Psychological Review

Point Estimate Observers: A New Class of Models for Perceptual Decision Making

Heiko H. Schütt, Aspen H. Yoo, Joshua Calder-Travis, and Wei Ji Ma

Online First Publication, February 20, 2023. https://dx.doi.org/10.1037/rev0000402

CITATION

Schütt, H. H., Yoo, A. H., Calder-Travis, J., & Ma, W. J. (2023, February 20). Point Estimate Observers: A New Class of Models for Perceptual Decision Making. *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000402

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000402

Point Estimate Observers: A New Class of Models for Perceptual Decision Making

Heiko H. Schütt[1], Aspen H. Yoo[1], Joshua Calder-Travis[2], and Wei Ji Ma[1]
[1] Center for Neural Science and Department of Psychology, New York University
[2] Department of Experimental Psychology, University of Oxford

Bayesian optimal inference is often heralded as a principled, general framework for human perception. However, optimal inference requires integration over all possible world states, which quickly becomes intractable in complex real-world settings. Additionally, deviations from optimal inference have been observed in human decisions. A number of approximation methods have previously been suggested, such as sampling methods. In this study, we additionally propose *point estimate observers*, which evaluate only a single best estimate of the world state per response category. We compare the predicted behavior of these model observers to human decisions in five perceptual categorization tasks. Compared to the Bayesian observer, the point estimate observer loses decisively in one task, ties in two and wins in two tasks. Two sampling observers also improve upon the Bayesian observer, but in a different set of tasks. Thus, none of the existing general observer models appears to fit human perceptual decisions in all situations, but the point estimate observer is competitive with other observer models and may provide another stepping stone for future model development.

Keywords: perceptual decision making, point estimate observer, Bayesian observer, observer model

A central question in cognitive science is how humans make decisions based on uncertain information about the world. This question has been studied extensively in the realm of perceptual inference. This area is particularly suited to precise quantitative modeling of decision making, as perceptual tasks can be carefully controlled and large amounts of data can be collected efficiently. Nonetheless, few models provide a unifying explanation for decisions across many different perceptual tasks.

This most successful framework for explaining human perceptual decision making is based on the Bayesian optimal solution to the task faced by participants (Banks et al., 1987; Burge & Geisler, 2011; Ernst & Banks, 2002; Geisler, 1989, 2011). In this framework, one assumes perfect use of all available information as a starting point and then adds restrictions to this optimal observer to adjust the model to human behavior. Similar models have also been successfully applied in other contexts, such as cognitive decision making (Griffiths & Tenenbaum, 2006; Tenenbaum & Griffiths, 2001; Vul et al., 2014) and motor control (Körding & Wolpert, 2004; Najemnik & Geisler, 2005; Wolpert et al., 1995). There is strong evidence that humans take into account prior information (Adams et al., 2004; Tassinari et al., 2006) and the uncertainty associated with sensory variables (Adler & Ma, 2018; Denison et al., 2018; Ernst & Banks, 2002), which are two predictions of Bayesian models. However, the success of Bayesian optimal observer models does not necessarily imply that humans perform full Bayesian probabilistic computation (Ma, 2012; Maloney & Mamassian, 2009), as near-optimal performance can be achieved by other means (Jones & Love, 2011; Ma & Jazayeri, 2014).

A key operation of the Bayesian approach is marginalization. Marginalization refers to a mathematical procedure which allows one to compute a probability distribution over a specific variable of interest (referred to as a marginal distribution), from a probability distribution over a larger number of variables (joint distribution). This is achieved by integrating "out" all variables that are not of interest. Marginalization scales badly with the dimensionality of the distributions involved. This is because the required integration can

Heiko H. Schütt https://orcid.org/0000-0002-2491-5710
Aspen H. Yoo https://orcid.org/0000-0002-3175-1881
Joshua Calder-Travis https://orcid.org/0000-0003-3764-2042
Wei Ji Ma https://orcid.org/0000-0002-9835-9083

Aspen H. Yoo is now available at Department of Psychology, University of California, Berkeley and Joshua Calder-Travis is now available at Institute of Neurophysiology and Pathophysiology, Universitätsklinikum Hamburg-Eppendorf.

The authors thank Will Adler, Shan Shen, and Yanli Zhou for sharing the data for their respective experiments, which made this project possible.

This study was funded in part by the German Research Foundation (DFG) through Grant SCHU 3351/1-1 to Heiko H. Schütt. Joshua Calder-Travis would like to thank the Grand Union ESRC Doctoral Training Partnership, and St John's College, Oxford, for DPhil studentship funding. Wei Ji Ma was supported by Grants R01EY020958 and R01EY026927 from the National Institutes of Health. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Data sets and analysis code are available at https://osf.io/x8q6j/ and https://github.com/NYUMaLab/AISP respectively. This study was not preregistered.

A preprint of this work was posted to PsyArXiv and is available at https://psyarxiv.com/bqkf4/.

Correspondence concerning this article should be addressed to Heiko H. Schütt, Center for Neural Science and Department of Psychology, New York University, 4 Washington Place, NY 10003, United States. Email: heiko.schuett@nyu.edu

often only be performed by summing the probabilities of all possible combinations of the irrelevant variables,[1] an example of the phenomenon known as the curse of dimensionality (Bishop, 2006; Hinrichs et al., 2014). While performing marginalization may be feasible in simple experimental settings (where only a few variables need to be integrated out), even simple perceptual inferences become intractable in a world with many objects, whose many features interact in complicated ways (Pouget et al., 2013). For example, Bayesian inference for the color of a single surface under a single source of illumination is not trivial, but manageable (Brainard & Freeman, 1997). Bayesian inference for the surface color of many objects, that reflect light onto each other, implies handling the joint distribution over the surface reflectances and positions of all objects. Thus, it seems unlikely that the marginalization assumed by the Bayesian observer is a good mechanistic explanation of human decision making. Indeed, for the example of inferences about color, models that work in naturalistic environments are not Bayesian (Kraft & Brainard, 1999).

Here, we present *point estimate observers*, a class of general models for decision making that avoid marginalization. A point estimate observer bases their response on values obtained by maximizing over irrelevant variables, rather than values obtained by integrating them out. To be more precise, the Bayesian observer finds the probability a possible response is correct by summing (through integration) the probabilities of all possible world states associated with that response (i.e., all possible values of the irrelevant variables). The point estimate observer evaluates a response by maximization to find the most probable world state associated with that response (i.e., most probable combination of values for the irrelevant variables).[2] By using maximization, instead of marginalization over the space of all possible world states, point estimate based inference is computationally cheaper than full Bayesian inference. Despite this difference in computation, point estimate observers can reach near-optimal performance.

One can arrive at point estimate observers through at least three routes. First, point estimate observers represent an approximation to the Bayesian observer and similar approximations are sometimes used in statistics. Second, point estimate observers may be understood as performing frequentist statistical inference. In frequentist model comparisons, models are fit and evaluated based only on the best-fitting parameters, just as the point estimate observer does for the response categories. In contrast to typical frequentist analyses, though, the point estimate observer takes the prior into account for fitting and evaluating the model. Finally, we can understand the point estimate observer as the best approximation of the posterior with a point mass. In all cases, the point estimate observer is a particularly simple or reduced incarnation of the framework, corroborating the idea that the point estimate observer implements a theoretically simple and computationally cheap solution.

By arguing for the point estimate observer on the basis of lower computational complexity, we assume that optimization is indeed easier than marginalization or integration. For general computer algorithms solving these problems, this is almost universally true (Quarteroni et al., 2000, compare Chapters 7 and 9). For convex functions for example, optimization algorithms like gradient descent achieve quadratic convergence, that is, their error scales with k^{-2} with the number of steps k, while the amount of computation scales linearly with the number of dimensions and can be parallelized across these (Boyd & Vandenberghe, 2004). In contrast, the error of sampling-based algorithms for integration scales only with $k^{-1/2}$ and the number of dimensions enters with a higher exponent than 1 ($\frac{5}{4}$ for Hamiltonian sampling e.g., see Neal, 2011). Therefore, sampling and numerical integration are certainly computationally more expensive than optimization, and we are not aware of any situation in which finding the maximum is more difficult than integration. Nonetheless, it requires a leap of faith to assume that this is also true for the brain solving real-world problems. After all, there are also problems for which marginalization or optimization have closed form solutions that require virtually no computation, and specialized methods for specific probabilistic graphical models can perform much better than the general purpose algorithms (Koller & Friedman, 2009) As we do not know the exact form of the problem, the brain solves and much less the algorithm and implementation it uses, it remains possible that the brain employs a clever combination of problem formulation and solving algorithm that avoids the computational inefficiencies of the general purpose algorithms.

