manipulate the co-occurrence of context and token.

rather than considering the overall performance like in previous experiments. The influence on performance from the SNLI training data is irrelevant since it will affect all four augmented models equally. Therefore we can neglect SNLI performance and carry out our experiments using fine-tuned augmentation rather than full retraining (see the Section 'Augmenting the learning environment' for details on these). This is computationally a lot cheaper.

All three models received the same number of training examples, with equal numbers of sentence pairs from both contexts C_1 and C_2. They all also saw T^* noun phrases appear in both contexts. The three models only differed in which contexts T^* noun phrases appeared during training. The control model never saw T^* noun phrases, Exposed–C_1T^* only saw them in the C_1 context and Exposed–C_2T^* only saw them in the C_2 context. All of these were then tested on the same held-out test set. We see from Table 5.13 that all three models generalize well to held-out test examples involving previously unobserved combinations of T^0 noun phrases in both contexts (first row). This is consistent with our initial results on augmentation (see section 'Augmenting the learning environment'). Further, the control (zero-shot reasoning) condition that never saw T^* noun phrases in training generalizes well to all the test examples with T^* noun phrases (first column). This is consistent with our results on zero-shot generalization (see section 'Zero-shot reasoning').

We now turn to generalization performance when tokens were seen before but only in a specific context (second and third columns in Table 5.13). We discuss the results for the model Exposed–C_1T^* (that saw T^* noun phrases in C_1 type comparisons), a symmetric discussion applies also to Exposed–C_2T^*. W found that Exposed–C_1T^* performs well on held-out test examples from the C_1T^* category (99.7 %) – as consistent with our original experiments with augmentation. However, we found that it fails to generalize very well to T^* type noun phrases in the C_2 context, with a significant drop in performance (67.71 %). The crucial comparison is that this low performance is also significantly worse that that of the zero-shot control on the same test set (95.78 %). Neither of these have seen T^* phrases in the C_2 context – yet the control generalizes very well, while the Exposed–C_1T^* fails to. This indicates that while the representations learned can generalize well to previously unseen tokens, this generalization is poorer to tokens that have in fact been seen before, but only in a different context.

This indicates that our representations do learn something akin to belief bias, where the context in which tokens have been seen (even within the right syntactic category) can influence how abstract logical rules (like in Table 5.1) generalize to them. This suggests potential directions for research on modeling how belief bias in humans arises. However, it is crucial to point out that although humans

do exhibit such context tying, the effects are mostly observed in children[92] and under time pressure / cognitive load[90]. The co-existence of such a fast heuristic strategy (that potentially suffers from belief bias), and a slower deliberative strategy (that can perform abstract reasoning) is a well-studied and popular model for representations and decision rules in humans[90,188,152]. Thus, although people have a tendency towards belief bias, they are able to overcome it and engage in abstract reasoning, which our machine-learned representations cannot do.

This raises a new concern about the scalability of augmentation as a general approach to learning systematic representations in such tabula rasa machine-learning systems. There are infinitely many possible sentences that all follow the rules of syntax, so observing tokens in contexts that one has not often seen them in, but where they are syntactically valid, is likely to occur often. Our new findings show that while zero-shot reasoning to previously unobserved tokens works in certain cases, these tabula rasa systems may tie an observed token to the small fraction of contexts in which it has been seen. This hinders generalization to cases where this token occurs in a new context. In order for every token to have been observed in every context, a combinatorially large amount of augmented training data would be required, potentially making this approach unfeasible for achieving the kinds of systematic representations humans have.

5.7 Discussion and Future Work

In this paper, we carried out a case study in the use of methods from cognitive science and psycholinguistics to better understand machine-learned representations. We developed minimal cases in a natural language inference task that test for some aspects of abstract relational structure in sentences. We used this diagnostic tool on large-scale state-of-the-art sentence embeddings[51] to not only demonstrate its lack of abstract composable structure, but also provide insight into the representations and decision criteria actually learned. This approach led us to isolate the use of some simple heuristics, which we then traced to structural regularities in the training distribution. This allowed us to demonstrate the strong effect the training data has on the representations learned. We then augmented this training

environment with so-called adversarial examples such that simple heuristics like the ones we found are no longer ecologically valid. We found that such augmentation leads the system to learn some forms of abstract relational structure. Notably, we found that one of the traditional holy grails of systematicity —zero-shot generalization of learned rules to new, previously unseen words—can be partially achieved using appropriate augmentation. Further tests, however, revealed limitations to the breadth of this generalization. We found that while zero-shot generalization to previously unseen words works, generalizations to words that have previously been seen in a different context, suffers. This gives us another measure for the extent of systematicity in representations—a phenomena we call 'context-tying'. We discussed the relationship between this effect and findings in human cognitive psychology where semantic beliefs about the real-world can interfere with flexible inferences supported by abstract logical representations[88]. This parallel suggests new ways to model this psychological phenomena[59]. The presence of context-tying in the machine-learned representations indicates that combinatorially large amounts of augmentation will likely be required for a tabula rasa unstructured neural network model to learn an entirely systematic representation from data.

These results suggest many directions for future work. We showed how the issue of context-tying bodes poorly for the scalability of using only training set augmentations to achieve human-like systematic representations. Recent work, however, suggests such adversarial mechanisms in the human brain[114]. This motivates further research on how this approach might be made more scalable. We studied the representations learned from a fixed amount of augmentation and training. An important step forward is to better understand how systematicity in these representations evolves over the course of augmented training, and exactly how much augmentation is really needed. Another important problem is to understand what augmentations work best. To that end, a promising direction is to integrate our approach, where augmentations are generated using existing knowledge about analogous representations in humans, with approaches that learn to generate such adversarial augmentations[191,136,394].

Human infants are not as tabula rasa as models like InferSent but rather encode useful inductive biases[249,278,49,224,323]. Building such biases into our models[211,106,79,13] is a promising direction to-

wards scalably learning systematic representations. We also showed how analysis and controlled testing for heuristic strategies in the learning environment can provide rich insights into the representations learned. Such analyses could also be used to improve learning and subsequent performance by leveraging this underlying structure[336,337,133,239]. Finally, we leverage methods from cognitive psychology to introduce a new structured test dataset (the Comparisons dataset) as well as a new metric (context-tying) for sentence representations. Rather than the traditional single-dimensional metrics of the accuracy achieved on ad-hoc test datasets, our approach provides insights into the kinds of mistakes made and therefore a more principled and nuanced ways to benchmark artificial systems against humans[386,234,208,226,243,134]. A metric like context-tying is not bound to the domain of language, and can also be used to benchmark systematicity in other domains that benefit from abstract compositional representations – like scene understanding[267,182] or structured planning[43,338]. Future work should pursue other such diagnostic metrics, to build towards a comprehensive suite of testable criteria for exactly what constitutes human-like representations, and also to further inform which aspects of these we wish to emulate in artificial systems.

6

Learning amortized alorithms in machines

Discovering and exploiting the causal structure in the environment is a crucial challenge for intelligent agents. However, there is much debate about the origins and form of such causal reasoning in natural intelligence. Here, we investigate the emergence of causal reasoning and intervention strategies from simpler reinforcement learning algorithms using a meta-learning framework. We find that agents learn strategies that effectively probe, uncover, and leverage the specific kinds of causal structure in their environment to perform causal reasoning in related, held-out tasks in order to obtain rewards, select informative interventions, draw causal inferences from observational data, and make counterfactual predictions. Empirical findings in human behavioural research suggest interesting connections between our model and the development and implementation of causal reasoning in humans. This work also lays the groundwork for causally directed, structured exploration in reinforcement learning, using agents that can perform and causally interpret experiments, and advances the research program

of learning causality from statistical structure.

6.1 Introduction

Real-world situations often require us to reason about cause and effect. Although causal reasoning has commonly been touted as an essential component of natural intelligence, characterizing these abilities in humans and understanding how they emerge and develop through childhood are still active areas of research in cognitive science and psychology[378,46].

Empirical work in human developmental research suggests that causal knowledge, and the ability to acquire and exploit it, does not reflect the operation of some general and innate algorithm, but instead emerges through learning[318,246,32,44]. Evidence from studies in adult causal reasoning also show that this acquired theory is not entirely normative, and is instead graded and often tends towards associative reasoning[294,295,94,96]. Further, these observed behavioral patterns are not consistent and show significant variation depending on mechanisms[227], and exposure[207]. The theory that causality is learned from experience offers a potential explanation for these findings – different experiences potentially support different kinds and extents of causal reasoning, and exact normative causal inference may not universally be the best adaptation to all aspects of the world humans operate in.

This gives rise to the question of what learning mechanisms allow causal understanding to be acquired from experience. In this work, we demonstrate how causal reasoning can arise in agents trained using meta-learning simply through interaction with environments that contain causal structure. In particular, we use a "meta-reinforcement learning" framework[78,379]. We chose reinforcement learning (RL) as the base learning paradigm since RL is based on interactions of an agent with the environment through actions. This allows for *interventions* which are an essential part of causal reasoning. This methodology has also been shown to give rise to complex policies that exploit structure in the task distribution, such as negotiating the explore-exploit trade-off in bandits[379,380], using episodic memory[302], and amortizing Bayesian filtering to solve sequential problems[271].

A key prediction of learning causality from experience, like in our framework, is that the (causal)

inference algorithm learned should reflect the structure of the environment and the data received by the agent. If normative causal reasoning provides an advantage, and is possible given the observed data and the structure of the environment, then an agent should be able to learn it. However, other kinds of experiences might lead to different algorithms that vary on the spectrum of how 'causally-aware' they are. In this paper, we test these predictions in 5 experiments. We see that architecturally identical agents can learn different strategies for reasoning about causal structure depending on the kinds of experiences gathered during training.

Finally, formal approaches to causal identification (determining the causal graph from data) often require large amounts of data[109,345,371], and inference in the constructed causal graphs is also computationally expensive[186]. In real-world environments, humans operate under time, data, and resource constraints, dealing with uncertainty in model structure as well as non-stationarity. Agents that learn aspects of the learning algorithm directly from experience will adapt to statistical structure in their specific environment and task, and could utilize useful abstract priors (or inductive biases) from other episodes that can be difficult to formally specify. Such adaptations amortize much of the computation over previous experience and could allow better performance than formal approaches under ecological constraints e.g.[60,120,130,216,361].

