Where do hypotheses come from?

Ishita Dasgupta
Department of Physics and Center for Brain Science
Harvard University

Eric Schulz
Department of Experimental Psychology
University College London

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

Abstract

Why are human inferences sometimes remarkably close to the Bayesian ideal and other times systematically biased? In particular, why do humans make near-rational inferences in some natural domains where the candidate hypotheses are explicitly available, whereas tasks in similar domains requiring the self-generation of hypotheses produce systematic deviations from rational inference. We propose that these deviations arise from algorithmic processes approximating Bayes' rule. Specifically in our account, hypotheses are generated stochastically from a sampling process, such that the sampled hypotheses form a Monte Carlo approximation of the posterior. While this approximation will converge to the true posterior in the limit of infinite samples, we take a small number of samples as we expect that the number of samples humans take is limited.We show that this model recreates several well-documented experimental findings such as anchoring and adjustment, subadditivity, superadditivity, the crowd within as well as the self-generation effect, the weak evidence, and the dud alternative effects. We confirm the model's prediction that superadditivity and subadditivity can be induced within the same paradigm by manipulating the unpacking and typicality of hypotheses. We also partially confirm our model's prediction about the effect of time pressure and cognitive load on these effects.

Keywords: hypothesis generation, Bayesian inference, Monte Carlo methods

---

Correspondence:
Ishita Dasgupta
Email: idasgupta@physics.harvard.edu
Department of Physics and Center for Brain Science
Harvard University
52 Oxford Street, room 295.08
Cambridge, MA 02138

Introduction

In his preface to *Astronomia Nova* (1609), Johannes Kepler described how he struggled to find an accurate mathematical description of planetary motion. Like most of his contemporaries, he started with the hypothesis that planets move in perfect circles. This necessitated extraordinary labor to reconcile the equations of motion with his other assumptions, "because I had bound them to millstones (as it were) of circularity, under the spell of common opinion." It was not the case that Kepler simply favored circles over ellipses (which he ultimately accepted), since he considered several other alternatives prior to ellipses. Kepler's problem was that he failed to generate the right hypothesis.[1]

Kepler is not alone: the history of science is replete with examples of "unconceived alternatives" (Stanford, 2010), 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, $P(d|h)$:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}. \tag{1}$$

Many studies have found that when $\mathcal{H}$ is supplied explicitly, humans can come close to the Bayesian ideal (e.g., Frank & Goodman, 2012; Griffiths & Tenenbaum, 2006, 2011; Oaksford & Chater, 2007; Petzschner, Glasauer, & Stephan, 2015).[2] However, when humans must generate the set of hypotheses themselves, they cannot generate them all and instead generate only a subset, leading to judgment biases (Carroll & Kemp, 2015; Dougherty & Hunter, 2003; Gettys & Fisher, 1979; Koriat, Lichtenstein, & Fischhoff, 1980; Thomas, Dougherty, Sprenger, & Harbison, 2008; Weber, Böckenholt, Hilton, & Wallace, 1993). Some prominent biases of this kind are listed in Table 1.

Most previously proposed models of hypothesis generation rely on cued recall from memory based on similarity to previously observed scenarios (c.f. Gennaioli & Shleifer, 2010; Thomas et al., 2008). 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,

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

[2] 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), typically in domains where people have strong intuitive knowledge about the relative probabilities of hypotheses; asking humans to reason consciously about unnatural problems like randomness or rare events (see Chater, Tenenbaum, & Yuille, 2006, for discussion), or carry out explicit updating calculations (Peterson & Beach, 1967), tends to produce deviations from the Bayesian ideal.

Table 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).	Fox and Tversky (1998)
Superadditivity	Perceived probability of a hypothesis is lower when the hypothesis is described as a disjunction of atypical component hypotheses (unpacked to atypical examples).	Sloman, Rottenstreich, Wisniewski, Hadjichristidis, and Fox (2004), Hadjichristidis, Stibel, Sloman, Over, and Stevenson (1999)
Weak evidence effect	The probability of an outcome is judged to be lower when positive evidence for a weak cause is presented	Fernbach, Darlow, and Sloman (2011)
Dud alternative effect	The judged probability of a focal outcome is higher when implausible alternatives are presented	Windschitl and Chambers (2004)
Self-generation effect	The probability judgment over hypotheses that participants have generated themselves is lower as compared to the same hypotheses generated by others	Koehler (1994); Koriat et al. (1980)
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.	Vul and Pashler (2008)
Anchoring and Adjustment	Generated hypotheses are biased by the hypothesis that is prompted at the start.	Tversky and Kahneman (1974)

the number of possible hypotheses is vast and only ever sparsely observed. Goodman, Tenenbaum, Feldman, and Griffiths (2008) 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 hy-

potheses only from the manageably small subset of previously observed hypotheses in memory and instead are able to generate hypotheses from the formidably large combinatorial space of all the conceivable possibilities. Given how large this space is, resource constraints at the time of inference suggest that only a subset are actually generated.

In this paper, we develop a normative theoretical framework for hypothesis generation in the domain of probabilistic inference, given fixed data, arguing that the brain copes with the intractability of inference by stochastically sampling hypotheses from the combinatorial space of possibilities (see also Sanborn & Chater, 2016). 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. Such biases are "computationally rational" in the sense that they result from a trade-off between the costs and benefits of computation—i.e., they are an emergent property of the expected utility calculus when costs of computation are taken into account (S. J. Gershman, Horvitz, & Tenenbaum, 2015; Lieder, Griffiths, Huys, & Goodman, 2017a; Vul, Goodman, Griffiths, & Tenenbaum, 2014). We propose that the framing of a query leads to sampling specific hypotheses first, which biases the rest of the hypothesis generation process through correlations in the sampling process. We discuss the properties of various sampler designs to explore the space of possible algorithms, and choose a specific design that can reproduce all the phenomena listed in Table 1. We then test our theory's novel predictions in four experiments.

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 (Chater et al., 2006). 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" (Marr & Poggio, 1976), 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 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 Griffiths, Vul, & Sanborn, 2012, 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.

Monte Carlo methods

In their simplest form, sample-based approximations (also known as *Monte Carlo* approximations; Robert & Casella, 2013), take the following form:

$$P(h|d) \approx \hat{P}_N(h|d) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[h_n = h], \tag{2}$$

where $\mathbb{I}[\cdot] = 1$ when its argument is true (0 otherwise) and h_n is a random hypothesis drawn from some distribution $Q_n(h)$.[3] 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. Some of these properties are shared with other sampling mechanisms, and others make MCMC more uniquely amenable.

First, all Monte Carlo approximations including MCMC, are stochastic in the finite sample regime, producing "posterior probability matching" (Denison, Bonawitz, Gopnik, & Griffiths, 2013; Moreno-Bote, Knill, & Pouget, 2011; Vul et al., 2014; Wozny, Beierholm, & Shams, 2010): 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. This is consistent with evidence that humans represent probabilities on a relative scale (Stewart, Chater, & Brown, 2006). While this property is not true of all samplers, it is shared with a large class of sampling mechanisms based on importance sampling. 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 experimental observations (Hamrick, Smith, Griffiths, & Vul, 2015). Here the properties of MCMC diverge more significantly from parallel sampling methods like importance sampling, where hypotheses are generated independently. Fourth, Markov chains (unlike parallel sampling mechanisms such as importance sampling) generate autocorrelated samples. This is consistent with autocorrelation in hypothesis generation (Bonawitz, Denison, Gopnik, & Griffiths, 2014; S. J. Gershman, Vul, & Tenenbaum, 2012; Vul & Pashler, 2008). Correlation between consecutive hypotheses that manifest as anchoring effects (where judgments are biased by the initial hypothesis; Tversky & Kahneman, 1974) are replicated by MCMC approximations that are transiently biased (during the"burn-in" period) by their initial hypothesis (Lieder, Griffiths, Huys, & Goodman, 2017b; Lieder et al., 2017a). This seems to hold also true for the way in which participants update their internal models in causal learning tasks (Bramley, Dayan, Griffiths, & Lagnado, 2017). Finally, work in theoretical neuroscience has shown

[3] 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 $H \subset \mathcal{H}$.

how MCMC algorithms could be realized in cortical circuits (Buesing, Bill, Nessler, & Maass, 2011; Moreno-Bote et al., 2011; Pecevski, Buesing, & Maass, 2011).