Other approximate solutions to the marginalization problem have been proposed as models of human decision making before. Perhaps, the most common proposal is sampling from the posterior (Berkes et al., 2011; Deneve, 2008; Haefner et al., 2016), an idea that is based on a common method for computing posteriors in statistical analyses (Markov chain Monte Carlo [MCMC]; Gelman et al., 2013). While this proposal reduces the impact of the argument that marginalization is hard, the approximation converges to the full Bayesian optimal solution, such that the behavioral predictions with many samples are identical to the Bayesian observer. To make predictions that differ from the Bayesian observer, sampling accounts assume that only few samples are taken (Lieder & Griffiths, 2019; Vul et al., 2014).

Another way to avoid complex marginalization problems is to approximate the true posterior with a factorized distribution. By *factorized distribution,* we mean a probability distribution that is the product of independent distributions for each variable (Bishop, 2006). Hence, any patterns in the joint distribution of variables—for example, correlations—are lost when a factorized distribution is used as an approximation. This type of representation is attractive, because it allows a representation of uncertainty about each stimulus dimension and can remove the need to marginalize: A factorized distribution already contains a distribution over the relevant variable, independent from the other variables. Several techniques have been proposed to find factorized approximations. Two popular techniques, which are related to each other, are variational inference and expectation propagation (Minka, 2005). Variational inference has been proposed as a general inference scheme humans might employ in the context of the "free energy principle" (Friston, 2010). In perception research, variational inference is mostly discussed as a normative explanation for interactions within and between brain areas, that show similarities to the messages passed in message passing implementations of variational inference (Friston, 2008, 2010; Friston & Kiebel, 2009). In cognition research, the focus is more on the effects variational approximations have on decision

[1] The same criticism applies to the computation of various other aspects of high dimensional distributions like expected values, variances, entropies, or expected rewards, but for brevity, we will discuss things as if we always aim for a marginal distribution over some variable of interest.

[2] Point estimate observers could also choose the point estimate they evaluate based on different criteria than probability, but here, we consider only optimization based on the probability density.

behavior, due to splitting the representation into dimensions that are represented separately (Sanborn, 2017; Sanborn & Silva, 2013). Expectation propagation is discussed less, but has the advantage that the inferred distributions for the stimulus dimensions converge to the marginal distributions of the correct posterior, such that a better approximation of full Bayesian inference is achieved.

To qualify as a theory or explanation, a model needs to apply to a large collection of decision making tasks. At least, such a theory should capture all perceptual categorization tasks, that is, any categorical decision made on the basis of incoming perceptual information. This encompasses a wide range of decisions from simple detection (e.g., is the object I am searching for at the locations I am looking at?) through typical categorization examples (e.g., what kind of animal is that? Is this fruit edible?) to inferring a general context (e.g., what situation am I in?).

We will first describe and analyze the different observer models on a theoretical level. In this theoretical analysis, we will notice that the variational observer cannot handle situations where some combinations of variable values are consistent with one category, but are impossible under another category. By "category," we refer to sets of world states between which the observer is deliberating. This restriction excludes the application of the variational observer to a broad range of situations. Furthermore, we find that the expectation propagation observer makes the same predictions as full Bayesian inference in our tasks. We find two different sampling observers that each apply to all our tasks. The *importance sampling* observer samples from the prior under each category and uses the samples to estimate the evidence in favor of the category. The *joint sampling* observer samples from the joint posterior over the category and the proximate stimulus.

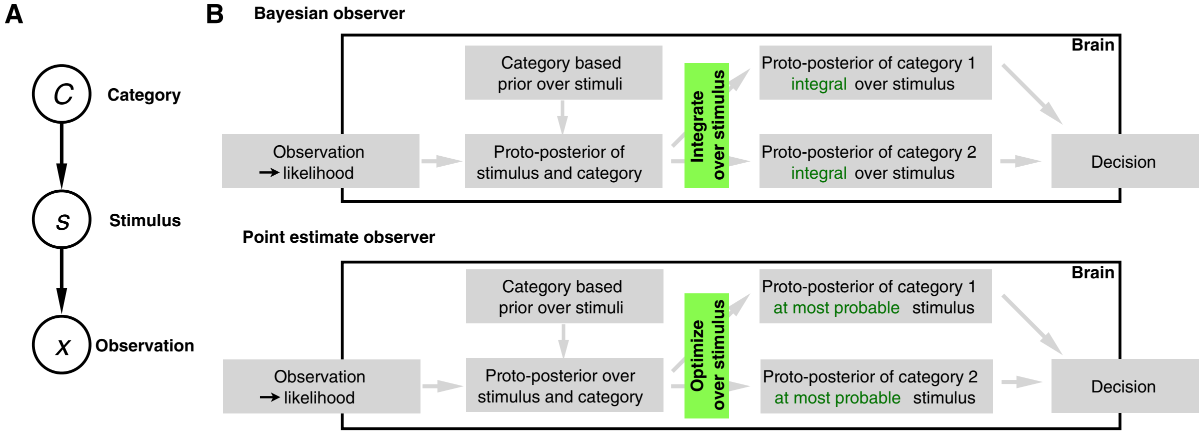
Next, we test the point estimate observer against its competitors. To make this exercise tractable, we only consider competitor models that apply to a broad range of tasks and make different predictions to the Bayesian observer without further constraints. Concretely, this rules out the variational inference observer model and the expectation propagation observer model respectively. We are thus left with the point estimate observer the two sampling observers and the Bayesian observer for empirical comparisons; for completeness, we keep the variational observer model for the one task to which it applies. To empirically test these observer models, we compare their predictions to human behavioral data. This comparison would be futile if all human behavior was well explained by the Bayesian observer model already. However, optimal observer models have been found to make predictions that deviate from human behavior (Rahnev & Denison, 2018) and some of these deviations are significantly better explained by other (task specific) models than by the Bayesian model (Adler & Ma, 2018; Stengård & van den Berg, 2019). Thus, we know there are some deviations to be explained. Additionally, there are many more data sets that were never formally tested for such deviations, partially due to the lack of a serious competing model (Bowers & Davis, 2012).

Theoretical Analysis

A simple testing ground for point estimate observers is provided by binary categorization tasks. These tasks have the following structure (see Figure 1A). There are two categories, and in each trial one of these is the true category (C). Each category is defined by a probability distribution $p(s|C)$ over a stimulus variable denoted by s. The observer makes a noisy observation of s, denoted by x. Based on this observation, the observer decides which category was presented.

Figure 1

General Model Overview


A
B
Bayesian observer
C
Category
s
Stimulus
x
Observation
Brain
Category based prior over stimuli
Observation → likelihood
Proto-posterior of stimulus and category
Integrate over stimulus
Proto-posterior of category 1 integral over stimulus
Proto-posterior of category 2 integral over stimulus
Decision
Point estimate observer
Brain
Category based prior over stimuli
Observation → likelihood
Proto-posterior over stimulus and category
Optimize over stimulus
Proto-posterior of category 1 at most probable stimulus
Proto-posterior of category 2 at most probable stimulus
Decision

Note. A: Graphical model for the general task structure. There are categories of stimuli C, to be discriminated by the participants. Each category defines a distribution over true stimulus values s. The participant makes a noisy observation x of s. Based on this observation, the participant determines which category the stimulus belongs to. B: Flow diagram for the Bayesian observer model and the point estimate observer model. Both model observers operate on the joint posterior over category and stimulus. The Bayesian observer integrates evidence over all stimulus values to compute the marginal posterior over categories. By contrast, the point estimate observer finds the most probable stimulus for each category and bases their decision on the posterior density at these stimulus values. Proto-posterior over stimulus and category here refers to the unnormalized distribution $\tilde{p}(s, C|x) = p(x|s)p(s|C)p(C)$, which is used by our observer models instead of the posterior, because the normalization is a potentially costly further processing step, which does not change the final decision. See the online article for the color version of this figure.

Bayesian Observer

The Bayesian observer bases their decision on the log-posterior ratio between the two categories:

$$d_B = \log\frac{p(C=1|x)}{p(C=0|x)} = \log\frac{p(C=1)p(x|C=1)}{p(C=0)p(x|C=0)}. \qquad (1)$$

This is the optimal decision variable, that is, choosing Category 1 if d_B is greater than 0 and Category 0 otherwise, maximizes the probability that the decision is correct. In Equation 1, the category-conditioned distributions of the observations, $p(x|C=0)$ and $p(x|C=1)$, are not directly known. Therefore, we need to rearrange this equation to replace these distributions with distributions that we do know. Per the rules of probability calculus, Equation 1 becomes

$$d_B = \log\frac{p(C=1)\int p(x|s)p(s|C=1)ds}{p(C=0)\int p(x|s)p(s|C=0)ds}, \qquad (2)$$

using the information that x is independent of C given s. Computing this decision variable requires the computation of integrals over possible stimuli, s. This may be computationally hard and scales poorly with the complexity of the stimulus (or more generally, of the world state).