The purpose of this work is not to propose a new algorithmic solution to causal inference per se. Rather, we argue that our meta-learning approach has compelling links to human causal reasoning in terms of a) how a theory of causality could be learned, b) the graded notion of causality in humans, and c) resource efficiency by meta-learning inductive biases. Resource efficient causal inference based on leveraging statistical structure, is also useful for and an active area of research in machine learning e.g.[19,163,231,273,250].

6.2 Related Work

Goodman et al.[138] demonstrated how an abstract notion of causality in humans can be learned from experience, with hierarchical Bayesian inference. Our approach is similar to this as meta-learning can

also be framed as hierarchical Bayesian inference[140]. However, these approaches provide complementary advantages. While formal theory learning (as in[138]) is systematic and generalizes across domains, it requires the pre-specification of discrete primitives and an expensive zero order (stochastic search) optimization to learn the correct theory built from these primitives[321,36]. A restrictive choice of primitives limits the space of possible theories, while a generous choice makes the optimization very expensive. This approach also leaves open the question of the origin of these discrete primitives and how they might be plausibly implemented in the brain. Our method avoids these assumptions and instead uses a first order (gradient-based) optimization method that leverages learning signals from the environment, thus discovering emergent structure directly from experience[241]. Since our model is implemented with a deep neural network, which can be universal approximators[332,175], it can implement different graded causal theories that don't conform to purely normative accounts, in a neurally-plausible distributed representation. This could give rise to graded causal reasoning behaviors analogous to those seen in humans[294,295,94,96] .

Bengio et al[19] propose a meta-learning approach to utilize explicit, pre-specified statistical properties of interventions to isolate and disentangle causal variables in a supervised learning setting. Our work shows how a spectrum of 'causally-aware algorithms' can arise from utilizing several different kinds of implicit, unspecified statistical structure in the environment. Our reinforcement learning approach further allows the agent to directly interact with the environment to also simultaneously learn an experimental policy that utilizes this underlying structure. Denil et al[67] showed that deep reinforcement learning agents can learn to perform actions to gain knowledge about latent, physical properties of objects, but do not explore explicit causal inference.

6.3 Problem Specification

Our goal is to demonstrate that causal reasoning can arise from meta-reinforcement learning. Further, we demonstrate that depending on the kinds of data the agents see during training, the kind of causal reasoning learned varies. Our agents learn to leverage statistical structure in different kinds of available

information, to carry out different kinds of causal reasoning. In this section, we first briefly formalize causal inference and how it depends on the kinds of data (See Supplementary Materials for more details).

Causal relationships among random variables can be expressed using *causal Bayesian networks* (CBNs $\mathcal{G}$)[276,345,64]. Each node X_i corresponds to a random variable, and the joint distribution $p(X_1, \ldots, X_N)$ is given by the product of conditional distributions of each node X_i given its parent nodes $\text{pa}(X_i)$, i.e. $p(X_{1:N}) = \prod_{i=1}^{N} p(X_i|\text{pa}(X_i))$.

The edges of $\mathcal{G}$ encode causal semantics: a directed path from X_c (cause) to X_e (effect) is called a causal path. The causal effect of X_c on X_e is the conditional distribution of X_c given X_e restricted to only causal paths. This restriction is an essential caveat, since the simple conditional distribution $p(X_e|X_c)$ encodes only correlations (i.e. associative reasoning). Intervening on a node X_c corresponds to removing its connection to its parent nodes $\text{pa}(X_c)$, and fixing it to some value C yielding a new CBN $\mathcal{G}_{\rightarrow X_c=C}$. The causal effect of X_c on X_e is given by the conditional distribution in this new CBN. This distribution is denoted $p_{\rightarrow X_c=C}(X_e|X_c = C)$

Different kinds of environments support different kinds of causal reasoning. It is often possible to compute $p_{\rightarrow X_c=C}(X_e|X_c = C)$ (i.e. causal reasoning) using observations from $\mathcal{G}^*$. We investigate this kind of causal reasoning in Experiment 1 (Observational Environments). However, in the presence of unobserved confounders (an unobserved variable that affects both X_c and X_e), this is, in general , no longer possible[276]. The only way to compute causal effects $p_{\rightarrow X_c=C}(X_e|X_c = C)$ in this case is by collecting observations directly from the intervened graph $\mathcal{G}_{\rightarrow X_c=C}$. In Experiment 2 (Interventional Environments), we investigate this kind of causal reasoning, by allowing agent to perform interventions on the environment. An additional level of sophistication comes from *counterfactual* environments – we provide results concerning this and other settings in the Supplementary Material.

---

*When the CBN $\mathcal{G}$ is known, this process can be formalized as do-calculus[276,277]. In our case the CBN is not directly provided, and the agent must simultaneously perform causal identification using samples from $\mathcal{G}$[163].

6.4 Task Setup and Agent Architecture

In our experiments, we use a simple framework that has some key properties relevant to ecologically realistic causal reasoning. First, the number of variables over which inference is carried out is small. Second, the amount of data available is limited. Third, agents can actively seek out information by interacting with the environment rather than only receiving passive input. This facilitates future work in drawing parallels to human causal reasoning, as well as permits a simple and clear demonstration of the effects of interest.

In each episode the agent interacts with a different CBN $\mathcal{G}$ with N variables. The structure of $\mathcal{G}$ is drawn randomly from the space of constraints described below. Each episode consists of T steps, which are divided into two phases: an *information phase* and a *quiz phase*. The information phase corresponds to the first $T-1$ steps and allows the agent to collect information from $\mathcal{G}$. Note that $\mathcal{G}$ is never directly provided to the agent, but is only observed through $T-1$ samples. Further, the agents in the different experiments are architecturally identical, and give rise to different behavior soley due to the data they receive in the information phase. The quiz phase, corresponding to the final step T, requires the agent to exploit the causal knowledge it accumulated during the information phase. In particular, the agent needs to select the node with the highest value under a random external intervention. The structure of the quiz phase is exactly the same for all agents in all experiments.

Causal Graphs, Observations, and Actions

We generate graphs that have $N = 5$ nodes and sample the adjacency matrix to have non-zero entries only in its upper triangular part (this guarantees that all the graphs obtained are acyclic). Edge weights w_{ji} are uniformly sampled from $\{-1, 0, 1\}$. This yields $3^{N(N-1)/2} = 59049$ unique graphs. These can be divided into equivalence classes, i.e. sets of graphs that are structurally identical but differ in the permutation order of the node labels. Our held-out test set consists of 12 random graphs plus all other graphs in the corresponding equivalence classes, yielding 408 total graphs in the test set. Thus,

none of the graphs in the test set (or any graphs equivalent to these) have been seen during training.

We sample each node, $X_i \in \mathbb{R}$, as a Gaussian random variable. The distribution of parentless nodes is $\mathcal{N}(\mu = 0.0, \sigma = 0.1)$, while for a node X_i with parents $\text{pa}(X_i)$ we use the conditional distribution $p(X_i|\text{pa}(X_i)) = \mathcal{N}(\mu = \sum_j w_{ji}X_j, \sigma = 0.1)$ with $X_j \in \text{pa}(X_i)$. We also tested graphs with non-linear causal effects and larger graphs of size N = 6, see the Supplementary Material for details.

A root node of $\mathcal{G}$ is always hidden, to allow for unobserved confounders, and the agent can therefore only ever see the values of the other 4 nodes. These 4 nodes are henceforth referred to as the 'visible nodes'. The concatenated values of the nodes, v_t, and a one-hot vector indicating the external intervention during the quiz phase, m_t, (explained below) form the observation vector provided to the agent at step t, $o_t = [v_t, m_t]^*$.

In both phases, at each step t, the agent chooses to take one out of $2(N-1)$ actions. The first $N-1$ actions are *information actions*, and the second $N-1$ actions are *quiz actions*. Both information and quiz actions are associated with selecting the $N-1$ visible nodes, but can only be legally used in the appropriate phase of the task. If used in the wrong phase, a penalty is applied and the action produces no effect.

Information Phase. The information phase differs depending on the kind of environment the agent is in – observational or interventional. Here, we discuss the case of the interventional environment.

An information action $a_t = i$ causes an intervention on the i-th node, setting the value of $X_{a_t} = X_i = 5$ (the value 5 is outside the likely range of sampled observations and thus facilitates learning the causal graph). The node values v_t are then obtained by sampling from $p_{\to X_i=5}(X_{1:N\setminus i}|X_i = 5)$ (where $X_{1:N\setminus i}$ indicates the set of all nodes except X_i), i.e. from the intervened CBN $\mathcal{G}_{\to X_{a_t}=5}$. If a quiz action is chosen during the information phase, it is ignored, i.e. the node values are sampled from $\mathcal{G}$ as if no intervention has been made. Furthermore, the agent is given a penalty of $r_t = -10$ in order to encourage it to take quiz actions during the quiz phase. There is no other reward during the

*'Observation' o_t refers to the reinforcement learning term, i.e. the input from the environment to the agent. This is distinct from observations in the causal sense which we refer to as observational data.

information phase.

The default length an episode is fixed to be $T = N = 5$, giving an information phase of length of $T - 1 = 4$. This episode length was chosen because in the noise-free limit, a minimum of $N - 1 = 4$ interventions, one on each visible node, is required in general to resolve the causal structure.

Quiz Phase. The quiz phase remains the same for all the different environments and agents. In the quiz phase, one visible node X_j is selected at random to be intervened on by the environment. Its value is set to -5. We chose -5 to disallow the agent from memorizing the results of interventions in the information phase (which are fixed to $+5$) in order to perform well on the quiz phase. The agent is informed which node received this external intervention via the one-hot vector m_t as part of the observation from the the final pre-quiz phase timestep, $T - 1$. For steps $t < T - 1$, m_t is the zero vector. The agent's reward on this step is the sampled value of the node it selected during the quiz phase. In other words, $r_T = X_i = X_{a_T-(N-1)}$ if the action selected is a quiz action (otherwise, the agent is given a penalty of $r_T = -10$).

Active vs Random conditions. Our agents have to perform two distinct tasks during the information phase: a) actively choose which nodes to act on and b) perform casual reasoning based on the observations. We refer to this setup as the "active" condition. To better understand the role of (a), we include comparisons with a baseline agent in the "random" condition where the environment ignores the agents actions and randomly chooses a visible node to intervene upon at each step of the information phase. Note again that the only difference between agents in these two conditions is the kind of data the environment provides them.