We will show how some of the biases in Table 1 can be replicated with samplers that have some subsets of these properties. Importantly, we will also show how a particular MCMC sampler can capture all of the biases in Table 1.

Computational rationality of sampling. We have emphasized properties that emerge in the finite sample regime because people tend to only generate a small number of hypotheses (Dougherty, Gettys, & Thomas, 1997; Gettys & Fisher, 1979; Klein, 1999; Ross & Murphy, 1996; Weber et al., 1993). Although this may seem to be manifestly sub-optimal, it can be justified within a "computational rationality" or "resource-rational" framework (S. J. Gershman et al., 2015; Griffiths, Lieder, & Goodman, 2015; Schulz, Speekenbrink, & Meder, 2016; Vul et al., 2014). 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. (2015) showed that people generated more hypotheses when they were more uncertain. By the same token, cognitive load (Sprenger et al., 2011) or response time pressure (Dougherty & Hunter, 2003) act as disincentives, reducing the number of generated hypotheses.

Despite our focus on the finite sample regime, it is also important to consider the asymptotic regime in order to explain the cases where human inference comes close to the Bayesian ideal. Monte Carlo algorithms are typically asymptotically exact; thus, they can accommodate unbiased inference when adequate cognitive resources are available. We do not claim, however, that all biases in human inference arise from adaptive allocation of cognitive resources. It seems likely that evolution has endowed the mind with some hardwired heuristics in order to avoid the cost of adaptive resource allocation (Gigerenzer & Brighton, 2009).

Comparison with particle filtering. A key feature of MCMC is that it produces hypotheses sequentially. As mentioned above, this gives it properties that distinguish it from parallel sampling mechanisms like importance sampling—specifically, the feedback between the generation and evaluation processes, and the autocorrelation of samples. It is therefore useful to compare MCMC with *particle filtering*, another Monte Carlo algorithm that generates hypotheses sequentially, and which has also been fruitfully applied to a number of domains in psychology, such as multiple object tracking (Vul, Alvarez, Tenenbaum, & Black, 2009), categorization (Sanborn, Griffiths, & Navarro, 2010), and change detection (Brown & Steyvers, 2009). In order to clarify the distinction between the sequential nature of particle filtering and MCMC, we note that the sequential structure of particle filtering 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. Particle filters can also be used for inferring static parameters (Chopin, 2002), updating the Monte Carlo approximation as new data arrive. Note that in this case the generative process is still inherently sequential. In contrast, MCMC always involves sequential hypothesis generation, regardless of the structure of the generative process.

MCMC can also be used in conjunction with particle filters: the samples generated by the particle filter can be "rejuvenated" by applying a Markov chain operator that preserves the target distribution (Abbott & Griffiths, 2011; Thaker, Tenenbaum, & Gershman, 2017). This process prevents degeneracy (collapse of the Monte Carlo approximation onto a few samples), a common problem in particle filtering. Here, the sequential nature of the Markov chain is

relevant only locally to each step of the particle filter, orthogonal to the sequential nature with which the particle filter processes new data. In this paper, we focus on non-sequential generative models, with no online updating of data, in order to retain clarity on this point.

A specific Markov chain Monte Carlo algorithm

The space of MCMC algorithms is vast (Robert & Casella, 2013), 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:

$$P(h_{n+1} = h'|h_n) = \min\left[1, \frac{P(d|h')P(h')Q(h_n|h')}{P(d|h_n)P(h_n)Q(h'|h_n)}\right]. \tag{3}$$

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 is known 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. In the limit of large N, the amount of time the chain spends at a particular hypothesis is proportional to its posterior probability. If N is not large enough, then the samples are affected by the initialization, leading to biased estimates of the 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[4]

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:

$$P(h_{n+1} = h'|h_n) = \min\left[1, \frac{P(d|h')P(h')}{P(d|h_n)P(h_n)}\right]. \tag{4}$$

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

[4]A low acceptance rate only implies that proposals are lower probability than the current state of the Markov chain, not that the current hypothesis necessarily has a high probability globally. There may always be higher probability hypotheses that the proposal distribution fails to propose. Conversely, a high acceptance rate does not necessarily imply a poor current hypothesis. For example, if the proposal distribution is proportional to the posterior distribution, then all proposals will be accepted.

the hypothesis generated next is close to the current one with high probability. The alternative 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 as long as the acceptance ratio is less than one (irrespective of the details of the proposal distribution), because the same state occurs consecutively when a proposal is 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 (Abbott, Austerweil, & Griffiths, 2015) and combinatorial spaces (Smith, Huber, & Vul, 2013). This locality has been shown to be optimal as a foraging strategy (Hills, Jones, & Todd, 2012) 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 Lieder, Griffiths, and Goodman (2013); 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 limitations and future extensions.

Finally, we need to specify how the chain is initialized. For cases where a hypothesis is presented explicitly or primed in the query, we assume that the chain starts at that hypotheses. If there are several hypotheses (say n in number) that have been presented explicitly in the query, we assume that a different chain starts from each of these hypotheses and runs for $\frac{N}{n}$ steps each, giving a total of N samples. However, in cases where no hypotheses are explicitly prompted, we assume that the initial hypothesis is drawn from the prior over the hypotheses instead of initializing at a prompted example. This assumption is consistent with evidence that hypotheses with high base rates are more likely to be generated (Weber et al., 1993). In order to maximize similarity to the corresponding "explicitly prompted" version of the question and keep the number of new initializations the same, n such chains are run for $\frac{N}{n}$ steps to give a total of N samples. 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 limitations and future extensions.

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.

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 (Cooper, 1990). To address this problem, approximate probabilistic inference algorithms (including Monte Carlo methods) are now widely-established (e.g., Heckerman, 1990; Jaakkola & Jordan, 1999; Shwe & Cooper, 1991). It is therefore 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 1 have been documented in clinical settings (Elstein, Shulman, & Sprafka, 1978; Redelmeier, Koehler, Liberman, & Tversky, 1995; Weber et al., 1993); 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 Shwe et al. (1991):

$$P(k_s = 1|h) = 1 - (1 - \epsilon) \prod_{d=1}^{D} (1 - w_{ds})^{h_d}, \tag{5}$$

where $k_s = 1$ when symptom $s = 1, \ldots, S$ is present (0 otherwise), $\epsilon \in [0, 1]$ 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). 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.

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 (Vul et al., 2014).

Table 2
Parameters used for noisy-or model.

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

This gives rise to *subadditivity* (see Table 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) (Dougherty & Hunter, 2003; Tversky & Koehler, 1994), 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 example components of the focal space are unpacked and explicitly prompted in the question (food poisoning and stomach cancer) and presented along with a catch-all hypothesis (any other gastrointestinal disease).[5]

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 are not explicitly prompted and instead a packed hypothesis is presented, the chain initializes with a random sample from

[5]In this paper, we study what is termed "implicit" subadditivity, where the unpacked query is framed as a conjunction of mutually exclusive sub-hypotheses, in contrast to "explicit" subadditivity, where each mutually exclusive sub-hypothesis is queried separately and the numerical estimates from each query are then added together. Explicit subadditivity could be modeled the same way as implicit subadditivity if we assume that the number of samples generated over the separately queried sub-hypotheses is equal to the net number of hypotheses generated in response to the conjunction, and that no samples are carried over in between the separately queried hypotheses.