Point Estimate Observer

The point estimate observer replaces the integral over the world state s with a maximization:

$$d_P = \log\frac{p(C=1)\max_s[p(x|s)p(s|C=1)]}{p(C=0)\max_s[p(x|s)p(s|C=0)]}. \qquad (3)$$

The primary computation necessary for the point estimate observer is optimization, while for the Bayesian observer, it is integration or marginalization. Optimization avoids the computational cost of marginalization, because it can be solved by gradually adapting the estimated world state to the observations and the prior knowledge. This is computationally far cheaper than evaluating a wide range of possible world states, especially in high dimensions.

The point estimate observer uses a separate estimate of the world state for each category (note the two separate maximisations over s in Equation 3). At first glance, an observer model that estimates the world state only once may seem attractive, but this observer model is ill defined for tasks that contain categories that restrict the range of possible stimuli to different sets. First, if the optimization ignored the restrictions given by the categories, it would sometimes yield world state estimates which are impossible according to all categories which means that its behavior is undefined. Second, if the optimization allowed all stimulus values that are possible under any category, it could still yield estimates of the world state that are impossible under some of the categories. In this case, the observer model would suddenly be infinitely confident in its decision and would loose its probabilistic interpretation. This second case contains taking the average prior over all categories $p(s) = \sum_i p(s|C=i)$ as a special case. Finally, if the optimization allowed only stimuli that are possible under all categories, the response probability for a category would depend mostly on its overlap with the shared space with other categories. This is problematic because (a) adding another category could then change the response probabilities for the existing categories completely and (b) the categories can be entirely disjoint, such that there are no stimuli which are possible under all categories. In fact, because there is no adequate way to take these restrictions into account within a single optimization, an observer model with only one optimization would not be able to do four of our five tasks.

By looking only at the maximum of $p(x|s)p(s|C=i)p(C=i)$ in Equation 3, the point estimate observer ignores how broad this distribution is. When one category allows a broader range of stimuli than the other, a priori (i.e., $p(s=i)$ is wider), this leads to a systematic bias in the decisions of the observer. For any single condition, a bias can be accommodated through the decision criterion. However, the direction and magnitude of this bias depends on the task and on the amount of noise in the observation x, such that different conditions in an experiment are biased to different extents. We created two variants of the point estimate observer: one with a fixed criterion, such that the model inherits the bias pattern described, and one with an optimally adjusted criterion that optimizes task performance, such that the model is equally biased or unbiased in all conditions. In statistics, such corrections are often discussed in the context of model selection. We can view observers as doing a kind of "model selection" when performing the categorization tasks studied here. In statistics, a full Bayesian solution (based on the model evidence) automatically accounts for the bias that would otherwise be introduced by model complexity, while an analysis that relies on maximum likelihood requires corrections as implemented in the Akaike information criterion (AIC) and Bayesian information criterion (BIC), the degrees of freedom in a likelihood ratio tests, or more data driven methods like cross-validation.

A related type of model called the self-consistent Bayesian observer has been proposed before (Luu & Stocker, 2018; Stocker & Simoncelli, 2008). In this model, the observer commits to a categorization of the stimulus, unlike the point estimate observer who commits to (two) values of the stimulus itself (through the two maximizations). This type of model was proposed to explain post-decision biases present when observers were first asked to categorize the stimulus, and then to report the stimulus. However, such biases are also observed and explainable when observers are not asked to categorize the stimuli (e.g., Zamboni et al., 2016). Such a scheme is suboptimal (Fleming et al., 2013), although it can be beneficial in the presence of later distortions of the stimulus representation (Qiu et al., 2020; and closely related Li et al., 2017). Also, in contrast to our point estimate observers, the original categorization decision is usually based on full Bayesian inference. Thus, the self-consistent Bayesian observer makes the same predictions for the categorization decision and is equally computationally expensive as the full Bayesian observer. A similar proposal in the cognitive literature can be found in the context of reasoning with uncertain categories (Chen et al., 2014; Murphy et al., 2012; Ross & Murphy, 1996), where humans sometimes appear to only rely on the most likely category for their decisions. In other cases, however, humans appear to take all possible categories into account (Chen et al., 2016; Murphy & Ross, 2010).

Sampling

A broad class of models for how humans may implement Bayesian decision making is based on sampling (e.g., Lieder et al., 2018; Vul et al., 2014). This kind of observer approximates the integrals in Equation 2 with a (potentially weighted) sum over samples from the posterior (or other appropriate distribution). Importantly, samples can be generated for the full joint posterior and marginalization can then be performed simply by ignoring all variables which are not of interest, so that no integrals need to be computed (Sanborn & Chater, 2016).

There are many different sampling algorithms, but typically, models of human decision making employ a form of MCMC sampling. The distinctive feature of these algorithms is that samples are taken sequentially and the next sample depends only on the current one. These sequential dependencies can explain anchoring effects if a small number of samples is taken for the approximation (Lieder et al., 2018). The origin of decision noise and some deviations from optimal decision making can also be explained by a small number of samples even if the samples are independent (Vul et al., 2014). In particular, the deviations caused by sampling can explain distortions in the handling of probabilities and reconcile the approximate Bayesian observer hypothesis with the observations that humans often make errors when handling probabilities (Zhu et al., 2020).

Most sampling accounts of human perceptual inference are concerned with estimating the proximate stimulus (s in our formulation, e.g., Moreno-Bote et al., 2011; Orbán et al., 2016; Vul et al., 2014). To do so, these approaches draw samples from the posterior $p(s|x)$. Such samples are not immediately usable for making decisions about the category C. To generate sampling observer models that apply to the decision about the category C, we had to adapt the idea slightly. We implement two sampling observer models: One important sampling observer that samples s values from the prior of each category, and a MCMC algorithm that samples from the joint posterior $p(C, s|x)$ and bases its decision on the sample frequencies of the two categories.

Importance Sampling

The first sampling-based observer we test is based on an importance sampling estimate of the integrals required for the Bayesian observer. Its implementation is straight forward. For all our experiments, we can directly sample stimulus values s_i from the prior distribution under each category $p(s|C)$ and compute an approximation to the probabilities in the ratio used by the Bayesian observer:

$$p(x|C) = \int p(x|s)p(s|C)ds \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x|s_i), \tag{4}$$

We then use the sampling estimates of these integrals for the two categories in the same way that we used the analytic solutions of these integrals for the full Bayesian observer.

Joint Posterior Sampling

The other sampling observer we implemented is based on sampling from the joint posterior over category and stimulus $p(C, s|x)$. Sampling from this posterior is not trivial, because the stimulus s may have different dimensionality for the different categories, and the stimulus distributions according to the two categories may not overlap. Thus, designing a proposal distribution for MCMC algorithms that can switch category is not trivial for general situations, especially if we additionally aim to produce proposals that are in some sense close to the current sample to achieve high acceptance rates. To avoid this problem, we simply use the prior $p(C, s)$ as a proposal distribution independent of the current sample. This choice allows the same model to apply to all tasks and has no additional parameters to be tuned to the distribution to be sampled.

Using a Metropolis–Hastings rejection sampler, this observer takes a fixed number of samples to make their decision. We then convert the number of samples for the two categories into a decision variable in analogy to the other observers as follows:

$$d_s = \log\frac{1 + \sum_i \mathbb{1}_{C_i=1}}{1 + \sum_i \mathbb{1}_{C_i=0}}, \tag{5}$$

Adding one to each category guarantees a valid decision variable for any sampling outcome. We then pass this decision variable through the same mapping to behavior as for all other observer models.

Variational Inference Observer

Variational inference is a general method that approximates complex posteriors with simpler distributions (Bishop, 2006; Blei et al., 2017; Minka, 2005). To do this, one chooses a family of simple distributions and optimizes within this family to find the distribution which is closest to the full posterior. This approach has been proposed prominently as part of the free energy principle framework (Friston, 2010), which aims not only to explain inference and decision making, but also how observers learn the world model, that is, how the observers understanding of the world is adjusted to match their oservations better. While there have been some attempts to experimentally test this type of observer model with neural data (e.g., Grabska-Barwińska et al., 2017), most discussion of the approach has remained on a theoretical level (Gershman, 2019).