Two Kinds of Learning. An "inner loop" of learning occurs within each episode where the agent is learning from the 4 samples it gathers during the information phase to perform well in the quiz phase. The same agent then enters a new episode, where it has to repeat the task on a different CBN. Test performance is reported on CBNs that the agent has never previously seen after all the weights of the RNN have been fixed. Hence, the only transfer from the training to test set (or the "outer loop" of learning) is a learned procedure for collecting evidence in the information phase to perform well in the

quiz phase. Exactly what this learned procedure is will depend on the training environment. We will show that this learned procedure can include performing different kinds of causal inference, as well as active information gathering. See the Supplementary Material for more details on meta-learning.

Agent Architecture and Training. For the agent, we use a long short-term memory (LSTM) network[174]. At each time-step t, the LSTM receives a concatenated vector containing $[o_t, a_{t-1}, r_{t-1}]$ as input. The LSTM outputs are a softmax over the number of actions and a scalar baseline. Learning is done by *asynchronous advantage actor-critic*[252]. See Supplementary Material for further details. In all experiments, the agent is tested with the learning rate set to zero using a held-out test set as discussed above.

6.5 Experiments

Our two experiments (observational and interventional environments) differ in the properties of the node values v_t that are observed by the agent during the information phase. This also limits the kinds of causal reasoning possible within each environment. We measure agent performance using a function of the reward earned in the quiz phase for held-out CBNs. Choosing a random node in the quiz phase results in an expected reward of $-5/4 = -1.25$ since one node (the externally intervened one) always has value -5 and the other nodes have on average 0 value. By learning to simply avoid the externally intervened node, the agent can earn on average 0 reward. Since the quiz phase requires the agent to predict the outcome of a previously unobserved intervention, consistently good performance on this task in general requires the agent to perform causal reasoning. We will see that performance reflects different extents of causal reasoning and depends on the kinds of environments agents experience. The rewards are normalized by the maximum possible reward achievable with exact causal reasoning on that test set. Henceforth, we refer to this measure as the "(normalized) performance". The maximal possible reward is calculated by computing the true maximum mean value among all the nodes in $\mathcal{G}_{\to X_j}$, where X_j is the node externally intervened upon in the quiz phase. We train 8 copies of each agent and report the average performance across 1632 episodes (408 held-out test CBNs,

with 4 possible external interventions). 95% confidence intervals are indicated by the error bars.

6.5.1 Experiment 1: Observational Environments

In Experiment 1, the agents are in an environment that does not permit any interventions during the information phase, agents only receive observations from $\mathcal{G}$. This corresponds to passively observing the world. This setting permits some limited causal reasoning as outlined in Section 6.3, and we sought to test if our agents can learn this. We examine this hypothesis by comparing agent performance to that of an "Associative Baseline", i.e. the performance obtained by only using correlations in the environment. Two other standard RL baselines are included in the Supplementary Material.

In this experiment, we tested 4 agents: "Observational", "Long Observational", "Active Conditional" and "Random Conditional". All the agents have the same architecture and employ the same learning algorithms. The only difference between them is the kind of data they have access to.

Observational Agents: In the information phase, the actions of the agent are ignored. The agent always receives the values of the visible nodes sampled from the joint distribution associated with $\mathcal{G}$. In addition to the default $T = 5$ episode length, we also trained this agent with $4\times$ longer episode length (Long Observational Agent) in order to measure performance when the agent has access to more data.

Conditional Agents: In this case, agents are still not allowed to interact with the environment via interventions, but they are given access to more informative observations. Specifically, the information phase actions correspond to observing a world in which the selected node X_j is equal to $X_j = 5$, and the remaining nodes are sampled from the conditional distribution $p(X_{1:N\setminus j}|X_j = 5)$. This differs from intervening on the variable X_j by setting it to the value $X_j = 5$, since here we take a conditional sample from $\mathcal{G}$ rather than from $\mathcal{G}_{\to X_j=5}$. Therefore, this agent still has access to only observational data, but receives more informative data, since it can observe samples far outside the likely range of observations. We run active and random versions of this agent as described in Section 6.4. Comparing these two settings allows us to disentangle whether the agent can learn to exercise control over what

data it wishes to observe from the environment by accordingly choosing informative observations.

Associative Baseline: This baseline receives the true joint distribution $p(X_{1:N})$ implied by the CBN in that episode and therefore has knowledge of the correlation structure of the environment. In the quiz phase, this baseline acts solely on this correlational information and chooses the node that has the maximum value according to $p(X_j|X_i = -5)$ with X_i the node externally intervened upon.

Results

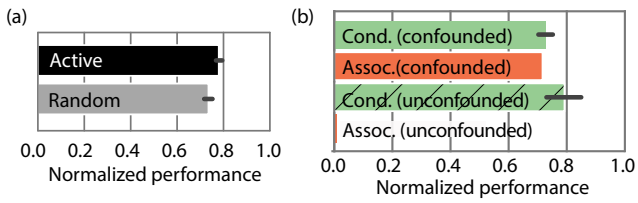


Figure 6.1: a) Active vs Random Conditional, b)Associative Baseline vs Active Conditional, where intervened node has a parent

The different agents in this experiment are given access to different kinds of data from the same underlying causal structure during the information phase. We are interested in understanding if agents learn to leverage this information to perform well on the quiz phase. The main conclusion we reach is that, when given access to informative observations, our agents can learn to perform a form of causal reasoning using observational data. The Associative Baseline tracks the best performance that can be achieved using only knowledge of correlations i.e. without causal knowledge. The Active-Conditional Agent outperforms this baseline by a non-trivial margin (Figure 6.2a).

To further demonstrate that this improvement is indeed due to causal reasoning, we partition the test cases by whether or not the node that was intervened on in the quiz phase has a parent (Figure 6.2b). If the intervened node X_j has no parents, then $\mathcal{G} = \mathcal{G}_{\to X_j}$, and doing causal reasoning should afford no advantage over doing associative reasoning. Indeed, the Active-Conditional Agent performs better than the Associative Baseline only when the intervened node has parents (hatched bars in Figure 6.2b). In Figure 6.2c, we show the quiz phase for an example test CBN. This highlights that the Associative Baseline chooses according to the node values predicted by $p(X_{1:N\setminus j}|X_j = -5)$, whereas the Active-Conditional Agent chooses according the node values predicted by $p_{\to X_j=5}(X_{1:N\setminus j}|X_j = 5)$).

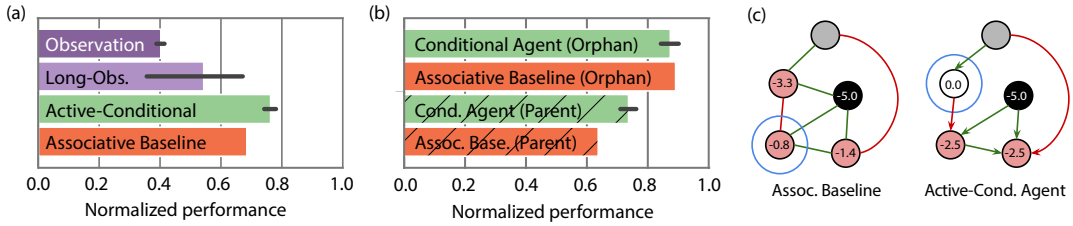


Figure 6.2: Experiment 1. Agents do causal reasoning from observational data. a) Average performance of the agents tested in this experiment. b) Performance split by the presence or absence of at least one parent (Parent and Orphan respectively) on the externally intervened node. c) Quiz phase for a test CBN. Green (red) edges indicate a weight of $+1$ (-1). Black represents the intervened node, green (red) nodes indicate a positive (negative) value, white indicates a zero value. The blue circles indicate the agent's choice. Left panel: The undirected version of $\mathcal{G}$ and the nodes taking the mean values prescribed by $p(X_{1:N\setminus j}|X_j = -5)$, including backward inference to the intervened node's parent. The Associative Baseline's choice is consistent with maximizing these (incorrect) node values. Right panel: $\mathcal{G}_{\to X_j=-5}$ and the nodes taking the mean values prescribed by $p_{\to X_j=-5}(X_{1:N\setminus j}|X_j = -5)$. The Active-Conditional Agent's choice is consistent with maximizing these (correct) node values.

Comparing the performances of the Active and Random versions of the Conditional Agents, we find that the active Agent's performance is slightly but significantly ($p = 0.003$, Figure 6.1a) higher than the Random Agent. This indicates that when permitted, the agent learns to generate informative observations. We also trained a third agent that employs the optimal information gathering policy in the noise-free limit (acting on each visible node exactly once), and obtained a performance slightly but significantly ($p = 0.008$, not shown) higher than the Active agent (although still significantly less than optimal causal reasoning), indicating that the policy learned by the Active Agent is not optimal. But the differences between the performances of agents with different information gathering policies is very small, indicating that learning a data-collection policy does not yield a critical benefit when receiving conditional samples in this small-data regime.

Agents that receive unconditional observations from $\mathcal{G}$, i.e. the Observational Agents ("Observation" and "Long-Obs" in Figure 6.2a) perform worse than the Active-Conditional Agent. Note that this is to be expected since these agents receive less diagnostic information during the information phase. However, the Observational agent is still able to leverage the information from the 4 unconditional samples it receives and perform better than the random baseline. Further, when given access to more data (the Long-Obs. agent) the same agent learns to utilize it, yielding better performance.

From Figure 6.2a, we see that while the Active-Conditional Agent performs significantly above the Associative baseline, it far from the performance utilizing full causal reasoning (= 1.0 on our scale). From Figure 6.2b, we see that this gap is driven mostly by test cases where the intervened node has a parent. While the Active-Conditional Agent's advantage over the baseline comes from these test cases, it is still not performing optimally on them. We hypothesize that this is due to the presence of unobserved confounders. As discussed in Section 6.3, full causal inference in the presence of confounders is in general not possible with just observational data. To further investigate this hypothesis, we partition the set of test cases into those where the intervened upon node has a confounded parent and those with unconfounded parents (Figure 6.1b). We see that the performance of the Active-Conditional Agent is significantly higher than the Associative baseline only in cases where the parent is not confounded. Causal inference in the presence of confounders is only in general possible with interventions. In the next experiment, we discuss the performance of our agents in an environment that permits interventions.