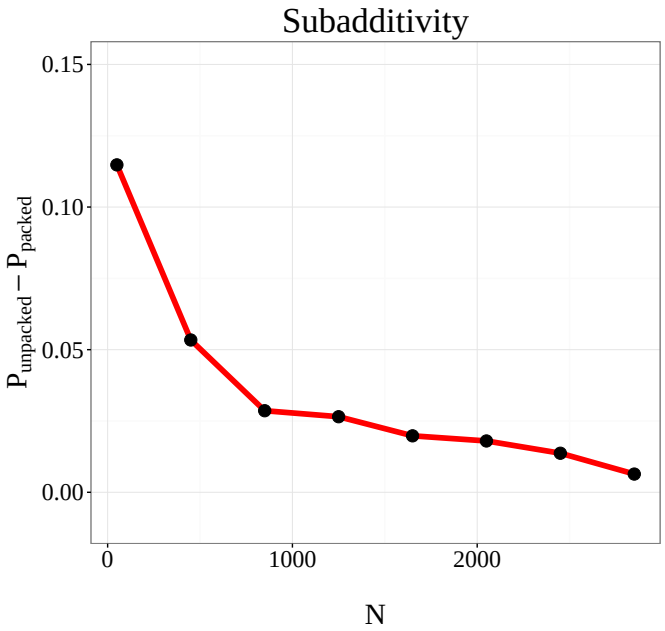

Figure 1. **Subadditivity**. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Packed: what is the probability that these symptoms were caused by the presence of a gastrointestinal disease? (b) Unpacked to typical examples: what is the probability 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.

the prior. The chain is thus likely to start from a fairly typical (high prior probability) hypothesis; however, with some probability it may fail to generate all the high probability hypotheses. Deterministically initializing the chain at a typical (high probability) hypothesis, ensures that the chain generates high probability hypotheses in the focal space and thus results in a larger probability judgment for that focal space. This effect can also be replicated by a parallel sampling algorithm as seen in Thomas et al. (2008). Here explicitly prompted hypotheses (under the unpacked condition) are appended to the other samples that would have been generated without prompting (under the packed condition), leading to more hypotheses in the focal space being generated in the unpacked condition and therefore raising the probability estimate under that condition.

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 1.

Experiments in Dougherty and Hunter (2003) 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 1.

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 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) (Hadjichristidis et al., 1999; Sloman et al., 2004), 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

$$\sum_{h\in\mathcal{A}} \hat{P}(h|d) = \sum_{h\in\mathcal{A}} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[h_n = h] = \frac{\sum_{h\in\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h]}{\sum_{h\in\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h] + \sum_{h'\notin\mathcal{A}} \sum_{n=1}^{N} \mathbb{I}[h_n = h']} \quad (6)$$

Being in $\mathcal{A}$ or not divides the total hypothesis space of $\mathcal{H}$ into two mutually exclusive parts. Therefore, the generation of more hypotheses outside the focal space (on average) when initialized at a consistently low probability (atypical) hypothesis in the focal space lowers the resulting probability estimate of the focal hypothesis space. This results in superadditive judgments.

To elucidate this effect in our medical diagnosis model, we use the following "unpacked to atypical examples" query: Given the symptoms *fever, nausea* and *fatigue*, what is the probability that these symptoms were caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease? The difference between the probability estimates from the two conditions is shown in Figure 2.

Previous accounts of subadditivity (e.g., Neil Bearden & Wallsten, 2004; Thomas et al., 2008) cannot explain superadditivity; any unpacked example only increases the probability judgment of the unpacked query with respect to the packed query. This weakness of

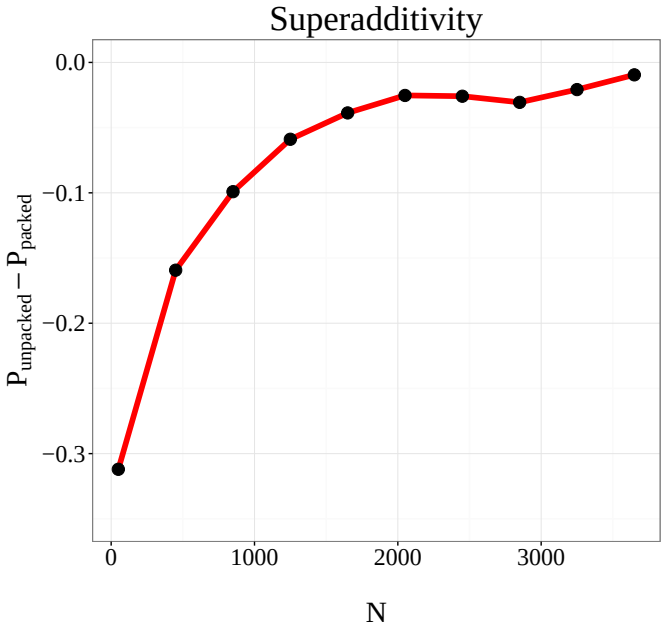

Figure 2. **Superadditivity**. MCMC-estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Packed: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Unpacked to atypical examples: What is the probability is that these symptoms were caused by the presence of gastroenteritis, stomach cancer, or any other gastrointestinal disease? The estimate for the unpacked condition is lower than that of the packed condition. The difference between these estimates is represented by the red line. This effect diminishes as the number of samples increases.

MINERVA-DM has been observed by Costello and Watts (2014) in the context of its failure to model binary complementarity—an effect which their noise-based analysis can capture. However, their analysis still fails to completely capture superadditivity, as it constrains unpacked judgments to be greater than (and, only for binary complements, equal to) the packed judgment, never less than the packed judgment. Our modeling of this effect hinges on the fact that MCMC allows for feedback between the generation and evaluation processes—the evaluated probability of already generated hypotheses influences how many new hypotheses will be generated. This property is not shared by parallel sampling algorithms. However, other samplers (besides MCMC algorithms) that exhibit correlated sampling may exhibit similar behaviors (see for example Bonawitz et al., 2014).

Sloman et al. (2004) explain superadditivity by suggesting that atypical examples divert attention from more typical examples and thus lower the probability estimate. But an explanation at the level of a 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. Fernbach et al. (2011) explain this effect as follows: mentioning evidence in support of a specific weak cause drives people to focus disproportionately on it and thus they fail to think about other good candidates in the focal space of possible causes. Our model replicates this effect by initializing at the weak cause, or low-probability hypothesis, resulting in a lower probability judgment of the

focal space by the same mechanism as in the superadditivity effect. However, the added evidence should normatively increase the probability of the cause it supports. Since the evidence is weak, this increase is small and the cause still remains low probability. Therefore, the superadditivity effect overwhelms this small increase in probability of the specific hypothesis and instead lowers the probability estimate of the focal space overall. This causes the final judged 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). While this small change is not expected to elicit a large difference in the probability 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 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 (Windschitl & Chambers, 2004). This can be viewed as the superadditivity effect in the complement (alternate) hypothesis space. The queries being contrasted here are initialized in the space complementary to the focal space—i.e., the space of alternatives. Initialization at a low probability alternative when it is explicitly prompted in the question results in a superadditive judgment (i.e., a lower probability judgment) of the complement space. This lower probability estimate for the complement space entails a higher probability estimate for the focal space. The assumption here is that the same chain is used to gauge the probability of both binary complements, by grouping the generated hypotheses into being either inside or outside the focal space and calculating the net- probability of each group. The framing simply alters the initialization of the chain. This assumption ensures that probability judgments over complementary spaces add up to one, in accordance with behavioral experiments that demonstrate binary complementarity in human judgments (Tversky & Koehler, 1994).

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)?