Here, we use the family of factorizing distributions as the family to optimize within. This means that we require our approximate posterior to be expressible as the product of separate distributions over s and C ($q(s)$ and $q(C)$). In other words, we require that the approximate posterior "factorizes over" the categorical variable C and the nuisance variable s. The best factorizing distribution is sometimes referred to as the mean field approximation. This model observer then bases their decision on the approximate posterior. To find the approximate posterior, we search for the $q(C, s) = q(C)q(s)$ that minimizes the Kullback–Leibler (KL) divergence between q and the true posterior p, $KL(q||p)$ (KL being a quantity that increases as two distributions become more dissimilar). We then define:

$$d_V = \frac{q(C = 1)}{q(C = 0)}. \tag{6}$$

In many tasks, and most of the tasks we discuss here, variational inference fails. Any distribution q that assigns a nonzero probability to a case which has zero probability under p is an infinitely bad approximation of p as measured by the KL divergence. In most tasks, the support of s differs between categories, that is, under p there are some values of s, which happen in category $C = 1$, but never in category $C = 0$ and/or vice versa. As we try to find a factorized approximation, a combination of C and s will only have zero probability according to the approximation if one of the corresponding factors $q(C)$ and $q(s)$ has zero probability (recall the approximation is $q(C, s) = q(C)q(s)$). Thus, for any s, which lie outside of the support of one category C, either $q(s)$ must be zero or $q(C)$ must be zero. As a consequence, all q with finite KL divergence commit fully to one of the categories and/or restrict s to the overlap of all supports (i.e., to values of s which are possible under both categories). If our approximation, q commits to one category completely, there is no distribution over C and the model observer is perfectly sure of their response, which is nonsensical. If s is restricted to the shared support, the inference will be strongly biased toward the narrower category, such that the model observer would always report that category, which is also nonsensical.

For example, take a simple classification task in which the participants distinguish between positive and negative values of s, like judging whether a stimulus is tilted left or right (illustrated in Task 2 of Figure 2). Then any value of s can only happen under one of the categories, that is, there is no overlap in the supports of the two probability distributions. Thus, all acceptable factorized solutions assign zero probability to one category and commit to the other category fully.

As another example, take the collinearity judgment task of Zhou et al. (2019); discussed in detail below. There, the two entries of s are equal in category $C = 1$ and are independently drawn in category $C = 0$. Thus, all distributions with $q(C = 1) > 0$ and $q(s_1 \neq s_2) > 0$ are considered to be infinitely bad approximations. If we restrict $q(s)$ to be nonzero only for $s_1 = s_2$ then $C = 1$ is infinitely more likely than $C = 0$. Thus, the model observer would always conclude with full confidence that $C = 1$.

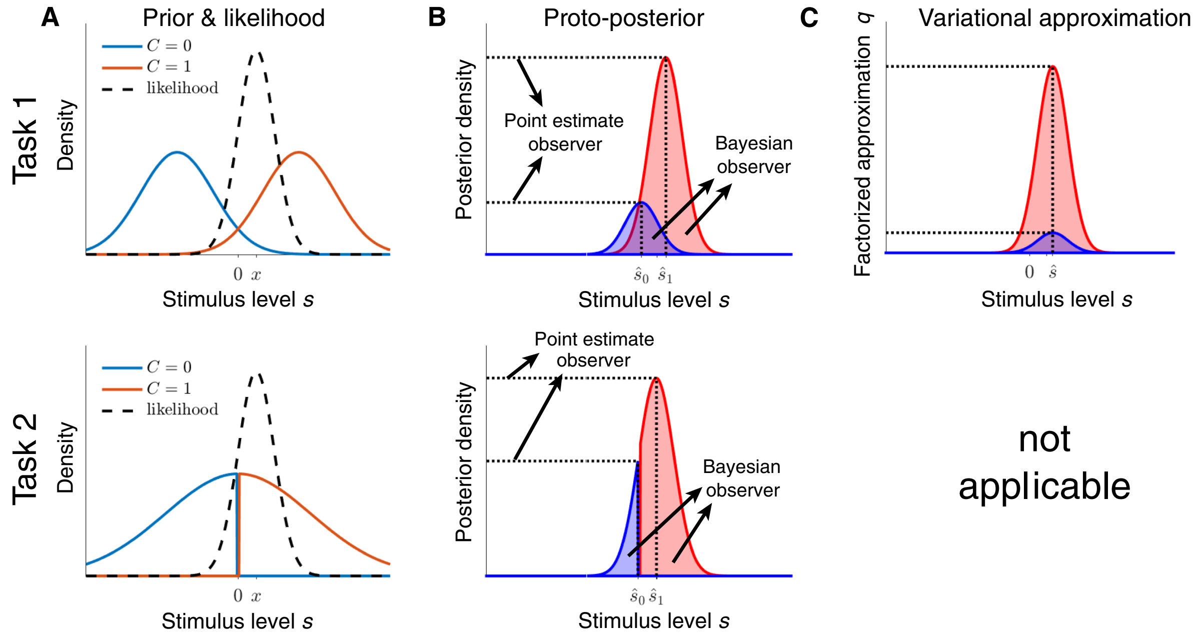
Thus, the variational inference scheme fundamentally fails as a general scheme for probabilistic inference for many typical cognitive tasks.

Expectation Propagation Observer

The expectation propagation observer is similar to the variational inference observer. It also tries to find a factorized distribution q that

Figure 2

Illustration of the Model Observers Using Two Simple Tasks (Rows)



Note. A: Prior distribution for the stimulus and the two categories, $P(s, C)$, and likelihood for an example observation x, $P(x|s)$. The colored lines each represent the distribution for s for one category. Together, they represent the joint prior over stimulus level and category. The area under the curves corresponds to the prior probability of the categories, which is equal here. In Task 1, the two categories are Gaussians with different means at ±μ. In Task 2, the two categories are the positive and negative halves of a single Gaussian. The black dashed lines represent the likelihood derived from an observation x, which all further illustrations are conditioned on. B: Joint posterior over the stimulus level and the category, $P(s, C|x)$, which can be computed by multiplying the prior and likelihood from the first column. Both curves together represent the joint posterior over category and stimulus level. The Bayesian optimal observer compares the posterior probability of the categories, which corresponds to the area under the curve, as for the prior (shaded). The point estimate observer instead compares the maximums of the posterior for the two categories (dashed lines). C: Variational mean field approximation of the posterior $q(s, C) = q(s)q(C)$, that is, an approximation which assumes a category independent distribution over the stimulus. In this factorized approximation, comparing the maximums, comparing the areas and using the computed marginal over categories all result in the same ratio between categories which is the basis for the decision of these observer models. In Task 2, the variational approximation fails entirely because variational inference never assigns any probability density in its factorized approximation to combinations of s and C for which the true posterior is 0. Expectation propagation (not shown) finds a factorized approximation in this task, but produces the exact same ratio of probabilities for the two categories as the Bayesian optimal observer. See the online article for the color version of this figure.

approximates the true posterior p. However, expectation propagation uses $KL(p||q)$ instead of $KL(q||p)$ as a measure for how well the true posterior is approximated. This might seem like a small difference, but it fundamentally changes what kind of approximation we search for. In particular, this change reverses the restriction on zeros, such that any combination which has nonzero probability under p also has to have nonzero probability under q (Minka, 2005).

Upon convergence to the global minimum, the approximation will reproduce the marginals of the true posterior distribution such that $q(C) = p(C|x)$. In all tasks presented here, participants were only asked about C. Because participants only responded to one stimulus, the expectation propagation observer would produce the same responses as the Bayesian observer (unless further approximations are enforced on the distribution, q). Thus, we do not explore expectation propagation further here.

We chose to call this model the "expectation propagation observer," as it has similarities to the approximate inference technique "expectation propagation." Technically, expectation propagation is the name for a specific message passing algorithm to compute the best factorized approximation according to the $KL(p||q)$ criterion. This algorithm requires the projection of the problem onto an exponential family distribution, which we do not perform here. As we are not aware of a succinct name for the best factorized approximation according to the $KL(p||q)$ criterion, and we do not commit to any algorithm for the computation of this approximation that we could name our observer model after, we settled on expectation propagation observer as a name, despite its inaccuracy.

Method

Formal Task Description

For all tasks, we have the following general layout: Trials come from two different categories which we index with a binary random variable C. Different categories result in different distributions over one or more true stimulus values s. For some tasks, there is an additional categorical variable L, which indexes the target location. We write $L = i$ to denote that the target location is i. (In this case, the point estimate observer maximizes over both stimulus values s and target location L).

We assume that all observer models base their inference on observations x, which are generated by adding normal or von Mises distributed sensory noise to s. The strength of the noise is measured by its standard deviation σ_n, or concentration parameter κ. This value may be different for different models and is estimated separately for each stimulus signal strength, which was varied using different approaches in the different tasks.

Probabilistic Response Model

This results in the following general probabilistic model which applies to all tasks and (a simplified version) is illustrated in Figure 1:

$$P(x, s, C, L) = P(x|s)P(s|C, L)P(C)P(L), \tag{7}$$

(Note that for the visual search task, we did not treat C and L as independent, although we did consider this choice. See Appendix B, for details.) In each experiment, the task for the observer was to infer the value of C based on x.