We also note that the Associative agent has higher performance when the parent of the intervened node is confounded than when it isn't (where the performance is not significantly above zero). This could point to other statistical structure in the environment – for example, if the intervened node has more *visible* parents (as is true for the graphs with unconfounded parents in Figure ??), there are more visible nodes strongly correlated with it due to (incorrect) backward inferences from child to parent. This could hinder the associative agent giving lower performance. These findings highlight that there are often unexpected statistical trends even in putatively formal settings like our distribution of simple CBNs, that could potentially be leveraged[179,176] by meta-learning agents.

6.5.2 Experiment 2: Interventional Environments

In this experiment, we test if agents can learn to perform causal inference from interventions. In particular, we are interested in performance in the presence of confounders. The interventional environment allows the agent to intervene on any visible node during the information phase. The agent's

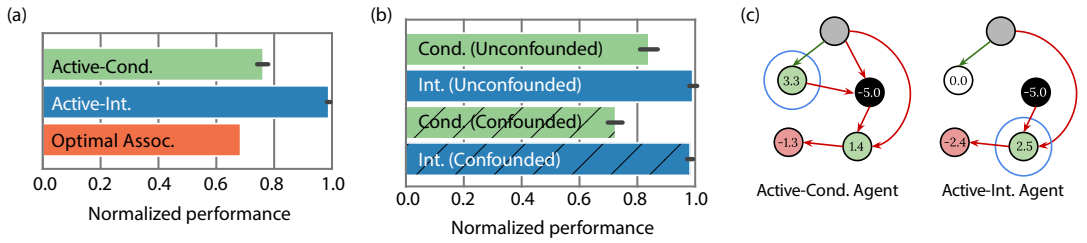


Figure 6.3: Experiment 2. Agents do causal reasoning from interventional data. a) Average performance of the agents tested in this experiment. See main text for details. b) Performance split by the presence or absence of unobserved confounders (abbreviated as Conf. and Unconf.). c) Quiz phase for a test CBN. See Figure 6.2 for a legend. Here, the left panel shows the full $\mathcal{G}$ and the nodes taking the mean values prescribed by $p(X_{1:N\setminus j}|X_j = -5)$. We see that the Active-Cond Agent's choice is consistent with choosing based on these (incorrect) node values. The right panel shows $\mathcal{G}_{\to X_j=-5}$ and the nodes taking the mean values prescribed by $p_{\to X_j=-5}(X_{1:N\setminus j}|X_j = -5)$. We see that the Active-Int. Agent's choice is consistent with maximizing on these (correct) node value.

actions correspond to performing an intervention on the selected node X_j and sampling from $\mathcal{G}_{\to X_j}$ (see Section 6.4). As discussed in Section 6.3, access to interventional data permits causal reasoning even in the presence of unobserved confounders, a feat in general impossible with access only to observational data. We test both active and random versions of the agent (see Section 6.4) to disentangle if the agent can also learn to select informative interventions when the environment permits.

Results

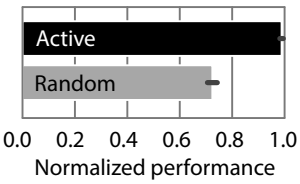


Figure 6.4: Active and Random Interventional Agents

The agents tested in this experiment differ from agents in previous experiments only in the kind of data that they have access to. We see in Figure 6.3a that the Active-Interventional Agent's performance is better than the Active-Conditional Agent, achieving close to optimal performance. This shows that when given access to interventions, the agent learns to leverage them to perform causal reasoning. Partitioning the test cases by whether any node has unobserved confounders with other nodes in the graph (Figure 6.3b), we see that the Active-Interventional Agent performs close to optimal on both confounded and unconfounded test cases. This confirms our hypothesis that the agent has learned to perform causal reasoning even in the presence of confounders which the Conditional agents in Experiment 1

could not do. This is highlighted by Figure 6.3c, which shows the quiz phase for an example CBN, where the Active-Conditional Agent is unable to resolve the unobserved confounder, whereas the Active-Interventional Agent is able to do so. We also see that while the performance of the Active-Conditional Agent is significantly higher in unconfounded cases than in confounded ones, it is not as high as the performance of the Interventional Agent, even though inference in the absence of confounders is in theory within reach of the conditional agent. This could be because causal inference from observations is more challenging than from interventions, in our setting. In our framework, the final quiz phase node values are the negative (with noise) of the values observed, if the quiz phase node is intervened on in the information phase*. This makes the decoding process significantly easier than if (as with the Conditional cases), information has to be integrated across several observations in the information phase to perform well in the quiz phase. When utilizing the statistical structure of the task and environment, interventions are easy to learn from. Evidence of such behavior has also been noted in humans[96,94].

Further, we find that the Active-Interventional agent learns to utilize the control it has over what interventions it does, to choose informative interventions: its performance is significantly better than the Random-Interventional Agent (Figure 6.4). This indicates that when permitted, the agents learns a good intervention policy to generate informative data. The difference betwee Active and Random is far greater than in the Conditional case, with the Active Interventional agent reaching ceiling performance. This indicates that in our domain, while causal inference is easier from interventions than observations, it is perhaps more sensitive to the right intervention policy – learning a policy for information gathering yields a critical benefit above a random policy, when learning from interventions, in our domain.

*We demonstrate in the Supplementary Material that our agents are able to infer from interventions even in non-linear cases where the decoding is more involved.

6.6 Discussion and Future Work

Learning abstract structural information about the world that generalizes across tasks is an important component of natural intelligence underlying its flexibility and data-efficiency. In this paper, we show that causal reasoning capabilities can arise from such hierarchical structure learning (i.e. meta-learning) simply through interaction with an environment that rewards and permits causal reasoning. An important prediction of our model is that different kinds and extents of causal reasoning can arise depending on existing structure in the environment. We find that when put in different environments, our agents learn to: 1) leverage observational data to make causal inferences, 2) leverage interventions to perform causal inference in the presence of unobserved confounders, and 3) perform active-learning, i.e. actively generate informative data when the environment permits it. In the Supplementary Material, we present results that showcase our agents performing counterfactual reasoning.

Even in this simple domain, we saw evidence of unspecified, non-trivial underlying statistical structure in the environment, as well as preliminary evidence that our agents utilize it. Future work could further examine the procedures being learned and the kinds of structure being utilized. In our analyses, we compared to baselines and study behaviour on diagnostic test-sets to characterize these. Other work on statistical approaches to learning causal structure[19,179,176], as well as methods from neuroscience[380], could provide further insights into what our agents learn, which could potentially be leveraged for more efficient causal reasoning. By using an RL framework, our agents learn to take actions that produce useful information—opening up possibilities for structured exploration, and optimal experiment design. In our work, we don't address the causal grounding problem—our agents are told what the relevant variables are. Using models that are more explicitly structured e.g.[5,13,108], and more advanced architectures e.g.[169,168,87], could allow us to scale up to directly inferring more systematic representations from unstructured input, and perform a larger range of tasks.

A crucial contribution of our work is to consider causal reasoning in natural intelligence not an end in and of itself but a means to better performance on some downstream task that is easier to specify, in

a world that contains causal structure. In our case this task is acquiring reward in an RL task, but could be generalized to any other task by simply changing the meta-learning objective. This is a reasonable assumption since causal reasoning exists in humans, and even chimpanzees and rats [27,139,289] without "formal instruction" on causality itself. This assumption allows us to frame the acquisition of causal reasoning as a meta-learning problem, and we highlight how this approach could also capture many qualitative empirical findings in how causal reasoning is learned and implemented in humans.

This direction of research opens up many interesting directions in cognitive science and psychology. We focused primarily on varying the kinds of data available to the agent, but there many other ways in which the agent's experience will inform the kind and extent of causal reasoning exhibited. In this study, we uniformly sample the space of CBNs and external interventions, but ecological distributions of causal structures and queries are not uniformly distributed and vary significantly from domain to domain. Our meta-learning framework adapts to such structure in the training distribution [78,271,379] and could parallel the domain/function specificity of human causal reasoning [207,227]. Different distributions of queries can also create situations where simpler associative strategies are largely indistinguishable from full causal reasoning [361,130]. Further, in most real-world tasks, causal inference is usually not useful in and of itself, but rather for some downstream task. The reward in our study also depended only indirectly on causal reasoning. While in our task, causal reasoning is still an optimal strategy, this may not always be the case. These factors may result in different optimal strategies that vary on the spectrum of how "causally-aware" they are, and allow parallels to the graded notions of causal inference in humans [94,96,294,295].

Supplementary materials

6.7 Agent architecture

We used a long short-term memory (LSTM) network [174] (with 192 hidden units) that, at each time-step t, receives a concatenated vector containing $[o_t, a_{t-1}, r_{t-1}, m_t]$ as input, where o_t is the observation, a_{t-1} is the previous action, r_{t-1} the previous reward and m_t indicates the external intervention. The

outputs, calculated as linear projections of the LSTM's hidden state, are a set of policy logits (with dimensionality equal to the number of available actions), plus a scalar baseline. The policy logits are transformed by a softmax function, and then sampled to give a selected action.

Learning was by *asynchronous advantage actor-critic*[252]. In this framework, the loss function consists of three terms – the policy gradient, the baseline cost and an entropy cost. The baseline cost was weighted by 0.05 relative to the policy gradient cost. The weighting of the entropy cost was annealed over the course of training from 0.25 to 0. Optimization was via RMSProp with $\varepsilon = 10^{-5}$, momentum = 0.9 and decay = 0.95. Learning rate was annealed from 9×10^{-6} to 0, with a discount of 0.93. Hyperparameters were optimized by performing a coarse grid search (2-4 values) over learning rate, discount factor, and the number of hidden units in the LSTM. Unless otherwise stated, training was done for 1×10^7 steps using batched environments with a batch size of 1024, using a distributed architecture with roughly 4000 CPUs for 5 days.

6.8 Formalism for Memory-based Meta-learning

Meta-learning refers to a broad range of approaches in which aspects of the learning algorithm itself are learned from the data. Many individual components of deep learning algorithms have been successfully meta-learned, including the optimizer[7], initial weight parameters,[97], a metric space[373], and use of external memory[316].