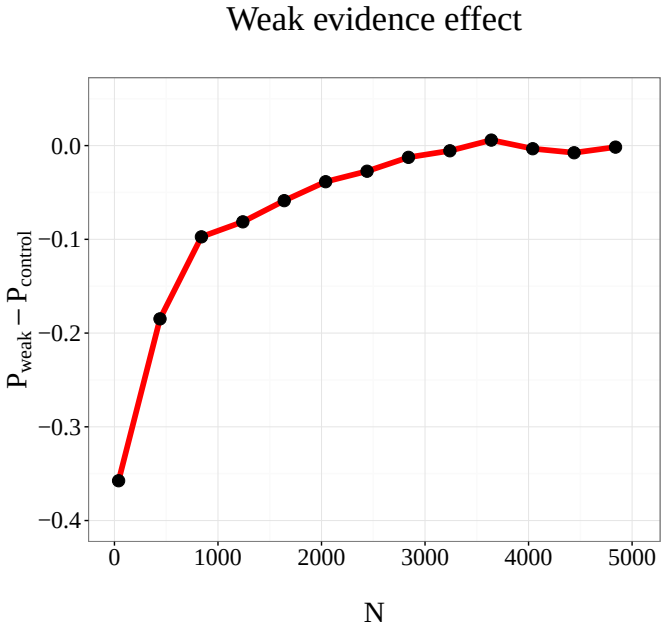

Figure 3. **Weak evidence effect**. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Control: What is the probability that these symptoms were caused by the presence of gastrointestinal disease? (b) Evidence for a weak cause: What is the probability 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.

We see in Figure 4 that the model predicts that the scenario with dud alternatives produces higher probability judgments than the control. Findings in Windschitl and Chambers (2004) 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 of samples is reached.

Our model currently only captures cases of binary complementarity where it's obvious to participants that complementarity holds. If this complementarity is obvious, then they can use the same chain, and if the complementarity isn't obvious, then they use a new chain. If this new chain is not suggestively unpacked, approximate binary complementarity should still hold. It is an interesting challenge to understand when humans might re-use the same chain and when they might use a new chain, and when they might use some intermediate between the two. This is part of our ongoing research.

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 (Koehler, 1994; Koriat et

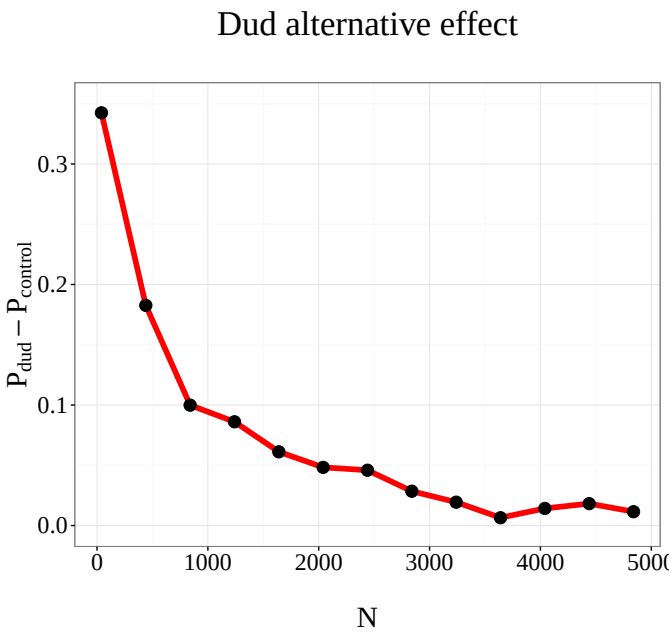

Figure 4. **Dud alternative effect**. MCMC estimates were made for the following queries: Given the symptoms *fever, nausea* and *fatigue*, (a) Control: What is the probability that the patient has a respiratory disease (as opposed to the symptoms being caused by the presence of a gastrointestinal disease)?, (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

al., 1980). 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 generation of fewer alternate hypotheses. As in the previous section, fewer alternate hypotheses lead to a higher probability judgment.

We simulate an experiment analogous to the experiments in Koehler (1994) 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 5.

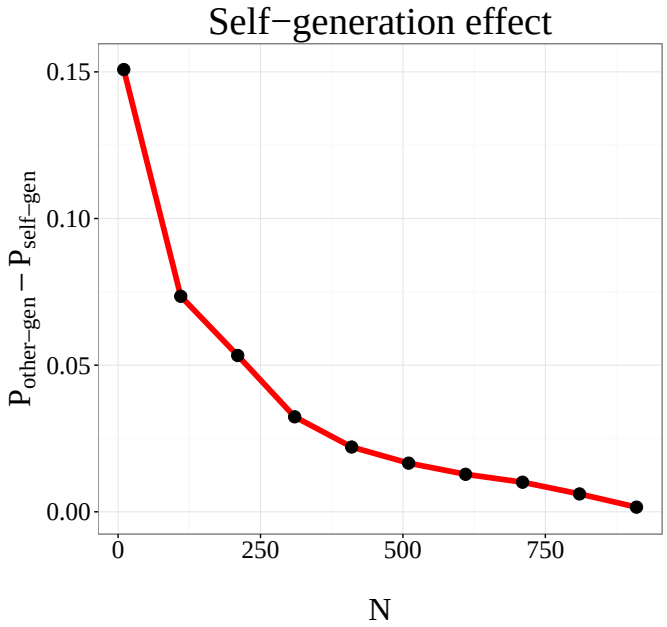

Figure 5. **Self-generation effect**. MCMC estimates for the following query: Given the symptoms *fever* and *fatigue*, (a) Self-generated: What are the two most likely respiratory diseases to have caused these symptoms? Estimate the probability that these symptoms are caused by either of these two diseases. (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

While this effect has previously been understood in terms of the generation of alternatives (Koehler, 1994), a rational process model specifying a mechanism for this differential generation of alternatives is novel. Our explanation of this effect is also contingent upon a property unique to MCMC – the 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 because most new proposals are rejected, resulting in fewer generated alternatives.

Anchoring and adjustment

In a classic experiment, Tversky and Kahneman (1974) had participants observe a roulette wheel that was predetermined to stop on either 10 or 65. Participants subsequently had to guess 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 and Kahneman (1974) 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. (2017a) showed that the anchoring and adjustment heuristic is a basic consequence of MCMC algorithms, due to the inherent autocorrelation of samples. Consistent

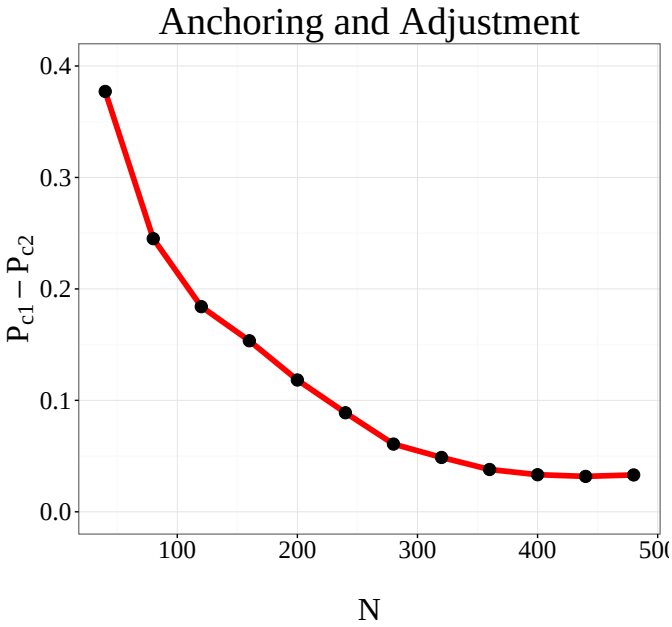

Figure 6. **Anchoring and adjustment**. The y axis represents the difference in the probabilities of respiratory flu and stomach flu given the symptoms *fever* and *fatigue* as returned by two different chains that are initialized differently. The chains are initialized in the two different clusters, 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.

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

We illustrate this effect in Figure 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.

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* (Surowiecki, 2005). 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 Vul and Pashler (2008) 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 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[6], 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 Vul and Pashler (2008)—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 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.

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:

$$\hat{P}_N(h|d) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[h_n = d]w_n, \tag{7}$$

[6]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 have a lower error on average than the first one. Vul and Pashler (2008) find this to not be the case empirically.