We add decision noise (Luce, 1959; Mueller & Weidemann, 2008), a bias (Green & Swets, 1966), and a lapse rate (Wichmann & Hill, 2001) to all our models. These additions make the observer suboptimal and are thus not justified by considerations of optimality. These additions do, however, improve the match between data and observer models as has been shown in general and, in many cases, for the tasks and data, we model here in the original publications on the data sets (Adler & Ma, 2018; Calder-Travis & Ma, 2020; Shen & Ma, 2016; Yoo et al., 2021; Zhou et al., 2019). We ensured our conclusions depended minimally on these additional components by adding them to all inference schemes considered: If, for example, bias really does exist, then because this is a component of all models, no model in particular will be favored. Hence, we would not incorrectly favor one perceptual inference scheme over another.

To implement these mechanisms, we start with the decision variable d which each observer generates. This decision variable is transformed into response probabilities by adding a bias term β_0, soft-max noise, and a lapse rate λ. This results in the following formula for the transformation:

$$P(\text{respond } C = 1) = \frac{\lambda}{2} + (1 - \lambda)\frac{\exp(\beta d + \beta_0)}{1 + \exp(\beta d + \beta_0)}. \tag{8}$$

Alternatively and equivalently, this type of noise can be implemented by exponentiating the probabilities of the categories and renormalizing to yield the probability for each response (Sanborn, Griffiths, & Shiffrin, 2010; Vulkan, 2000).

The observer models we consider do not have any additional parameters. That is, their free parameters are the sensory noise standard deviations σ_n for the different conditions, β, β_0 and λ.

A truly "optimal" observer would not lapse ($\lambda = 0$), would not have a bias ($\beta_0 = 0$) and would not have decision noise ($\beta \to \infty$). We use the term "Bayesian Observer" to refer to observer models based on this decision variable, allowing for an additive bias, decision noise, and lapses. Note that in building this "Bayesian observer," we added substantive and important additional mechanisms beyond simply assuming optimal inference (Rahnev & Denison, 2018).

For the point estimate observer, we created a fixed criterion variant and an optimal-criterion variant: The fixed criterion variant takes d_p as defined above and thus inherits any noise-level dependent bias. To generate the optimal-criterion variant, we add a noise-level dependent term to d_p such that $\beta_0 = 0$ leads to maximal performance regardless of the noise level (for an otherwise optimal observer with $\beta \to \infty$, $\lambda = 0$). In most cases, we could not find a closed form solution for the noise-level dependent bias. To estimate it, we simulated 100,000 trials for each category and calculated d for each sample (or 50,000 trials when fitting the visual search data set due to computational cost). We then found the optimal criterion with a bisection search.

Model Fitting and Comparison

The parameters for each model and task are reported in Table 1. For each task, all our models have equivalent parameters and the two sampling observers have one additional parameter for the number of samples.

For some of the models, we use in this article, we cannot directly compute likelihoods as this would require us to integrate out the measurement x, and this integral can be intractable. Thus, we chose

Table 1

Parameters for the Tasks With Limits for the Bayesian Adaptive Direct Search Algorithm for Optimization

Task	Parameter	Bounds	Probable bounds
Simple categorization	σ_i	$[e^{-3}, e^{6}]$	$[e^{-1}, e^{5}]$
	β	$[e^{-2}, e^{5}]$	$[e^{-2}, e^{5}]$
	β_0	[10, 10]	[−3, 10]
	λ	[ε, 0.25]	[0.01, 0.1]
	N_s	[1, 1000]	[1, 100]
Collinearity	σ_i	[0.1, 100]	[1, 25]
	β	[0.01, 25]	[0.1, 5]
	β_0	[−10, 10]	[−5, 5]
	λ	[0.0001, 0.5]	[0.01, 0.05]
	N_s	[1, 1000]	[1, 100]
Visual search	$\log \kappa$	[−6, 7]	[−4, 4]
	β	[0.01, 25]	[0.1, 5]
	β_0	[−10, 10]	[−5, 5]
	λ	[0.005, 1]	[0.01, 0.2]
	N_s	[1, 1000]	[1, 1000]
Outlier classification	σ	[0, 50]	[0, 20]
	β	[0, 20]	[2, 10]
	β_0	[−5, 5]	[−5, 5]
	λ	[0.0001, 0.5]	[0.0001, 0.5]
	N_s	[1, 1000]	[1, 100]
Change detection	$\bar{J}_{\text{high}}$	[0.1, 100]	[20, 40]
	$\bar{J}_{\text{low}}$	[0.1, 100]	[5, 25]
	τ	[0.1, 100]	[1, 25]
	β	[0.01, 25]	[0.1, 5]
	β_0	[−10, 10]	[−3, 3]
	λ	[0.0001, 0.5]	[0.01, 0.05]
	N_s	[1, 1000]	[1, 10]

Note. ε stands for the smallest number different from 0 in machine precision. In the visual search task, there is one $\log \kappa$ parameter for each possible value of the number of items in the display.

to estimate all log-likelihoods (LLs) using Inverse Binomial Sampling (van Opheusden et al., 2020). We ran inverse binomial sampling repetitions until the estimated variance of the log-likelihood became smaller than 4.

We fit the parameters of each model to each participant separately, by maximizing the estimated log-likelihood using Bayesian adaptive direct search (Acerbi & Ma, 2017). As bounds for the parameter optimization, we chose values reported in Table 1. We initially used 20 different starting positions, uniformly choosing positions within the plausible bounds, to reduce the probability of finding only a local maximum. For each optimization, we aimed for at least five other starting positions leading to an optimization result within two log-likelihood points of the maximum log-likelihood found, so that we are confident that model differences were not due to failed optimisations. We continued to add more starting points, if we did not satisfy the log-likelihood criterion. There was an exception to this policy: For two participant-model combinations in the visual search task, even after running the fitting from 180 starting points, there were not five runs which ended within two log-likelihoods of the maximum log-likelihood found.

To estimate the log-likelihoods at the candidate optima more accurately, we reevaluated the log-likelihood using 100 repetitions of our inverse sampling approach, which each generated estimates of variance 4 or less. Hence, after 100 repetitions, the estimated variance of the log-likelihood at the optimum is 0.04, that is, the standard deviation is 0.2. We then took the parameter combination corresponding to the minimum negative log-likelihood over our runs as the maximum likelihood estimated parameters. We report differences in raw log-likelihood values for model comparison, which are equivalent to both AIC and BIC values, as the number of parameters does not differ between models.

For plotting model fits against behavior, we simulated new data sets of the same size as the original data sets. On a participant-by-participant basis we took the best fitting parameters and used these to generate simulated responses to the stimuli that were used in the real experiment.

To keep the computational requirements for the sampling observers manageable, we restricted the number of samples taken to be less than 1,000. Restricting the number of samples is also of theoretical interest; with a sufficiently large number of samples, sampling observers should converge to the Bayesian observer. Because we are interested in how these models may diverge, specifically, how sampling observers may fit human data better than a Bayesian observer, we are interested in lower sampling regimes. When the fits of these observer models converge to this bound while still yielding worse goodness-of-fit than the Bayesian observer, we interpret this as evidence that the sampling observers would eventually converge to the Bayesian observer if we had more computational resources available.

Data

Data sets and analysis code are available at https://osf.io/x8q6j/ and https://github.com/NYUMaLab/AISP or dx.doi.org/10.17605/OSF.IO/X8Q6J. This study was not preregistered, and all data were collected for previous publications. Ethics approval from the institutional review boards were obtained for the original publications and is reported in these articles (Adler & Ma, 2018; Calder-Travis & Ma, 2020; Shen & Ma, 2016; Yoo et al., 2021; Zhou et al., 2019).

Model Evaluations

As we aim to find observer models that generalize to a wide range of tasks, we chose five typical perceptual decision making tasks to test the point estimate observer on. For all five tasks, data had been collected and published before. For the first task, we chose the categorization task by Adler and Ma (2018), which showed evidence for non-Bayesian decision making and confidence reports. Second, we chose a study on colinearity judgements by Zhou et al. (2019), which again showed a slight preference for non-Bayesian decision making and emphasized the necessity to include decision noise and lapse rates. Third, we chose a visual search task from Calder-Travis and Ma (2020), to collect more information on how information is pooled from multiple items in a single stimulus. Previous Bayesian and non-Bayesian models had fit this data set approximately equally well. Fourth, we chose an outlier classification task by Shen and Ma (2016), which provided strong evidence in favor of Bayesian inference and especially for near-optimal integration over target locations. Finally, we chose a working memory change detection task by Yoo et al. (2021), which showed evidence in favor of Bayesian decision making. Both Bayesian and non-Bayesian models had performed well in this data set. These studies cover a range of tasks that are often employed, computations required over different subhypotheses, and degrees of support for Bayesian optimal decision making. We provide details of the individual experiments in Appendix A and derivations of the decision rules in Appendix B.