Following the approach of[78,379], the entire inner loop of learning is implemented by a recurrent neural network (RNN), and we train the weights of the RNN with model-free reinforcement learning (RL). The RNN is trained on a broad distribution of problems which each require learning. Consider a distribution $\mathcal{D}$ over Markov Decision Processes (MDPs). We train an agent with memory (in our case an RNN-based agent) on this distribution. In each episode, we sample a task $m \sim \mathcal{D}$. At each step t within an episode, the agent sees an observation o_t, executes an action a_t, and receives a reward r_t. Both a_{t-1} and r_{t-1} are given as additional inputs to the network. Thus, via the recurrence of the network, each action is a function of the entire trajectory $\mathcal{H}_t = \{o_0, a_0, r_0, \ldots, o_{t-1}, a_{t-1}, r_{t-1}, o_t\}$

of the episode. Because this function is implemented by the neural network, its complexity is limited only by the size of the network. When trained in this way, the RNN is able to implement a learning algorithm capable of efficiently solving novel learning problems in or near the training distribution.

Learning the weights of the RNN by model-free RL can be thought of as the "outer loop" of learning. The outer loop shapes the weights of the RNN into an "inner loop" learning algorithm, which plays out in the activation dynamics of the RNN and can continue learning even when the weights of the network are frozen. The inner loop algorithm can also have very different properties from the outer loop algorithm used to train it. For example, this approach has been used to negotiate the exploration-exploitation tradeoff in multi-armed bandits[78,379], learn algorithms which dynamically adjust their own learning rates[379,380], and perform one-shot learning using external memory[316]. In the present work we explore the possibility of obtaining a causally-aware inner-loop learning algorithm.

6.9 Formalisms for causal inference

6.9.1 Causal Bayes Nets

By combining graph theory and probability theory, the causal Bayesian network framework provides us with a graphical tool to formalize and test different levels of causal reasoning. This section introduces the main definitions underlying this framework and explains how to visually test for statistical independence[275,26,202,9,256].

A graph is a collection of nodes and links connecting pairs of nodes. The links may be directed or undirected, giving rise to directed or undirected graphs respectively.

A path from node X_i to node X_j is a sequence of linked nodes starting at X_i and ending at X_j. A directed path is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

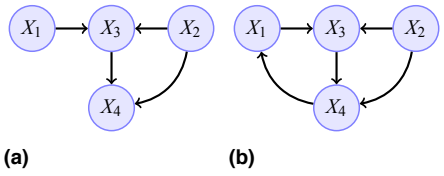


Figure 6.5: (a): Directed acyclic graph. The node X_3 is a collider on the path $X_1 \to X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \to X_3 \to X_4$. (b): Cyclic graph obtained from (a) by adding a link from X_4 to X_1.

A directed acyclic graph is a directed graph with no directed paths starting and ending at the same node. For example, the directed graph in Figure 6.5(a) is acyclic. The addition of a link from X_4 to X_1 gives rise to a cyclic graph (Figure 6.5(b)).

A node X_i with a directed link to X_j is called parent of X_j. In this case, X_j is called child of X_i.

A node is a collider on a specified path if it has (at least) two parents on that path. Notice that a node can be a collider on a path and a non-collider on another path. For example, in Figure 6.5(a) X_3 is a collider on the path $X_1 \to X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \to X_3 \to X_4$.

A node X_i is an ancestor of a node X_j if there exists a directed path from X_i to X_j. In this case, X_j is a descendant of X_i.

A graphical model is a graph in which nodes represent random variables and links express statistical relationships between the variables.

A Bayesian network is a directed acyclic graphical model in which each node X_i is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ indicates the parents of X_i. The joint distribution of all nodes in the graph, $p(X_{1:N})$, is given by the product of all conditional distributions, i.e. $p(X_{1:N}) = \prod_{i=1}^{N} p(X_i|\text{pa}(X_i))$.

When equipped with causal semantic, namely when describing the process underlying the data generation, a Bayesian network expresses both causal and statistical relationships among random variables—in such a case the network is called causal.

Assessing statistical independence in Bayesian networks. Given the sets of random variables $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{Y}$ are statistically independent given $\mathcal{Z}$ if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are closed (or blocked). A path is closed if at least one of the following

conditions is satisfied:

(i) There is a non-collider on the path which belongs to the conditioning set $\mathcal{Z}$.

(ii) There is a collider on the path such that neither the collider nor any of its descendants belong to $\mathcal{Z}$.

6.9.2 An intuitive example of cause-effect reasoning in a CBN

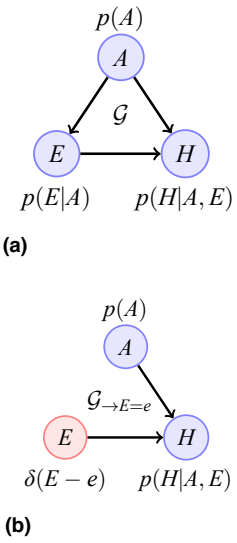


Figure 6.6: (a): A CBN $\mathcal{G}$ with a confounder for the effect of exercise (E) on heath (H) given by age (A). (b): Intervened CBN $\mathcal{G}_{\to E=e}$.

An example of CBN $\mathcal{G}$ is given in Figure 6.6a, where E represents hours of exercise in a week, H cardiac health, and A age. Random variables are denoted by capital letters (e.g., E) and their values by small letters (e.g., e). The causal effect of E on H is the conditional distribution restricted to the path $E \to H$, i.e. excluding the path $E \leftarrow A \to H$. The variable A is called a *confounder*, as it confounds the causal effect with non-causal statistical influence.

Simply observing cardiac health conditioning on exercise level from $p(H|E)$ (associative reasoning) cannot answer if change in exercise levels cause changes in cardiac health (cause-effect reasoning), since there is always the possibility that correlation between the two is because of the common confounder of age.

The causal effect of $E = e$ can be seen as the conditional distribution $p_{\to E=e}(H|E = e)^{\star}$ on the *intervened* CBN $\mathcal{G}_{\to E=e}$ resulting from replacing $p(E|A)$ with a delta distribution $\delta(E - e)$ (thereby removing the link from A to E) and leaving the remaining conditional distributions $p(H|E, A)$ and

[*] In the causality literature, this distribution would most often be indicated with $p(H|\text{do}(E = e))$. We prefer to use $p_{\to E=e}(H|E = e)$ to highlight that intervening on E results in changing the original distribution p, by structurally altering the CBN.

$p(A)$ unaltered (Figure 6.6b). The rules of do-calculus[276,277] tell us how to compute $p_{\to E=e}(H|E=e)$ using observations from $\mathcal{G}$. In this case $p_{\to E=e}(H|E=e) = \sum_A p(H|E=e,A)p(A)^*$. Therefore, do-calculus enables us to reason in the intervened graph $\mathcal{G}_{\to E=e}$ even if our observations are from $\mathcal{G}$. This is the kind of causal reasoning possible in our observational data setting.

Such inferences are always possible if the confounders are observed, but in the presence of unobserved confounders, for many CBN structures the only way to compute causal effects is by collecting observations directly from the intervened graph, e.g. from $\mathcal{G}_{\to E=e}$ by fixing the value of the variable $E=e$ and observing the remaining variables—we call this process performing an actual intervention in the environment. In our interventional data setting the agent has access to such interventions.

6.9.3 Counterfactual reasoning

Cause-effect reasoning can be used to correctly answer predictive questions of the type "Does exercising improve cardiac health?" by accounting for causal structure and confounding. However, it cannot answer retrospective questions about what *would have* happened. For example, given an individual i who has died of a heart attack, this method would not be able to answer questions of the type "What would the cardiac health of this individual have been had she done more exercise?". This type of question requires reasoning about a counterfactual world (that did not happen). To do this, we can first use the observations from the factual world and knowledge about the CBN to get an estimate of the specific latent randomness in the makeup of individual i (for example information about this specific patient's blood pressure and other variables as inferred by her having had a heart attack). Then, we can use this estimate to compute cardiac health under intervention on exercise. This procedure is called the *Abduction-Action-Prediction Method*[277] and is described below.

Assume, for example, the following model for $\mathcal{G}$ in Figure 6.6: $E = w_{AE}A + \eta$, $H = w_{AH}A + w_{EH}E + \varepsilon$, where the weights w_{ij} represent the known causal effects in $\mathcal{G}$ and ε and η are terms of (e.g.) Gaussian noise that represent the latent randomness in the makeup of each individual. These noise

*Notice that conditioning on $E=e$ would instead give $p(H|E=e) = \sum_A p(H|E=e,A)p(A|E=e)$.

variables are zero in expectation, so without access to their value for an individual we simply use $\mathcal{G}$: $E = w_{AE}A, H = w_{AH}A + w_{EH}E$ to make causal predictions. Suppose that for individual i we observe: $A = a^i, E = e^i, H = h^i$. We can answer the counterfactual question of "What if individual i had done more exercise, i.e. $E = e'$, instead?" by: a) *Abduction:* estimate the individual's specific makeup with $\varepsilon^i = h^i - w_{AH}a^i - w_{EH}e^i$, b) *Action:* set E to more exercise e', c) *Prediction:* predict a new value for cardiac health as $h' = w_{AH}a^i + w_{EH}e' + \varepsilon^i$.

6.10 RL Baselines

We can also compare the performance of these agents to two standard model-free RL baselines. The Q-total Agent learns a Q-value for each action across all steps for all the episodes. The Q-episode Agent learns a Q-value for each action conditioned on the input at each time step $[o_t, a_{t-1}, r_{t-1}]$, but with no LSTM memory to store previous actions and observations. Since the relationship between action and reward is random between episodes, Q-total was equivalent to selecting actions randomly, resulting in a considerably negative reward (-1.247 ± 2.940). The Q-episode agent essentially makes sure to not choose the arm that is indicated by m_t to be the external intervention (which is assured to be equal to -5), and essentially chooses randomly otherwise, giving a reward close to o (0.080 ± 2.077).

6.11 Additional Experiments

The purview of the previous experiments was to show a proof of concept on a simple tractable system, demonstrating that causal induction and inference can be learned and implemented via a meta-learned agent. In the following, we additionally demonstrate counterfactual reasoning, and scale up our results to more complex systems in two new experiments.

6.II.I Experiment 3: Counterfactual Setting

In Experiment 3, the agent was again allowed to make interventions as in Experiment 2, but in this case the quiz phase task entailed answering a counterfactual question. We explain here what a counterfactual question in our experimental domain looks like. Assume $X_i = \sum_j w_{ji} X_j + \varepsilon_i$ where ε_i is distributed as $\mathcal{N}(0.0, 0.1)$ (giving the conditional distribution $p(X_i|\text{pa}(X_i)) = \mathcal{N}(\sum_j w_{ji} X_j, 0.1)$ as described in Section 3). After observing the nodes $X_{2:N}$ (X_1 is hidden) in the CBN in one sample, we can infer this latent randomness ε_i for each observable node X_i (i.e. *abduction*) and answer counterfactual questions like "What would the values of the nodes be, had X_i instead taken on a different value than what we observed?", for any of the observable nodes X_i. We test three new agents, two of which are learned: "Active Counterfactual", "Random Counterfactual", and "Optimal Counterfactual Baseline" (not learned).