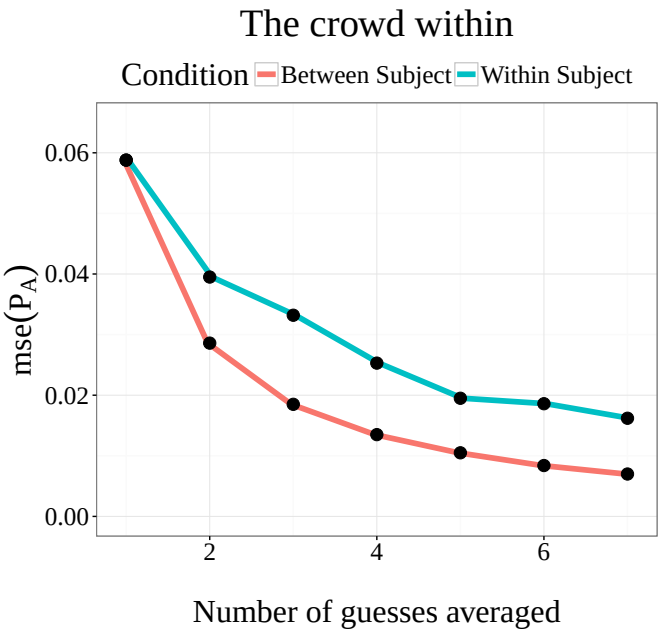

Figure 7. **The crowd within**. Errors in the MCMC estimates for the following query: Given the symptoms *nausea* and *shortness of breath*, what is the probability that these were caused by the presence of a respiratory disease? The estimates are averaged either over samples from the same individual (blue) or over samples from different individuals (red)

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

$$w_n \propto \frac{P(h_n, d)}{Q(h_n)}. \tag{8}$$

Intuitively, the importance weight corrects for the fact that the importance sampler draws samples from the wrong distribution. Shi, Griffiths, Feldman, and Sanborn (2010) 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 and Griffiths (2009) demonstrated how importance sampling could be realized in a biologically plausible neural circuit (see also Abbott, Hamrick, & Griffiths, 2013).

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. (2008) 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 Koehler (1994) 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 importance 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 1 that rely on correlated sampling, such as the anchoring effect and the crowd within.

It is also valuable to contrast MCMC with anchoring and adjustment schemes that involve incremental changes to a numerical estimate in the direction of the target value. Although MCMC produces autocorrelation of samples, it does not require changes to be incremental; MCMC allows the proposal distribution to be non-local. In fact, substantial evidence suggests that some of these changes can be quite dramatic, as in perceptual multistability (S. Gershman, Hoffman, & Blei, 2012) and insight problem-solving (Sternberg & Davidson, 1995).

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 scene perception and understanding. Extensive research suggests that the visual system encodes information about natural scene statistics (Barlow, 2001; Simoncelli & Olshausen, 2001). Although these low-level scene statistics like the distribution of oriented edges are not consciously accessible, statistics at the level of objects, for example object co-occurrence statistics in natural scenes studied in Greene (2013), can be used to inform a generative model that can act as a proxy for

one aspect of human scene understanding. We can then leverage such models to test theories of hypothesis generation in this domain.

Specifically, Greene (2013) provides a database of natural scenes with hand-labeled objects. We fit the latent Dirichlet allocation (LDA) model (Blei, Ng, & Jordan, 2003) to this dataset, allowing us to capture the distribution of co-occurrences of different objects in terms of latent "topics" (distributions over objects). Each scene is modeled as a probabilistic mixture of topics. The LDA model captures the fact that microwaves are likely to co-occur with toasters, and cars are likely to co-occur with mailboxes. The marginal distribution of objects provides a natural empirical prior over objects. We do not fit any free parameters to the dataset; all hyperparameters are set to the values described in Blei et al. (2003).

For our purposes, the important point is that we can use our 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.

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 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, 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 8.

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 mean experimental probability estimates and the probability estimates from the model. This estimate was used to generate confidence intervals. The value of k that best fit the data was $k = 6$ with negligible uncertainty, and the number of samples $N = 230$ with a 95% confidence interval $[191, 269]$. 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 Greene (2013). This value for N as the number of samples is higher than numbers found in some previous work like Vul et al. (2014) 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.

The details of the proposal distribution could also influence the individual and relative magnitudes of the observed subadditivity and superadditivity effects, and perhaps different

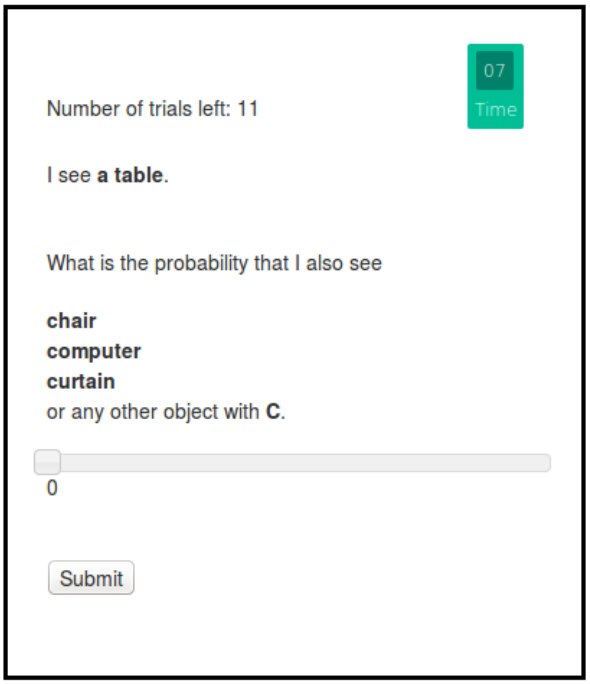

Figure 8. **Experimental setup**. Participants were asked to estimate the conditional probability using a slider bar within a 20-second time limit.

parameters for N and k. Instead of making strong assumptions about locality in this particular hypothesis space, we use a uniform proposal distribution.

Results and discussion. We compared the mean probability judgments for each condition (Figure 9). Consistent with our hypothesis, we found subadditivity in the unpacked-typical condition, with significantly higher probability estimates compared to the control condition $[t(58) = 4.53, p < 0.01]$, and superadditivity in the unpacked-atypical condition, with significantly lower probability estimates compared to the control condition $[t(58) = -4.97, p < 0.01]$. 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 Sloman et al. (2004), 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.

Experiment 2

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

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

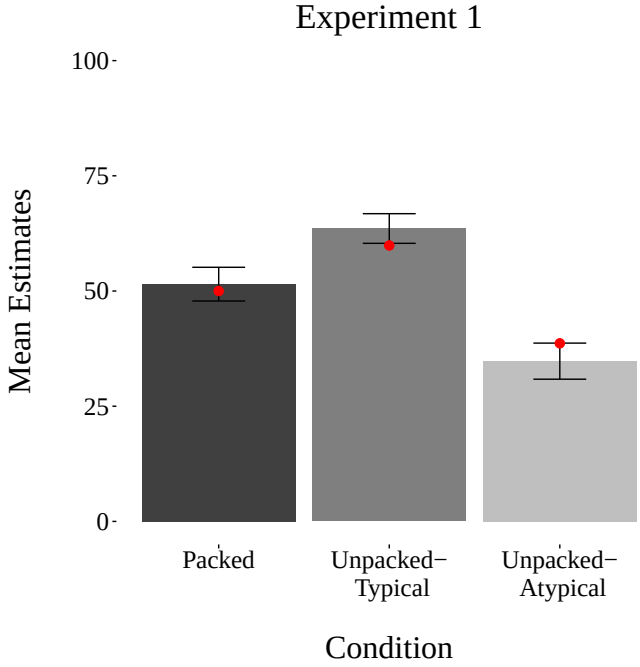

Figure 9. **Experiment 1 results**. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene.