Simple Categorization

In the simple categorization task (Adler & Ma, 2018), participants were asked to report whether a Gabor or ellipse stimulus came from a narrow central category or from a wider category with the same mean (Figure 3A). The precision of the observation was varied by changing the aspect ratio of the ellipse or by changing the contrast of the Gabor patch. For both types of stimuli, there were six precision levels. Each participant saw either Gabors or ellipses only.

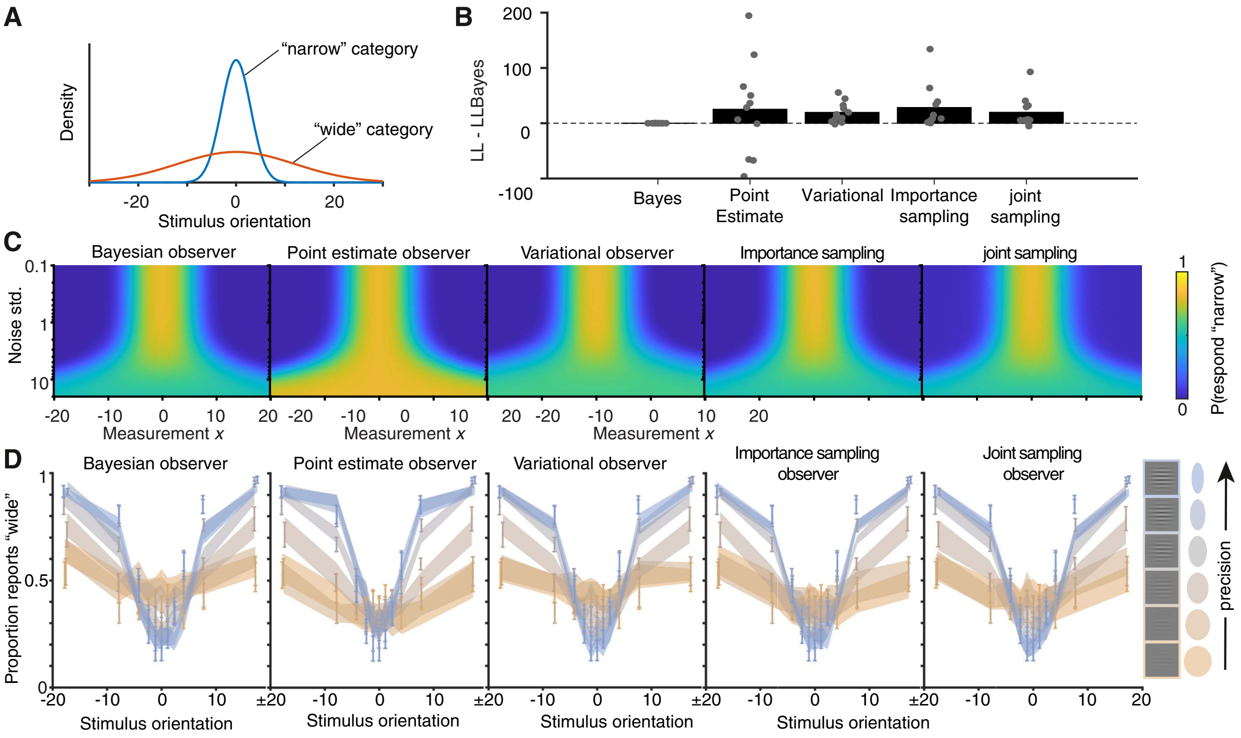
Qualitatively, participants showed the patterns expected for rational behavior (Figure 3C). At all reliability levels, participants increasingly preferred the wider Category 2 when presented with more tilted stimuli. Also, with decreasing precision this dependance became flatter and, at small precisions, the preference for the narrow Category 1 at small tilts diminished. These patterns are all qualitatively consistent with optimal decision behavior. However, the relationship between the broadening of the curve and the scaling of the peak is quantitatively different from the optimal Bayesian observer prediction.

For this task, we evaluate the full Bayesian observer, the point estimate observer with a fixed criterion, the two sampling observers, and a variational inference observer. We do not evaluate the point estimate observer with the optimal criterion because, in this task, this observer is equivalent to the Bayesian observer (see Appendix B).

We find that all models can capture the qualitative trends in the data. In the formal model evaluation, the point estimate observer performs better (on average 25.1 log-likelihood points) than the Bayesian observer. However, the relative evaluations differ dramatically between participants, so that the conclusion is not consistent across participants. The variational inference observer has an average performance between the Bayesian and point estimate observers (on average 19.5 log-likelihood points better than the Bayesian observer and 5.6 points worse than the point estimate observer). The difference to the Bayesian observer for the variational observer is considerably less variable than for the point estimate observer, but

Figure 3

Model Fits for the Simple Categorization Task (Adler & Ma, 2018)



Note. A: Illustration of the stimulus distributions under the two categories: Participants were asked to judge whether a stimulus came from Category 1 or Category 2 based on the orientation of an ellipse or of a drifting Gabor patch. B: Log-likelihood comparison against the Bayesian observer. Gray dots represent individual participants, the black represents their average. The sampling observers have the number of samples as an additional parameter for each participant. Typical corrections like AIC or BIC are very small compared to the differences between models observed for this task. This is the only task that we consider where the variational inference observer is applicable. For this task, we do not show the point estimate observer with an optimal criterion because it is equivalent to the Bayesian observer. C: Illustration of the decision boundaries of the five different observers. For each observer model, the probability for responding Category 2 is plotted against the standard deviation of the noise and against the measurement, x, which is assumed to be distributed normally around the true stimulus level, s. D: Model predictions of the five observer models (shaded regions) plotted with the data (points in center of error bars). The shaded regions and error bars represent SEMs over participants. The level of perceptual noise was varied by changing the aspect ratio of the ellipses or the contrast of the Gabor patches, respectively, in six steps as illustrated on the right. AIC = Akaike information criterion; BIC = Bayesian information criterion; SEM = standard error of the mean. See the online article for the color version of this figure.

there are still participants who are slightly better fit by the Bayesian observer than the variational one. Interestingly, the two sampling observer models also perform better than the full Bayesian observer (on average 28.3 log-likelihood points better for the importance sampling observer, and 19.7 for the joint sampling observer). The sampling models have one parameter more for each subject than the other models. If we applied a correction for that like AIC (1 likelihood point) or BIC (3.8 likelihood points), importance sampling—the better of the two—would still marginally win against the point estimate observer model, but the two are very close.

For this task, Adler and Ma (2018) found poor performance of the Bayesian observer when this model was compared to ad hoc decision models specific to this task. It appears that the point estimate and variational inference observers can explain a part of these deviations from Bayesian decision making.

Collinearity Judgements

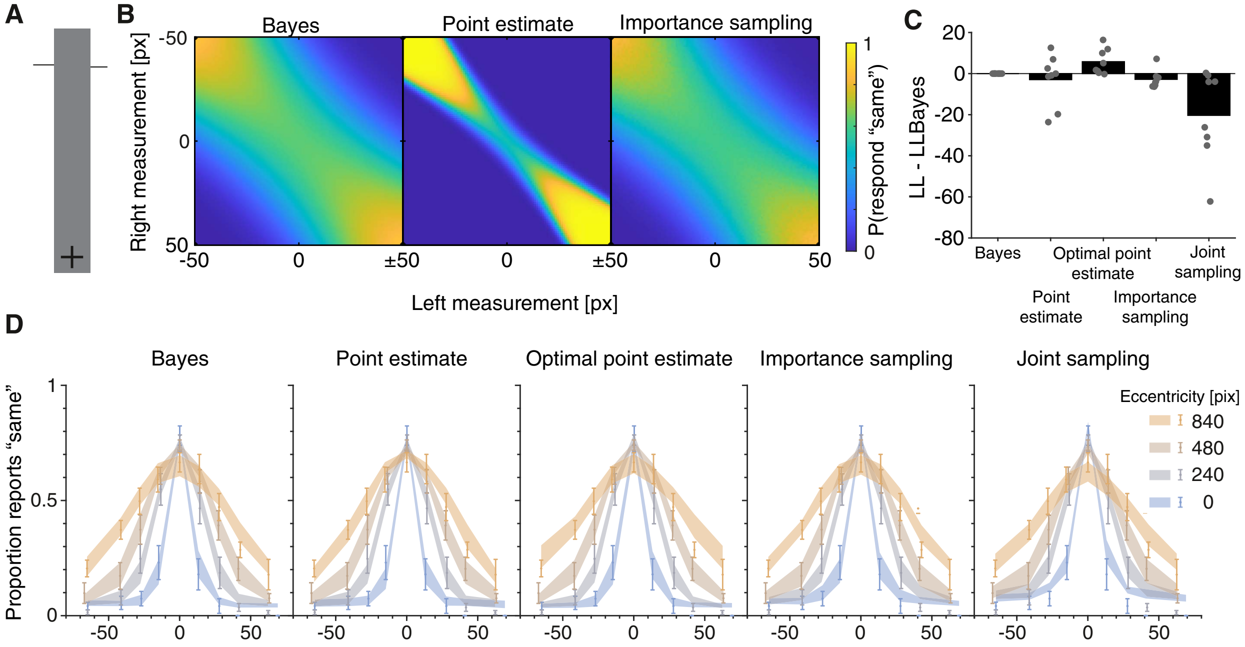
In this task, participants were asked to report whether two peripherally presented line segments disappearing behind an occluder were collinear, such that they could be part of the same straight line. The line positions were collinear in half the trials. In those trials, they were generated as a single draw from a normal distribution. In the other half of trials, they were drawn independently from the same normal distribution. The experiment was originally published by Zhou et al. (2019).