Counterfactual Agents: This agent is the same as the Interventional agent, but trained on tasks in which the latent randomness in the last information phase step $t = T - 1$ (where some $X_p = +5$) is stored and the same randomness is used in the quiz phase step $t = T$ (where some $X_f = -5$). While the question our agents have had to answer correctly so far in order to maximize their reward in the quiz phase was "Which of the nodes $X_{2:N}$ will have the highest value when X_f is set to -5?", in this setting, we ask "Which of the nodes $X_{2:N}$ would have had the highest value in the last step of the information phase, if instead of having the intervention $X_p = +5$, we had the intervention $X_f = -5$?". We run active and random versions of this agent as described in the main text.

Optimal Counterfactual Baseline: This baseline receives the true CBN and does exact abduction of the latent randomness based on observations from the penultimate step of the information phase, and combines this correctly with the appropriate interventional inference on the true CBN in the quiz phase.

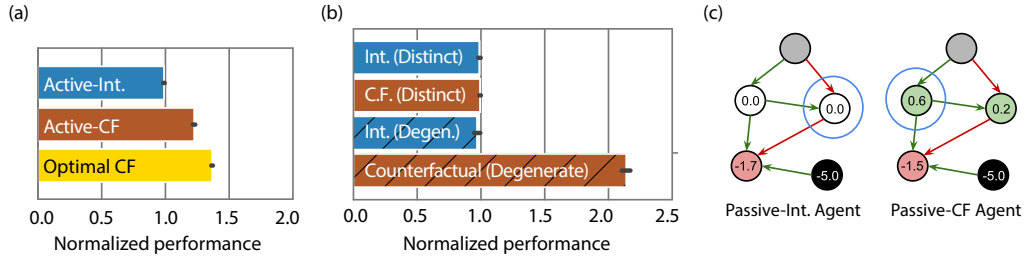


Figure 6.7: Experiment 3. Agents do counterfactual reasoning. a) Performance of the agents tested in this experiment. Note that performance can be above 1.0 since the counterfactual agent can theoretically perform better than the optimal interventional baseline, which doesn't have access to noise information. See main text for details. b) Performance split by if the maximum node value in the quiz phase is degenerate (Deg.) or distinct (Dist.). c) Quiz phase for an example test-CBN. See Figures in Main text for a legend. Here, the left panel shows $\mathcal{G}_{\to X_j=-5}$ and the nodes taking the mean values prescribed by $p_{\to X_j=-5}(X_{1:N\setminus j}|X_j=-5)$. We see that the Active-Int. Agent's choice is consistent with maximizing on these node values, where it makes a random choice between two nodes with the same value. The right panel panel shows $\mathcal{G}_{\to X_j=-5}$ and the nodes taking the exact values prescribed by the means of $p_{\to X_j=-5}(X_{1:N\setminus j}|X_j=-5)$, combined with the specific randomness inferred from the previous time step. As a result of accounting for the randomness, the two previously degenerate maximum values are now distinct. We see that the Active-CF. agent's choice is consistent with maximizing on these node values.

Results

We focus on two key questions in this experiment. (i) Do our agents learn to do counterfactual inference? The Active-Counterfactual Agent achieves higher performance than the maximum possible performance using only causal reasoning (Figure 6.7a). This indicates that the agent learns to infer and apply noise information from the last step of the information phase. To evaluate whether this difference is driven by the agent's use of abduction, we split the test set into two groups, depending on whether or not the decision for which node will have the highest value in the quiz phase is affected by the latent randomness, i.e. whether or not the node with the maximum value in the quiz phase changes if the noise is resampled. This is most prevalent in cases where the maximum expected reward is degenerate, i.e. where several nodes give the same maximum reward (denoted by hatched bars in Figure 6.7b). Here, agents with no access to the randomness have no basis for choosing one over the other, but different noise samples can give rise to significant differences in the actual values that these degenerate nodes have.

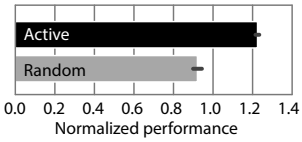


Figure 6.8: Active and Random Counterfactual Agents

We see indeed that there is no difference in the rewards received by the Active-Counterfactual and Active-Interventional Agents in the cases where the maximum values are distinct, however the Active-Counterfactual Agent significantly outperforms the Active-Interventional Agent in cases where there are degenerate maximum values. This performance increase is very high since in most cases where the maximum values are degenerate, this maximum value is close to 0.0. Thus, taking the noise into account gives the Counterfactual agent a huge relative advantage in these cases.

(ii) Do our agents learn to make useful interventions in the service of a counterfactual task? The Active-Counterfactual Agent's performance is significantly greater than the Random-Counterfactual Agent's (Figure 6.8). This indicates that when the agent is allowed to choose its actions, it makes tailored, non-random choices about the interventions it makes and the data it wants to observe – even in the service of a counterfactual objective.

6.11.2 Experiment 4: Non-linear Causal Graphs

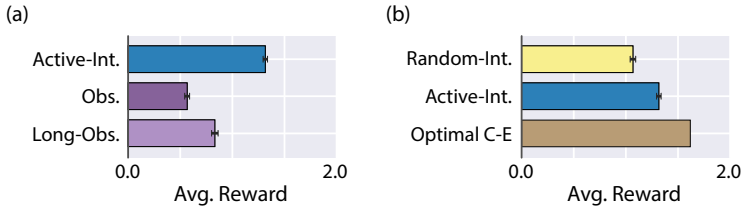


Figure 6.9: Results for non-linear graphs. (a) Comparing average episode reward for agents trained with different data. (b) Comparing information phase intervention policies.

In this experiment, we generalize some of our results to nonlinear, non-Gaussian causal graphs which are more typical of real-world causal graphs and to demonstrate that our results hold without loss of generality on such systems.

Here we investigate causal Bayesian networks (CBNs) with a quadratic dependence on the parents by changing the conditional distribution to $p(X_i|\text{pa}(X_i)) = \mathcal{N}(\frac{1}{N_i}\sum_j w_{ji}(X_j + X_j^2), \sigma)$. Here, al-

though each node is normally distributed given its parents, the joint distribution is not multivariate Gaussian due to the non-linearity in how the means are determined. We find that the Long-Observational Agent achieves more reward than the Observational Agent indicating that the agent is in fact learning the statistical dependencies between the nodes, within an episode. * We also find that the Active-Interventional Agent achieves reward well above the best agent with access to only observational data (Long-Observational in this case) indicating an ability to reason from interventions. We also see that the Active-Interventional Agent performs better than the Random-Interventional Agent, indicating an ability to choose informative interventions.

6.11.3 Experiment 5: Larger Causal Graphs

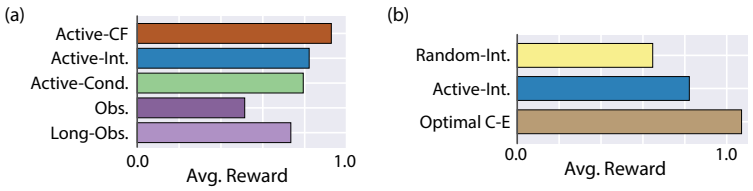


Figure 6.10: Results for $N = 6$ graphs. (a) Comparing average episode reward for agents trained with different data. (b) Comparing information phase intervention policies.

In this experiment we scaled up to larger graphs with $N = 6$ nodes, which afforded considerably more unique CBNs than with $N = 5$ (1.4×10^7 vs 5.9×10^4). As shown in Figure 6.10a, we find the same pattern of behavior noted in the main text where the rewards earned are ordered such that Observational agent $<$ Active-Conditional agent $<$ Active-Interventional agent $<$ Active-Counterfactual agent. We see additionally in Figure 6.10b that the Active-Interventional agent performs significantly better than the baseline Random-Interventional agent, indicating an ability to choose non-random, informative interventions.

*The conditional distribution $p(X_{1:N\setminus j}|X_j = 5)$, and therefore Conditional Agents, were non-trivial to calculate for the quadratic case, and was thus omitted.

References

[1] 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.

[2] 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).

[3] 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.

[4] 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.

[5] Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 39–48).

[6] 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.

[7] Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M., Pfau, D., Schaul, T., Shillingford, B., & Freitas, N. D. (2016). Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems* (pp. 3981–3989).

[8] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233.

[9] Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

[10] Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49, 307–343.

[11] Barbey, A. K. & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30(3), 241–254.

[12] Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, 24(04), 602–607.

[13] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

[14] Beach, L. R. & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3, 439–449.

[15] Beach, L. R., Wise, J. A., & Barclay, S. (1970). Sample proportions and subjective probability revisions. *Organizational Behavior and Human Performance*, 5(2), 183–190.

[16] Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological review*, 75(2), 127.

[17] Belinkov, Y. & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7, 49–72.

[18] 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).

[19] Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., & Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

[20] Benjamin, D. J. (2018). *Errors in probabilistic reasoning and judgment biases*. Technical report, National Bureau of Economic Research.

[21] 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.

[22] Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124(1), 1.

[23] Bird, S. & Loper, E. (2004). Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (pp.31).: Association for Computational Linguistics.

[24] Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology*, (pp. 85–94).

[25] 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.

[26] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[27] Blaisdell, A., Sawa, K., Leising, K., & Waldmann, M. (2006). Causal reasoning in rats. *Science*, 311(5763), 1020–1022.

[28] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003a). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

[29] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.

[30] 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.

[31] 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.

[32] Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? investigating the gap between prediction and action in toddlers? causal inferences. *Cognition*, 115(1), 104–117.

[33] Bordalo, P., Gennaioli, N., & Shleifer, A. (2017). *Memory, attention, and choice*. Technical report, National Bureau of Economic Research.

[34] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: Association for Computational Linguistics.

[35] Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological review*, 85(1), 1.

[36] 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*.

[37] 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.

[38] Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113(2), 409.

[39] Brown, S. D. & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58, 49–67.

[40] Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193.