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

Table 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 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 4) leads to significantly lower estimates than using "Unpacking 2" $[t(251) = 2.52, p < 0.01]$. Additionally,

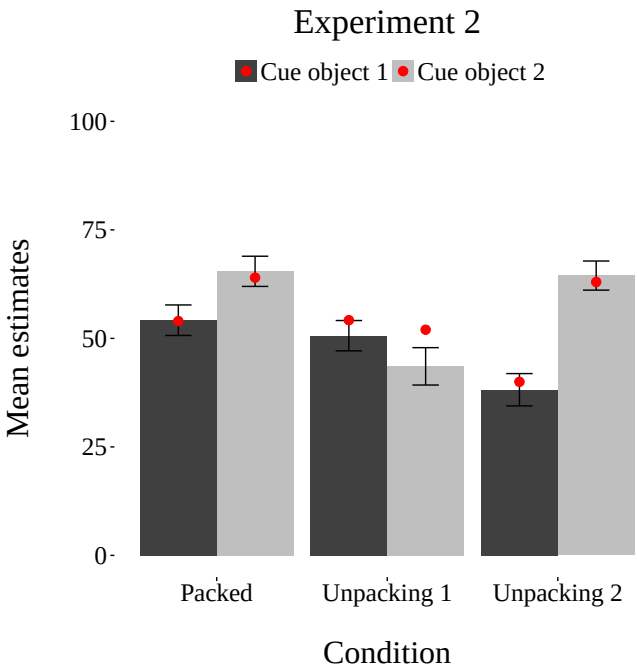

Figure 10. **Experiment 2 results**. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 230 samples, assuming 6 hidden objects in the scene. Unpacking 1 is typical for cue object 1 and atypical for cue object 2; unpacking 2 is typical for cue object 2 and atypical for cue object 1.

when conditioning on cue object 1, using "Unpacking 2" produces significantly lower estimates than using "Unpacking 1"; $[t(165) = -3.31, p < 0.001]$. These results show that typicality of the unpackings and, by proxy the sub- and super-additive effects, crucially depend on the conditioned cue object.

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

Experiment 3

A key prediction of our model is that the strength of subadditivity and superadditivity will decrease with the number of sampled hypotheses, as the chain approaches its stationary distribution. To test this prediction, we repeated Experiment 1, but reduced the time limit and incentivized participants to respond more quickly. We predicted that these changes would lead to stronger subadditivity and superadditivity effects.

Participants. 62 participants (34 females, mean age= 25.65, SD=12.36) were recruited via Amazon's Mechanical Turk web service and received \$0.5 for their participation plus a performance-dependent bonus.

Materials and procedure. Materials were the same as in Experiment 1. However, in this experiment participants had less time available per trial (5 seconds) and were asked to respond as quickly as possible. Participants were paid a baseline amount for their participation of \$0.5. Additionally, they were incentivized to respond quickly: they could gain more money the faster they responded on each trial (up to \$0.1 per trial) and gained an additional \$0.1 for every on time response per trial overall.

Results and Discussion. The mean estimates for the different conditions are shown in Figure 11. Replicating the results of Experiment 1, the estimates for the unpacked-atypical condition were significantly lower than for the packed condition $[t(57) = -4.8183, p < 0.01]$, and the estimates for the unpacked-typical condition were significantly higher than for the packed condition $[t(57) = 4.76, p < 0.01]$. Our hypothesis generation model fits the data well with parameter values $K = 3$ with negligible uncertainty and $N = 170$ with 95% confidence interval $[94, 246]$. We see that the best fit number of samples is substantially lower than that found in Experiment 1 ($N = 230$, with 95% confidence interval $[191, 269]$). The number of hidden objects K is also lower. These parameter estimates are consistent with the idea that time pressure results in fewer generated samples and fewer objects under consideration.

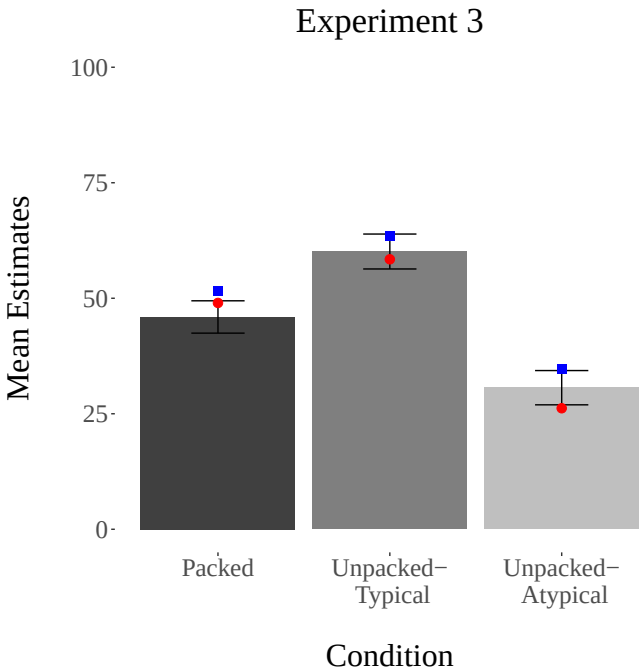

Figure 11. **Experiment 3 results**. Mean probability estimates for each condition. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 170 samples, assuming 3 hidden objects in the scene. Blue squares show means estimates of Experiment 1.

Next, we performed a median split based on the overall reaction times and thereby classified trials into slow and fast trials. The slow and fast trials were separately fit using the same value of K from the overall responses and adjusting N. We see that the data from the fast trials are better fit with a lower N ($N = 150$, with 95% confidence interval $[98, 202]$) than the slow

trials ($N = 190$, with 95% confidence interval $[125, 255]$). The estimates for the slow trials have a high overlap with the estimates from Experiment 1 ($N = 230$, with 95% confidence interval $[191, 269]$). However, the intervals for the fast response and the one from Experiment 1 have a small overlap of ~ 10 steps. The results are shown in Figure 12. We then performed an ANOVA, regressing the median time (fast or slow response), condition (packed, typically unpacked and atypically unpacked hypothesis) onto participants' probability estimates, where responses were nested within participants. Condition was a significant predictor of participants' responses ($\chi^2(1) = 157.8$, $p < 0.001$). The time variable alone was not a significant predictor of participants' responses ($\chi^2(1) = 3.9$, $p = 0.05$). This is expected since the subadditivity and superadditivity effects go in opposite directions. The interaction between time and condition was significant ($\chi(1) = 37.03$, $p < 0.01$) indicating that the time variable influences the estimates depending on condition. Further assessing this difference between the interactions again using a nested ANOVA showed that faster responses produced greater subadditivity effect as compared to slow responses ($t(248) = -2.1602$, $p < 0.05$). The difference in the superadditivity effect however was not significant ($t(213) = 0.78$, $p = 0.4$). Comparing the sub- and superadditivity effects of Experiment 3 to the effects of Experiment 1, we found that they were relatively similar overall ($t(453) = -1.353$,$p > 0.1$). However, comparing only the fast responses from Experiment 3 to the results of Experiment 1, we found an increased subadditivity effect ($t(102) = -2.46$, $p < 0.05$) but a similar superadditivity effect ($t(104) = -0.71$, $p = 0.48$). This is in agreement with the model fits.

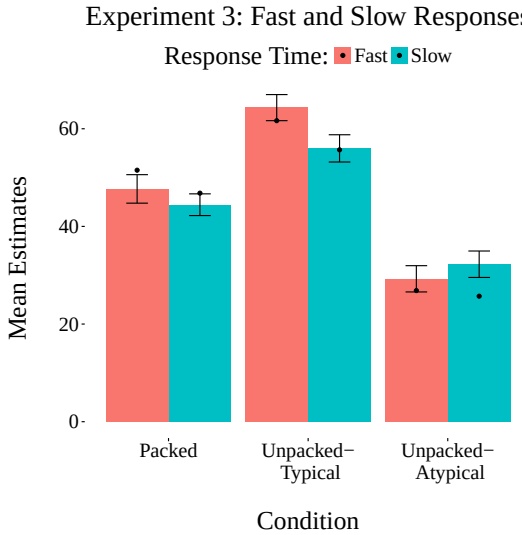

Figure 12. **Experiment 3 results: response time analysis**. Mean probability estimates for each condition divided into fast and slow trials based on a median split of the response times. Error bars represent the 95% confidence interval of the mean. Dots represent the model fits with model parameters $K = 3$, and $N = 150$ for the fast trials and $N = 190$ for slow trials.