Human behavior qualitatively follows the expectations for rational behavior again (Figure 4D): Larger offsets lead to fewer collinearity reports and this effect decreases with smaller precision. We evaluate the Bayesian observer, a point estimate observer with a fixed criterion and a point estimate observer with an optimal criterion. The variational observer has no sensible interpretation for this task (see Appendix B and discussion above).

For this task, all models can capture the data well both qualitatively and quantitatively. In the formal model comparison, the point estimate observer with optimal criterion has a slight advantage of 5.81 log-likelihood points, and the point estimate observer with a constant criterion had a slight disadvantage of 2.96 log-likelihood points, but these differences are very small. The two sampling observers perform worse than the Bayesian observer model (by 2.8 and 20.3 likelihood points for the importance sampling and joint sampling observer, respectively), even without penalties for their extra parameter. As these observers converge to the Bayesian observer model with increasing number of samples, this indicates that the limit of 1,000 samples we imposed, limited performance of these observers. Indeed many observers are fitted with a number of samples above 900 (4 of 8 for the importance sampling observer, 7 of 8 for the joint sampling observer). We take this as evidence that sampling does not explain any additional patterns in the participants' behavior, but just gives an imperfect approximation to the Bayesian observer model. These results are consistent with that of the original publication, where the authors found a slight advantage for a non-Bayesian ad hoc observer.

Figure 4

Model Fits for the Collinearity Task (Zhou et al., 2019)



Note. A: Illustration of the task: Participants were asked to judge whether two peripheral line segments were collinear or not. B: Proportion same responses plotted against the left and right measurements *x*. C: Log-likelihood comparison against the Bayesian observer. For this task, we evaluate the two point estimate observer models, the two sampling observers models and the Bayesian observer model. The variational observer does not work for this scenario, caused by the unequal support for the two categories. D: Predicted responses (shaded regions) and measured data (error bars) plotted against the offset between the two lines. See the online article for the color version of this figure.

Visual Search

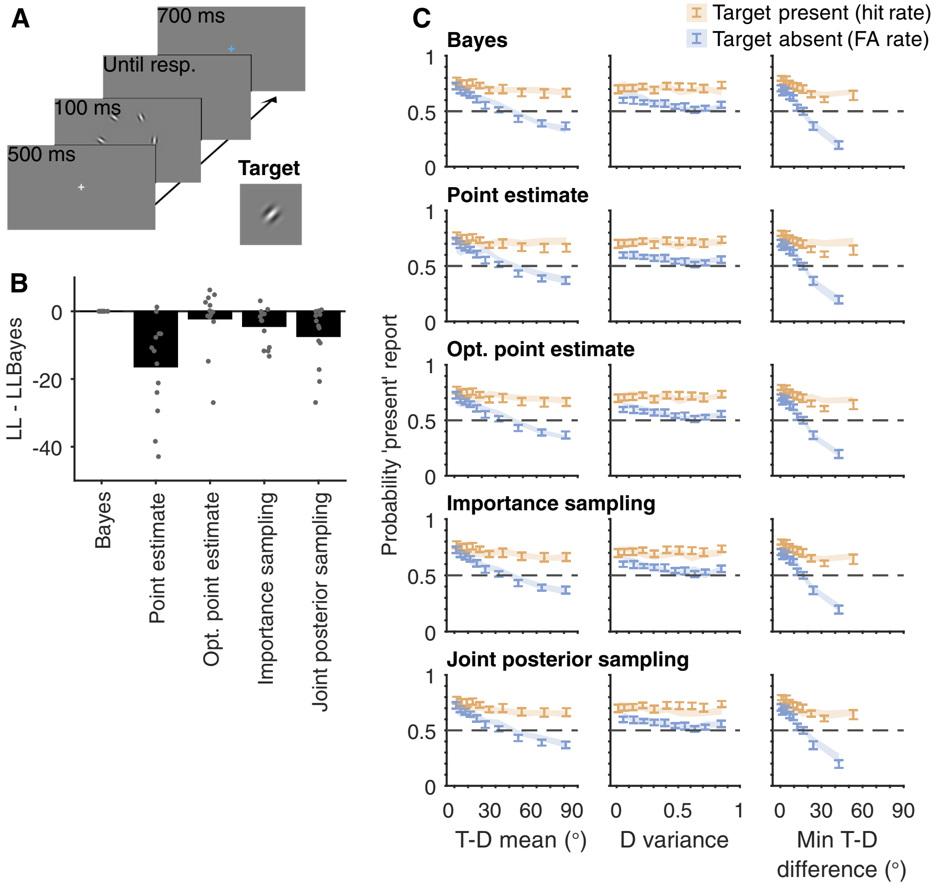
We next evaluated the models using data from a visual search task. Besides visual search being a heavily used cognitive function (Eckstein, 2011), it is interesting to study this task because participants must pool information from many items in a display. In this task, participants were asked to report the presence or absence of a target Gabor patch oriented at 45° clockwise from vertical, among distractor Gabors (Figure 5A). A display containing between 2 and 6 Gabor patches was presented for 100 ms, and then participants had as long as they wanted to respond whether the target was present using a key press. The orientation of each distractor was randomly drawn from either a uniform distribution, or a von Mises distribution centered on the target orientation, depending on the current block. Further details can be found in Appendix A, and full details in the study for which the data were originally collected (Calder-Travis & Ma, 2020).

Behavior in the task, as a function of three statistics summarizing the distractors presented on each trial, is plotted in Figure 5C. Behavior is plotted using error bars (note behavior plots are duplicated five times). On trials in which the mean of the distractors was further from the target orientation (*T–D* mean), and trials in which the difference between the target and the most similar distractor was larger (min *T–D* difference), participants were less likely to report that the target was present. The effect of distractor variance (*D* variance) was less clear.

For this task, we evaluated the performance of the Bayesian model, the point estimate observer, the optimal point estimate observer, and the two sampling observers. Again, variational inference did not provide a feasible approach (Appendix B). All models accounted well for the relationship between both hit and false alarm rate, and the three distractor statistics considered (Figure 5C). When evaluating the models using the mean maximum log-likelihood, the

Figure 5

Model Fits for the Visual Search Task (Calder-Travis & Ma, 2020)



Note. A: Illustration of the task: Participants reported the presence or absence of a target Gabor patch oriented at 45° clockwise from vertical, in a briefly presented display. B: Maximum log-likelihood found for each model, compared to the maximum log-likelihood of the Bayesian model. Dots represent the result for individual participant fits, and bars represent the mean. For this task, we evaluate the two point estimate observer models, the two sampling models, and the Bayesian observer. The variational observer does not work for this scenario. C: Behavioral data, and model fits. Data are plotted using error bars (±1 SEM, across participants), and model fits with error shading (width ±1 SEM across participants). All models captured the patterns in the data well. SEM = standard error of the mean. See the online article for the color version of this figure.

Bayesian model outperformed the point estimate and sampling models (Figure 5B; all models share the same number of parameters apart from the sampling observers which have an additional parameter). It is important to note that, while the mean maximum log-likelihood was slightly greater for the Bayesian model than the optimal point estimate model, the optimal point estimate model described many participants better than the Bayesian observer model. Even with the extra flexibility of an additional parameter, the sampling models did not in general outperform the Bayesian observer model. The fitted values for the free parameter governing how many samples the sampling observers used were very high. They often reached the upper limit set during the fitting, suggesting again that the deviation from Bayes-optimal performance, introduced by limiting the number of samples, is only detrimental to model fit.

Outlier Classification

In the outlier classification task, participants were shown four Gabor stimuli, three of which had the same orientation. They were then asked to report whether the one with a different orientation was tilted left or right. Orientations for the target and the distractors were drawn from the same normal distribution. This task was originally published by Shen and Ma (2016).