[41] 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.

[42] 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.

[43] Burridge, R. R., Rizzi, A. A., & Koditschek, D. E. (1999). Sequential composition of dynamically dexterous robot behaviors. *The International Journal of Robotics Research*, 18(6), 534–555.

[44] Carey, S. (2009). *The origin of concepts*. Oxford University Press.

[45] Carroll, C. D. & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition*, 139, 130–153.

[46] Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, 71(5), 805–819.

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

[48] Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter.

[49] Chomsky, N. & Lightfoot, D. W. (2002). *Syntactic structures*. Walter de Gruyter.

[50] Cohen, A. L., Sidlowski, S., & Staub, A. (2017). Beliefs and Bayesian reasoning. *Psychonomic Bulletin & Review*, 24(3), 972–978.

[51] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.

[52] Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3), 393–405.

[53] Costello, F. & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3), 463–80.

[54] Costello, F., Watts, P., & Fisher, C. (2018). Surprising rationality in probability judgment: Assessing two competing models. *Cognition*, 170, 280–297.

[55] Costello, F. J. & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, 100, 1–16.

[56] Dasgupta, I., Schulz, E., & Gershman, S. J. (2017a). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25.

[57] Dasgupta, I., Schulz, E., & Gershman, S. J. (2017b). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25.

[58] Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, 178, 67–81.

[59] Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2019a). A theory of learning to infer. *BioRxiv*, (pp. 644534).

[60] Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2019b). A theory of learning to infer. *bioRxiv*.

[61] 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.

[62] 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.

[63] Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 4(6), 396–400.

[64] Dawid, P. (2007). *Fundamentals of Statistical Causality*. Technical report, University Colledge London.

[65] Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7, 889–904.

[66] De Freitas, N., Højen-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.

[67] Denil, M., Agrawal, P., Kulkarni, T. D., Erez, T., Battaglia, P., & de Freitas, N. (2016). Learning to perform physics experiments via deep reinforcement learning. *arXiv preprint arXiv:1611.01843*.

[68] 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.

[69] Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013b). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126, 285–300.

[70] Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013c). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126, 285–300.

[71] Dhami, M. K. & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of behavioral decision making*, 14(2), 141–168.

[72] 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.

[73] 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.

[74] Dougherty, M. & Hunter, J. (2003a). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31, 968–982.

[75] 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.

[76] 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.

[77] Dougherty, M. R. & Hunter, J. E. (2003c). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113(3), 263–282.

[78] Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

[79] Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018). Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*.

[80] DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, 85, 66–74.

[81] Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. *Judgment under uncertainty: Heuristics and biases*, (pp. 249–267).

[82] Edwards, W. (1968). Conservatism in human information processing. *Formal Representation of Human Judgment*, 17, 51.

[83] Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). Medical problem solving an analysis of clinical reasoning.

[84] Erev, I. & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112, 912–931.

[85] 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.

[86] 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).

[87] Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*.

[88] Evans, J. S. B. (2013). *The psychology of deductive reasoning (Psychology revivals)*. Psychology Press.

[89] 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.

[90] Evans, J. S. B. & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4), 382–389.

[91] Evans, J. S. B., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in bayesian inference. *Memory & Cognition*, 30(2), 179–190.

[92] Evans, J. S. B. & Perry, T. S. (1995). Belief bias in children's reasoning. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*.

[93] 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.

[94] Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, 21(3), 329–336.

[95] 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.

[96] Fernbach, P. M. & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1), 64–88.

[97] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

[98] Fischhoff, B. & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition*, 12, 402–410.

[99] Fischhoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90, 239–260.

[100] Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational behavior and human performance*, 23(3), 339–359.

[101] Fleming, S. M. & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91.

[102] Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.

[103] Fox, C. R. & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science*, 44(7), 879–895.

[104] Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998–998.

[105] Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4, e1000211.

[106] Gandhi, K. & Lake, B. M. (2019). Mutual exclusivity as a challenge for neural networks. *arXiv preprint arXiv:1906.10197*.

[107] 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.

[108] Ganin, Y., Kulkarni, T., Babuschkin, I., Eslami, S. M., & Vinyals, O. (2018). Synthesizing programs for images using reinforced adversarial learning. *arXiv preprint arXiv:1804.01118*.

[109] Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5), 507–534.

[110] Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1–58.

[111] Gennaioli, N. & Shleifer, A. (2010). What comes to mind. *The Quarterly Journal of Economics*, 125, 1399–1433.

[112] Gershman, S. & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

[113] Gershman, S. J. (2017). On the blessing of abstraction. *The Quarterly Journal of Experimental Psychology*, 70(3), 361–365.

[114] Gershman, S. J. (2019a). The generative adversarial brain.

[115] Gershman, S. J. (2019b). What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*.

[116] Gershman, S. J. & Beck, J. M. (2017). Complex probabilistic inference. In A. Moustafa (Ed.), *Computational Models of Brain and Behavior*. John Wiley & Sons.

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012a). Multistability and Perceptual Inference. *Neural Computation*, 24(1), 1–24.

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

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

[125] Gershman, S. J. & Wilson, R. (2010). The neural costs of optimal control. In *Advances in neural information processing systems* (pp. 712–720).

[126] Gershman, S. J., Zhou, J., & Kommers, C. (2017). Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, 29(12), 2103–2113.

[127] Gettys, C. F. & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, 24, 93–110.

[128] Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103, 592–596.

[129] Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, 3(1), 20–29.

[130] Gigerenzer, G. & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143.

[131] Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62, 451–482.

[132] Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.

[133] Gigerenzer, G. & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

[134] Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

[135] Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227.

[136] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[137] 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.

[138] Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, 118(1), 110.

[139] Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1), 3.

[140] Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*.

[141] Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in psychology*, 4.

[142] Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95(3), 537–557.

[143] Griffin, D. & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.

[144] Griffiths, T. L. (2015). Revealing ontological commitments by magic. *Cognition*, 136, 43–48.

[145] 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.

[146] Griffiths, T. L. & Tenenbaum, J. B. (2006a). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.

[147] Griffiths, T. L. & Tenenbaum, J. B. (2006b). Optimal predictions in everyday cognition. *Psychological Science*, 17(9), 767–773.

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] Groves, P. M. & Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychological review*, 77(5), 419.

[153] Gu, S. S., Ghahramani, Z., & Turner, R. E. (2015). Neural adaptive sequential monte carlo. In *Advances in Neural Information Processing Systems* (pp. 2629–2637).

[154] 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*.

[155] Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90, 649–660.

[156] 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).

[157] 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*.

[158] 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.

[159] 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.

[160] Hayes, B. K., Fritz, K., & Heit, E. (2013). The relationship between memory and inductive reasoning: Does it develop? *Developmental Psychology*, 49(5), 848–860.

[161] Hayes, B. K. & Heit, E. (2013). How similar are recognition memory and inductive reasoning? *Memory & Cognition*, 41(5), 781–795.

[162] Heckerman, D. (1990). A Tractable Inference Algorithm for Diagnosing Multiple Diseases 1 The QMR model.

[163] Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3), 197–243.

[164] Heit, E. & Hayes, B. K. (2011). Predicting reasoning from memory. *Journal of Experimental Psychology: General*, 140(1), 76–101.

[165] 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.

[166] Hertwig, R. & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523.

[167] 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.

[168] Hessel, M., Soyer, H., Espeholt, L., Czarnecki, W., Schmitt, S., & van Hasselt, H. (2018). Multi-task deep reinforcement learning with popart. *arXiv preprint arXiv:1809.04474*.

[169] Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., et al. (2017). Deep q-learning from demonstrations. *arXiv preprint arXiv:1704.03732*.

[170] Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, 138(2), 211–237.

[171] Hill, F., Cho, K., & Korhonen, A. (2016). Learning Distributed Representations of Sentences from Unlabelled Data.

[172] Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431.

[173] Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268, 1158–1160.

[174] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

[175] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2), 251–257.

[176] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems* (pp. 689–696).

[177] Jaakkola, T. S. & Jordan, M. I. (1999). Variational Probabilistic Inference and the QMR-DT Network. *Journal of Artificial Intelligence Research*, 10, 291–322.

[178] 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.

[179] Janzing, D., Hoyer, P. O., & Schölkopf, B. (2009). Telling cause from effect based on high-dimensional observations. *arXiv preprint arXiv:0909.4386*.

[180] Jazayeri, M. & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138), 912–915.

[181] Johnson, E. J. & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, 31(4), 395–414.

[182] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2901–2910).

[183] Johnson-Laird, P. N. & Steedman, M. (1978). The psychology of syllogisms. *Cognitive psychology*, 10(1), 64–99.

[184] Jones, M. N. & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, 114(1), 1.

[185] 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.

[186] Jordan, M. I. & Weiss, Y. (2002). Graphical models: Probabilistic inference. *The handbook of brain theory and neural networks*, (pp. 490–496).

[187] Kádár, A., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4), 761–780.

[188] Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

[189] Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.

[190] Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.

[191] Kang, D., Khot, T., Sabharwal, A., & Hovy, E. (2018). Adventure: Adversarial training for textual entailment with knowledge-guided examples. *arXiv preprint arXiv:1805.04680*.

[192] Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

[193] Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment*, 15, 292–307.

[194] 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.

[195] Kingma, D. P. & Welling, M. (2013a). Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*.

[196] Kingma, D. P. & Welling, M. (2013b). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

[197] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294–3302).

[198] Klein, G. (1999). *Sources of Power: How People Make Decisions*. MIT press.

[199] Knill, D. C. & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

[200] Koehler, D. (1994). Hypothesis Generation And Confidence in Judgment. *Learning, Memory*.

[201] Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1), 1–17.

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

[203] Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 28(9), 1321–1333.

[204] Körding, K. P. & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7), 319–326.

[205] Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6, 107–118.

[206] 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.

[207] Krynski, T. R. & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430.

[208] Lake, B. M. & Baroni, M. (2017). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*.

[209] Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*.

[210] 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.

[211] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2018). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

[212] Landauer, T. K. & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

[213] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

[214] Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

[215] Li, Y., Turner, R. E., & Liu, Q. (2017). Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*.

[216] Lieder, F. & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124, 762–794.

[217] 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).

[218] 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).

[219] 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.

[220] 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.

[221] 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.

[222] 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.

[223] Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*.

[224] Lightfoot, D. & Julia, P. (1984). The language lottery: Toward a biology of grammars.

[225] Linzen, T. (2019). What can linguistics and deep learning contribute to each other? response to pater. *Language*.