Experiment 4

In our final experiment, we explored the possibility that cognitive load will reduce the number of samples, under the assumption that load consumes resources necessary for hypothesis generation. Therefore, we repeated Experiment 1, but put participants under cognitive load

while responding to the packed or unpacked queries. We predicted that subadditivity and super-additivity effects should become stronger under cognitive load. In addition, the effects should again depend on participants' response time, such that faster trials are expected to produce larger effects.

Participants

69 participants (28 females, mean age= 32.17, SD=7.64) were recruited via Amazon's Mechanical Turk web service and received $0.5 for their participation plus a bonus of $0.1 for every question they answered on time and $0.1 for every time they remembered whether or not an item shown after the question had a appeared within a sequence before the target question correctly.

Materials and Procedure

Materials were the same as in Experiment 1 and 3. Additionally, participants were put under cognitive load while performing the probability estimation task. On each trial, participants again first saw the cue object. Once they clicked "Next", a sequence of three random digits appeared, each remaining on the screen for 1 second before disappearing after which the next digit appeared. Participants were asked to remember these digits. Immediately afterwards, participants were asked to judge the probability of a hypothesis that could be either packed or unpacked (same as in Experiment 1). They were then shown another digit and had to indicate whether or not that digit had occurred within the sequence they had just been shown.

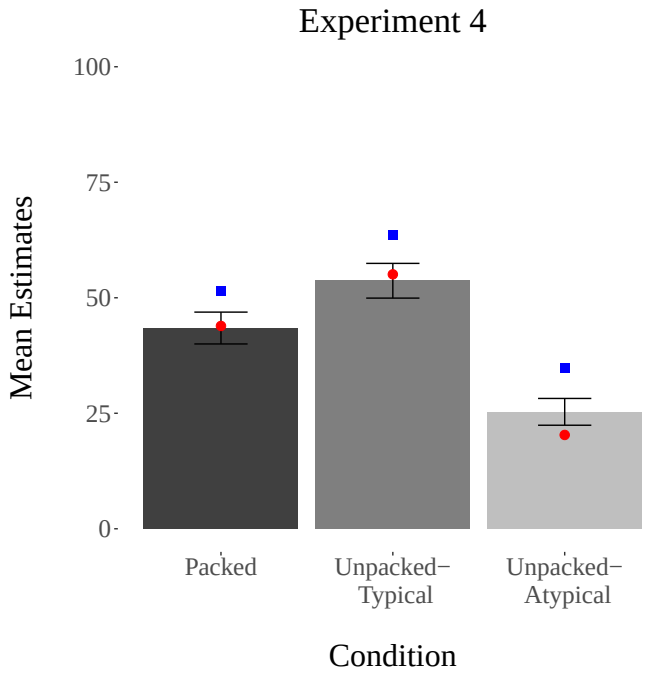

Figure 13. **Experiment 4 results**. Mean probability estimates for each condition when participants are put under cognitive load. Error bars represent the 95% confidence interval of the mean. Red dots show estimates from the MCMC model with 110 samples, assuming 2 hidden objects in the scene. Blue squares show means estimates of Experiment 1.

Results and Discussion

The mean probability estimates for each condition are shown in Figure 13. Again replicating Experiment 1, the estimates for the unpacked-atypical condition were significantly lower than for the packed condition $[t(68) = -7.31, p < 0.01]$, and the estimates for the unpacked-typical condition were significantly higher than for the packed condition $[t(68) = 4.18, p < 0.01]$. The model fits the data well with parameter values $K = 2$ and $N = 110$ with 95% confidence interval of $[74, 146]$. We see again that the best fit number of samples is substantially lower than that found in Experiment 1 ($N = 230$, with 95% confidence interval $[191, 269]$), with no overlap in the confidence intervals. The number of hidden objects K is also lower. Additionally, the cognitive load manipulation increased the effect of superadditivity (packed-atypical condition) as compared to Experiment 1 $[t(58) = 10.38, p < 0.001]$, but was not significantly different from Experiment 1 for the subadditivity effect (packed-typical condition) $[t(58) = -1.9, p > 0.05]$.

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 four experiments the model's prediction that subadditivity and superadditivity depend critically on the typicality of unpacked examples and that the superadditivity effect increases under time pressure and cognitive load.

Our work extends a line of research on using rational process models to understand cognitive biases. Most prominently, Thomas et al. (2008) 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 (Vul et al., 2014). Like other models based on Monte Carlo methods (e.g., S. J. Gershman et al., 2012; Lieder et al., 2017b, 2017a; Shi et al., 2010), our model predicts exact Bayesian inference in the limit of large sample sizes. However, cognitively bounded agents are expected to be *computationally rational* (S. J. Gershman et al., 2015): sampling takes time and effort, and hence the optimal sampling strategy will tend to generate relatively few hypotheses (Vul et al., 2014).

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 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, that are not captured by parallel sampling algorithms. First, our model posits an interplay between generation and eval-

uation 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 (Shi et al., 2010) or cued recall (Thomas et al., 2008) is that the generation and the evaluation processes are largely decoupled; the probabilities of the hypotheses already in the cache of generated hypotheses do not affect whether or not new hypotheses are generated.

The second key property of our model is the autocorrelation of hypotheses in the Markov chain. This autocorrelation arises from two sources: the non-zero rejection rate (which ensures that the chain sometimes stays at its current hypothesis for multiple time steps) and the locality of the proposal distribution (which ensures that proposed hypotheses are in the vicinity of the previously generated hypothesis). Previous models based on importance sampling or cued recall generate new candidate hypotheses independently of the hypotheses that have already been generated (i.e., the previously generated hypotheses have no impact on future hypotheses). Lieder et al. (2017a) 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 (Bonawitz et al., 2014; S. J. Gershman et al., 2012; Vul & Pashler, 2008). For example, the "crowd within" phenomenon (Vul & Pashler, 2008), which we also simulate, demonstrates that errors in numerical guesses are correlated in time, and this error is reduced if the guesses are spread out.

MCMC models with global proposal distributions will show much weaker autocorrelation compared to those with local proposal distributions, because any autocorrelation will depend entirely on rejection of proposals. Since efficient samplers have relatively low rejection rates (Robert & Casella, 2013), 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 (Abbott et al., 2015; Smith et al., 2013), further suggesting that humans use local proposal distributions.

Previous work demonstrating the effect of superadditivity (Sloman et al., 2004) 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 discrepancy is that, unlike the studies in Sloman et al. (2004), we impose a response deadline on participants. The size of the subadditivity and superadditivity effects decay with the number

[7]While an importance sampler does reproduce the dud alternative 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 (Koehler, 1994).

Table 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	?[7]	✓	✓
Crowd within		✓	✓
Anchoring & adjustment			✓

of hypothesis sampled. Subadditivity decays to almost zero with fewer samples than superadditivity as seen for the scene statistics model in Figure 14. The time pressure in Experiment 1, by restricting the number of samples, may have rendered subadditivity observable, whereas the superadditivity effect is observable in both. Time pressure in Experiment 3 and cognitive load in Experiment 4 strengthened some of the effects, but did not consistently strengthen both effects. Thus, more experimental work is needed to understand the role of time pressure and cognitive load.

Our results cannot be explained by simpler heuristics like anchoring and adjustment. Although anchoring to a low probability hypothesis can account for superadditivity (probability estimates are adjusted upwards), anchoring to a high probability example does not explain subadditivity, since the high probability hypothesis still has lower probability than the total probability of the focal space (e.g., the probability of "chair" is lower than the probability of seeing any object starting with the letter "c"). Thus, adjustment away from the low probability hypothesis towards the normatively correct probability cannot lead to a probability estimate higher than the answer to the packed query (where presumably no anchoring occurs).

Other effects like the conservatism bias could also potentially be captured by variants of our model. Conservatism bias has previously been modeled using noisy retrieval of memories (Dougherty, Gettys, & Ogden, 1999; Marchiori, Di Guida, & Erev, 2015) and can be reproduced in our model in the same spirit by allowing noisy initialization. Due to the discreteness and resulting low resolution of probability estimates allowed by a limited number of samples, even a few initial samples from the focal space might over-represent its probability. When queried focal space has low probability, the chain is initialized there and the few initial hypothesis generated from the focal space could give it higher probability than the true posterior. When the queried focal space instead has high probability, it will be under-represented (as predicted by conservatism) if there are more samples from its complement space. If we introduce noise that causes the chain to initialize in the complement space with some small probability, this will produce a higher probability for the complement space and thus a lower probability for the focal space—i.e., conservatism. That being said, the addition of noise might interfere with our explanations of other probability judgment biases, so further modeling work is needed to explore this hypothesis.

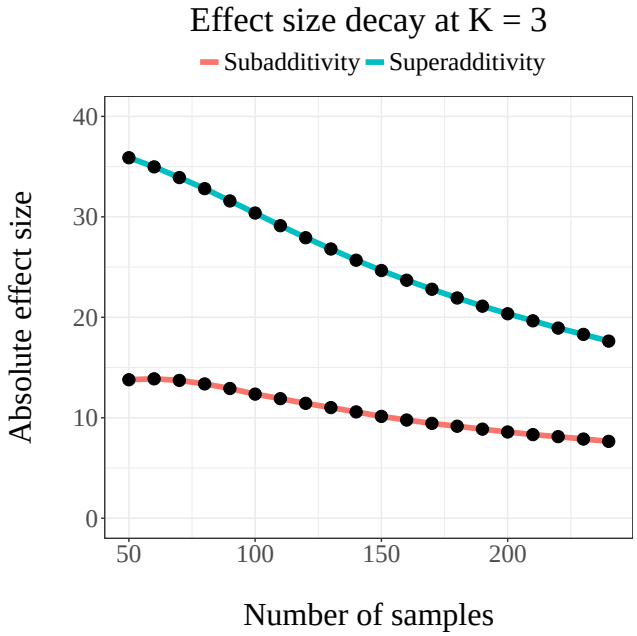

Figure 14. The effect size of subadditivity and superadditivity (calculated as the absolute difference between unpacked judgments and packed judgments, averaged over 200 chains) decays with increase in the number of samples taken. We plot this for $K = 3$ but this structure is maintained at all K. This plot shows that superadditivity decays faster than subadditivity with increase in the number of samples, and that subadditivity decays to close to zero with a smaller number of samples.

Limitations and future extensions

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 (S. J. Gershman & Goodman, 2014; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015). This mechanism might capture effects that hinge on the availability and representativeness heuristics. Our current model fails to replicate these effects because it assumes that all hypotheses are equally likely to be proposed, although they are accepted proportional to their probability. A proposal distribution that preferentially proposes certain hypotheses might help build a link between our stochastic sampling-based method and the literature on heuristics.

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 (S. J. Gershman et al., 2015; Griffiths et al., 2015; Vul et al., 2014). Studying this more directly would involve changing the incentive structure of the experiment in tandem with response deadlines and cognitive load manipulations.

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. It is important again to point out here that we expect our model to work only in domains in which humans have natural intuitions for relative probabilities of hypotheses, without requiring explicit calculation. For example, it has been shown that

humans are capable of simulating physical trajectories that they have never directly observed, making fairly accurate inferences when predicting the motion of a projectile (Téglás et al., 2011), judging mass in collisions (Sanborn & Griffiths, 2009), and judging the balance of block towers (Hamrick, Battaglia, & Tenenbaum, 2011). Furthermore, research also suggests that humans sample noisy simulations of future trajectories (Hamrick et al., 2015; Smith & Vul, 2013), 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.

Returning to the puzzle we started with, why is it that humans are sometimes so successful at probabilistic inference, and at other times so unsuccessful? We have identified one common source of inferential fallacies: computational constraints on hypothesis generation. Although this account can explain many departures from rationality, it remains puzzling why humans should fail at tasks where the hypothesis space is clearly and exhaustively enumerated—for example, in tasks that involve inferences about balls in urns (see Peterson & Beach, 1967, for a review). We suspect that people have difficulty with these tasks because their artificiality invokes an explicit, error-prone reasoning process (much like solving a math problem) rather than drawing upon intuitive knowledge of natural domains (Cohen, Sidlowski, & Staub, 2016; Evans, Handley, Over, & Perham, 2002). Studies have shown that difficult inference problems can be transformed into tractable ones simply by providing subjects with an intuitive causal structure (Cheng & Holyoak, 1985; Krynski & Tenenbaum, 2007; Tversky & Kahneman, 1980). A direct comparison between naturalistic and artificial versions of the tasks we used in our studies is challenging because scene knowledge is complex and high-dimensional (precisely why we were interested in this domain to begin with). Thus, our claims remain tentative until further experimentation yields decisive evidence.

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 and Falk Lieder for helpful comments and discussions.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference* (Vol. 122, p. 558).

Abbott, J. T., & Griffiths, T. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2950–2955.

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).

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

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.

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

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*, e1002211.

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

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

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391–416.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–552.

Cohen, A. L., Sidlowski, S., & Staub, A. (2016). Beliefs and Bayesian reasoning. *Psychonomic Bulletin & Review*, 1–7.

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

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

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, *126*(2), 280–300.

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

Dougherty, M., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*, 135–148.

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

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

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

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.

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

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

Gennaioli, N., & Shleifer, A. (2010). What comes to mind. *The Quarterly journal of economics*,

125(4), 1399–1433.

Gershman, S., Hoffman, M., & Blei, D. (2012). Nonparametric variational inference.

Gershman, S. J., & Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, *1*, 517–522.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

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

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

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

Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science: A Multidisciplinary Journal*, *32*(1), 108–154.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*(October), 777.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

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

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*, 725–743.

Griffiths, T. L., Vul, E., & Sanborn, a. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

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.*

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*, 1545–1550.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). 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.*

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

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

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

Klein, G. (1999). *Sources of power: How people make decisions*. MIT press.

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

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of*

Experimental Psychology: Human learning and memory, *6*, 107–118.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430–450.

Lieder, F., Griffiths, T., Huys, Q. J., & Goodman, N. D. (2017b). Empirical evidence for resource-rational anchoring and adjustment. *Retrieved from osf. io/zu4pt*.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2013). Burn-in , bias , and the rationality of anchoring. *Advances in Neural Information Processing Systems 25*, *25*, 1–9.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2017a). The anchoring bias reflects rational use of cognitive resources. *Psychological Review*.

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

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

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

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

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

Pecevski, D., Buesing, L., & Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, *7*, e1002294.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological bulletin*, *68*(1), 29.

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

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.

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

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.

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

Sanborn, A. N., & Griffiths, T. L. (2009). A Bayesian Framework for Modeling Intuitive Dynamics. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1145–1150).

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Schulz, E., Speekenbrink, M., & Meder, B. (2016). Simple trees in complex forests: Growing take the best by approximate Bayesian computation. *arXiv preprint arXiv:1605.01598*.

Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In *Advances in Neural Information Processing Systems* (pp.

1669–1677).

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

Shwe, M. A., & Cooper, G. (1991). An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Computers and biomedical research, an international journal*, *24*(5), 453–475.

Shwe, M. A., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., & Cooper, G. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of information in Medicine*, *30*(4), 241–255.

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

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573–582.

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

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

Sprenger, A. M., Dougherty, M., Atkins, S. M., Franco-Watkins, A. M., Thomas, R., Lange, N., & Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, *2*, 129.

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

Sternberg, R. J., & Davidson, J. E. (1995). *The nature of insight*. The MIT Press.

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

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

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–9.

Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20.

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

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

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

Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in Neural Information Processing Systems* (pp. 1955–1963).

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

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

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.

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 198.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, *6*, e1000871.

Yildirim, I., Kulkarni, T. D., 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 *Annual conference of the cognitive science society.*