[226] Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

[227] Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4), 303–332.

[228] Luu, L. & Stocker, A. A. (2016). Choice-dependent perceptual biases.

[229] Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4), 287–298.

[230] MacKay, D. J. (2003). Information theory, inference and learning algorithms.

[231] Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., & Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems* (pp. 10846–10856).

[232] Marchiori, D., Di Guida, S., & Erev, I. (2015). Noisy retrieval models of over-and undersensitivity to rare events. *Decision*, 2, 82–106.

[233] Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

[234] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 1–8).

[235] Marewski, J. N. & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5, 39–59.

[236] Marewski, J. N. & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, 118, 393–437.

[237] Marino, J., Yue, Y., & Mandt, S. (2018). Learning to infer.

[238] Marr, D. & Poggio, T. (1976). From understanding computation to understanding neural circuitry.

[239] Martignon, L. & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52(1), 29–71.

[240] Massey, C. & Wu, G. (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 51(6), 932–947.

[241] McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8), 348–356.

[242] 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.

[243] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

[244] Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207.

[245] Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269–275.

[246] Meltzoff, A. N. (2007). Infants' causal learning: Intervention, observation, imitation.

[247] Mercier, H. & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

[248] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

[249] Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ?

[250] Mitrovic, J., Sejdinovic, D., & Teh, Y. W. (2018). Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems* (pp. 6986–6994).

[251] Mnih, A. & Gregor, K. (2014). Neural variational inference and learning in belief networks. In *International Conference on Machine Learning* (pp. 1791–1799).

[252] Mnih, V., Badia, A., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783.

[253] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529.

[254] Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, 7, e32548.

[255] Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108, 12491–12496.

[256] Murphy, K. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press.

[257] Naesseth, C. A., Linderman, S. W., Ranganath, R., & Blei, D. M. (2017). Variational sequential monte carlo. *arXiv preprint arXiv:1705.11140*.

[258] Neil Bearden, J. & Wallsten, T. S. (2004). Minerva-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making*, 17(5), 349–363.

[259] 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.

[260] Nie, Y., Wang, Y., & Bansal, M. (2019). Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 (pp. 6867–6874).

[261] Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.

[262] Oakhill, J., Johnson-Laird, P., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, 31(2), 117–140.

[263] Oaksford, M. & Chater, N. (2007a). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

[264] Oaksford, M. & Chater, N. (2007b). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

[265] Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes*, 42, 343–363.

[266] Oliva, A. & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), 520–527.

[267] Ommer, B. & Buhmann, J. (2009). Learning the compositional nature of visual object categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 501–516.

[268] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[269] Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92, 530–543.

[270] Orhan, A. E. & Ma, W. J. (2017). Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, 8(1), 138.

[271] Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., Heess, N., Veness, J., Pritzel, A., Sprechmann, P., et al. (2019). Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*.

[272] Paige, B. & Wood, F. (2016). Inference networks for sequential Monte Carlo in graphical models. In *International Conference on Machine Learning* (pp. 3040–3049).

[273] Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2017). Learning independent causal mechanisms. *arXiv preprint arXiv:1712.00961*.

[274] Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, 102, 127–144.

[275] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[276] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[277] Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: a primer*. John Wiley & Sons.

[278] Pearl, L. & Goldwater, S. (2016). Statistical learning, inductive bias, and bayesian inference in language acquisition.

[279] 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.

[280] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

[281] Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, 33(3-4), 175–190.

[282] Peterson, C. & Miller, A. (1964). Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology*, 68, 363.

[283] Peterson, C. R., DuCharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology*, 76, 236–243.

[284] Peterson, C. R. & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70(1), 117.

[285] 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.

[286] Peterson, C. R. & Ulehla, Z. (1964). Uncertainty, inference difficulty, and probability learning. *Journal of Experimental Psychology*, 67, 523–530.

[287] Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19, 285–293.

[288] Phillips, L. D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72, 346.

[289] Premack, D. & Premack, A. J. (1994). Levels of causal understanding in chimpanzees and children. *Cognition*, 50(1-3), 347–362.

[290] Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics* (pp. 814–822).

[291] Rasmussen, C. & Ghahramani, Z. (2003). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 15* (pp. 505–512).: MIT Press.

[292] 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.

[293] Redington, M., Crater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, 22(4), 425–469.

[294] Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, 72, 54–107.

[295] Rehder, B. & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, 45(2), 245–260.

[296] 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.

[297] Rezende, D. & Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* (pp. 1530–1538).

[298] 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).

[299] Rieskamp, J. & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135, 207–236.

[300] 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).

[301] Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International Conference on Machine Learning* (pp. 2940–2949).

[302] Ritter, S., Wang, J. X., Kurth-Nelson, Z., Jayakumar, S. M., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been there, done that: Meta-learning with episodic recall. *arXiv preprint arXiv:1805.09692*.

[303] Robert, C. & Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

[304] Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo*, 4(10.1201).

[305] 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.

[306] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[307] Ruiz, F. J. R. & Titsia, M. K. (2019). A contrastive divergence for combining variational inference and mcmc. *arXiv preprint arXiv:1905.04062*.

[308] Rule, J., Schulz, E., Piantadosi, S. T., & Tenenbaum, J. B. (2018). Learning list concepts through program induction. *BioRxiv*, (pp. 321505).

[309] Rumelhart, D. E. & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, 5(1), 1–28.

[310] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

[311] Saeedi, A., Kulkarni, T. D., Mansinghka, V., & Gershman, S. J. (2017a). Variational particle approximations. *Journal of Machine Learning Research*, 18, 1–29.

[312] Saeedi, A., Kulkarni, T. D., Mansinghka, V. K., & Gershman, S. J. (2017b). Variational particle approximations. *The Journal of Machine Learning Research*, 18, 2328–2356.

[313] Samuels, R., Stich, S., & Bishop, M. (2012). Ending the rationality wars. *Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010*, 2, 191.

[314] Sanborn, A. N. & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20, 883–893.

[315] 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).

[316] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

[317] Sasaki, S. & Kawagoe, T. (2007). Belief updating in individual and social learning: A field experiment on the internet.

[318] Saxe, R. & Carey, S. (2006). The perception of causality in infancy. *Acta psychologica*, 123(1-2), 144–165.

[319] 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.

[320] Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, 99, 44–79.

[321] Schulz, L. (2012). Finding new facts; thinking new thoughts. In *Advances in child development and behavior*, volume 43 (pp. 269–294). Elsevier.

[322] 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.

[323] Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275(5306), 1599–1603.

[324] Shah, A. K. & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134, 207–222.

[325] Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science*, 3(2), 143–162.

[326] 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).

[327] 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.

[328] 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.

[329] Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.

[330] 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.

[331] 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.

[332] Siegelmann, H. T. & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of computer and system sciences*, 50(1), 132–150.

[333] Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.

[334] Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization science*, 2(1), 125–134.

[335] Simoncelli, E. P. & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1), 1193–1216.

[336] Şimşek, Ö. (2013). Linear decision rule as aspiration for simple decision heuristics. In *Advances in neural information processing systems* (pp. 2904–2912).

[337] Şimşek, Ö., Algorta, S., & Kothiyal, A. (2016). Why most decisions are easy in tetris-and perhaps in other sequential decision problems, as well. *Proceedings of Machine Learning Research*, 48, 1757–1765.

[338] Singh, S. P. (1992). Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8(3-4), 323–339.

[339] 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.

[340] 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.

[341] 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.

[342] Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, 128(1), 64–75.

[343] Smith, K. A. & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1), 185–199.

[344] Socher, R., Manning, C. D., & Ng, A. Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, volume 2010 (pp. 1–9).

[345] Spirtes, P., Glymour, C., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

[346] 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.

[347] Stanford, P. K. (2010). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford University Press.

[348] Stewart, N., Chater, N., & Brown, G. D. (2006a). Decision by sampling. *Cognitive Psychology*, 53, 1–26.

[349] Stewart, N., Chater, N., & Brown, G. D. (2006b). Decision by sampling. *Cognitive Psychology*, 53(1), 1–26.

[350] Steyvers, M. (2006). Multidimensional scaling. *Encyclopedia of cognitive science*.

[351] Stocker, A. A. & Simoncelli, E. P. (2008). A bayesian model of conditioned perception. In *Advances in Neural Information Processing Systems* (pp. 1409–1416).

[352] Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. In *Advances in Neural Information Processing Systems* (pp. 3048–3056).

[353] 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.

[354] Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

[355] 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.

[356] Tenenbaum, J. B. & Griffiths, T. (2001). The rational basis of representatives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

[357] Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*.

[358] Thomas, R. P., Dougherty, M., Sprenger, A. M., & Harbison, J. I. (2008a). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155–185.

[359] 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.

[360] 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.

[361] Todd, P. M. & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3), 167–171.

[362] 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).

[363] Trueblood, J. S. & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, 35(8), 1518–1552.

[364] Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.

[365] Tversky, A. & Kahneman, D. (1974a). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, 185(4157), 1124–1131.

[366] Tversky, A. & Kahneman, D. (1974b). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

[367] Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.

[368] Tversky, A. & Koehler, D. J. (1994a). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.

[369] Tversky, A. & Koehler, D. J. (1994b). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4), 547.

[370] 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.

[371] Verma, T. & Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department.

[372] 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.

[373] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems* (pp. 3630–3638).

[374] 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).

[375] 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.

[376] 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.

[377] Vul, E. & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19, 645–647.

[378] Waldmann, M. R. & Hagmayer, Y. (2013). Causal reasoning.

[379] Wang, J., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *CoRR*, abs/1611.05763.

[380] Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018a). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6), 860.

[381] Wang, T., Wu, Y., Moore, D., & Russell, S. J. (2018b). Meta-Learning MCMC Proposals. In *Advances in Neural Information Processing Systems* (pp. 4150–4160).

[382] 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.

[383] 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.

[384] 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.

[385] Wheeler, G. & Beach, L. R. (1968). Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, 3(1), 36–46.

[386] White, A. S., Rastogi, P., Duh, K., & Van Durme, B. (2017). Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 996–1005).

[387] Whittington, J. C. & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*.

[388] 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.

[389] Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, 6, e1000871.

[Yildirim & Kulkarni] Yildirim, I. & Kulkarni, T. D. Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations.

[391] 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*.

[392] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

[393] Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).: Springer.

[394] Zhao, Z., Dua, D., & Singh, S. (2017). Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

[395] Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2018). Bayesian inference causes incoherence in human probability judgments.