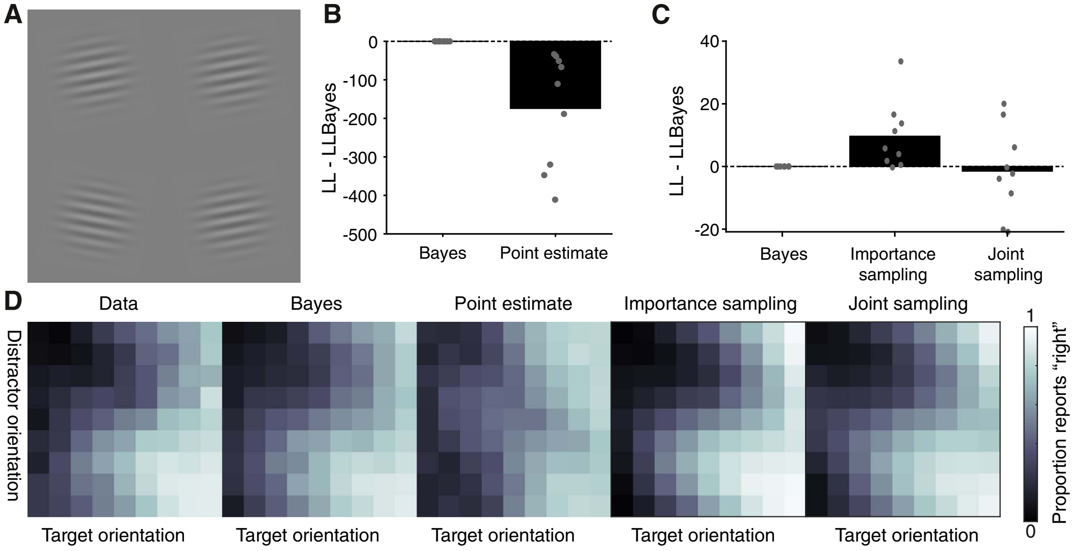
In this task, the optimal decision pattern is relatively complex, because the decision of which item is the target needs to include the orientations of all items in the display. Importantly, this leads to an S-shaped influence of the distractors' orientation on the judgements. When the distractors are only slightly tilted they can be confused with a target near the decision boundary, and thus bias the decision toward their orientation. However, strongly tilted distractors do not influence the decision, because they can only be confused with the target if the target clearly has the same tilt direction. Thus, the bias caused by the distractors diminishes at strong tilts. This pattern of decisions is also produced quite closely by participants in this task (Figure 6C).

For this task, we only evaluate the Bayesian observer, the two sampling observers, and the point estimate observer. In this case, the whole experiment was performed at a constant noise level, removing the necessity to implement the Optimal-Criterion Point Estimate Observer. Again, the variational observer does not work for this task because the different categories and target locations imply incompatible distributions for the stimulus *s*.

The fits for the models are displayed in Figure 6. Comparing the point estimate observer to the Bayesian observer, we find strong evidence in favor of the full Bayesian observer. The data plots show a clear discrepancy between the data and the prediction of the point estimate model. This is also visible in the formal model comparison: All participants are better fit by the Bayesian observer than by the point estimate observer, and the fit is on average 174 log-likelihood points better. The differences between the sampling observers and the Bayesian observer are so much smaller, that we need to plot them on a different scale to make them visible (Figure 6C). The importance sampling observer beats the Bayesian observer by a noticeable

Figure 6

Model Fits for the Outlier Classification Task (Shen & Ma, 2016)



Note. A: Illustration of the task: Participants were asked to report whether the Gabor with the different orientation was tilted left or right. B: Log-likelihood difference from the Bayesian observer, dots represent the individual participants. The bar represents the average. For this task, we compare only one version of the point estimate observer to the Bayesian one. The two variants of the point estimate observer are equal here because the noise was not varied in this experiment. The variational observer once again fails because the support of the categories does not overlap. C: As B, but for the sampling observers, separated to make the much smaller differences to the Bayesian observer visible. D: Proportion of "right" reports plotted against the orientation of the target and the orientation of the distractors, binned into nine quantiles of the normal distribution. See the online article for the color version of this figure.

margin (on average 9.6 log-likelihood points). The joint sampling observer performs about equal to the Bayesian observer (on average 1.4 log-likelihood points worse).

Change Detection

In this task, participants viewed four oriented ellipses, remembered their orientation over a working memory delay, then viewed four ellipses again. Participants indicated whether they believed the orientation of one of the ellipses changed. On every trial, there was a 0.5 probability that one of the ellipses changed orientation. If there was a change, the change was drawn from a uniform distribution and was equally probable to occur in any of the ellipses. Ellipses provided either high- or low-reliability orientation information, the high-reliability ellipses being longer and narrower. The probability of each ellipse being high reliability was 0.5, independent of the reliability of the other ellipses. This experiment was originally published as the "ellipse condition" by Yoo et al. (2021).

To maximize performance in this task, the participants should take into account the item-to-item uncertainty when making the change detection decision. For this task, we evaluate the Bayesian observer, the point estimate observer with a fixed criterion, the point estimate observer with an optimal criterion, and the two sampling observers. Again, the variational observer has no sensible interpretation for this task (see Appendix B).

All models provided similar fits, qualitatively (Figure 7C) and quantitatively (Figure 7B). All models perform similarly well across participants; the Bayesian model has an average log-likelihood two higher than point estimate model, 1.3 higher than the important sampling model, and 3.5 higher than the joint posterior sampling model. The Bayesian and optimal point estimate models perform almost identically (0.3 difference in log-likelihood). We do not consider any of the differences across models meaningful.

Performance Comparison

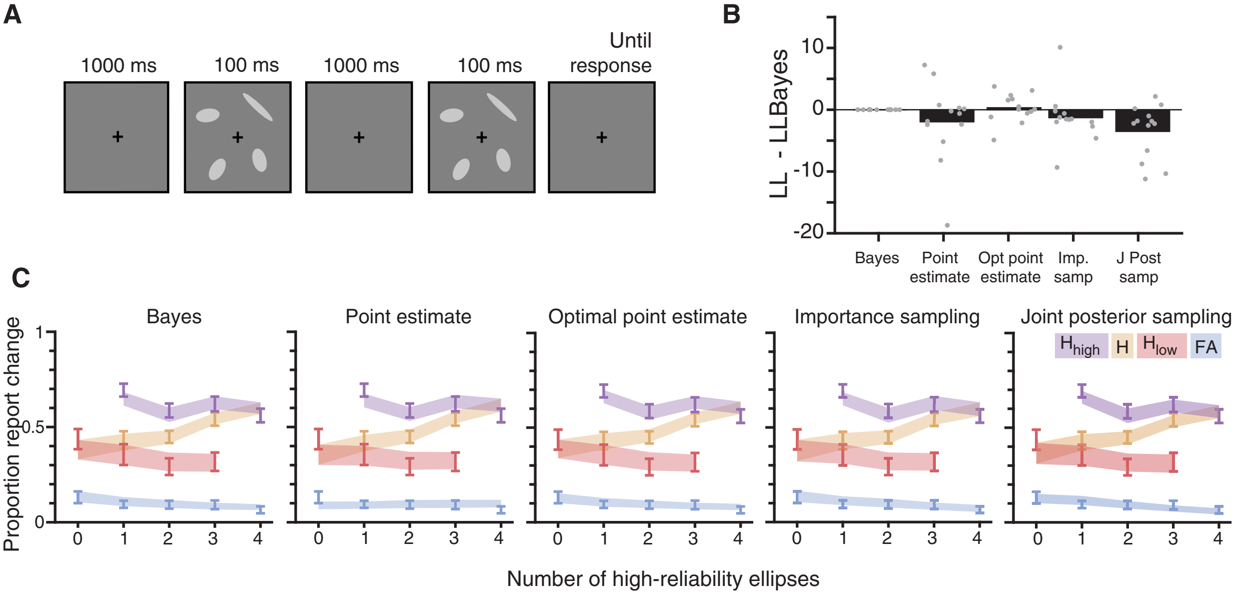
In discussing approximate observer models, one question of immediate interest is how well such approximate observers perform the tasks, that is, what level of task performance these observers achieve, as opposed to how well do these observer models match human behavior. In this section, we investigate how sensory noise affects the performance of each of these model observers, simulating responses to a large number of trials (Figure 8, 100,000 at each level).

For the plots, sensory noise was varied across a range of values. Lapse rate (λ) was set to 0, bias (β_0) was set to 0, number of samples in the sampling observer models (N_s) was set to 10. Other parameters were set to their mean value across participants from the Bayesian observer model fits. For the visual search task, we plot performance in one specific condition (3 Gabors in the display, uniformly distributed distractors).

In general, the optimal point estimate observer and the variational observer primarily lead to performance drops at the high noise levels that yield generally low performance (Figure 8B and E), if there were any performance drops for these models at all (see Figure 8C, D and F, for the collinearity, visual search, and detection tasks with

Figure 7

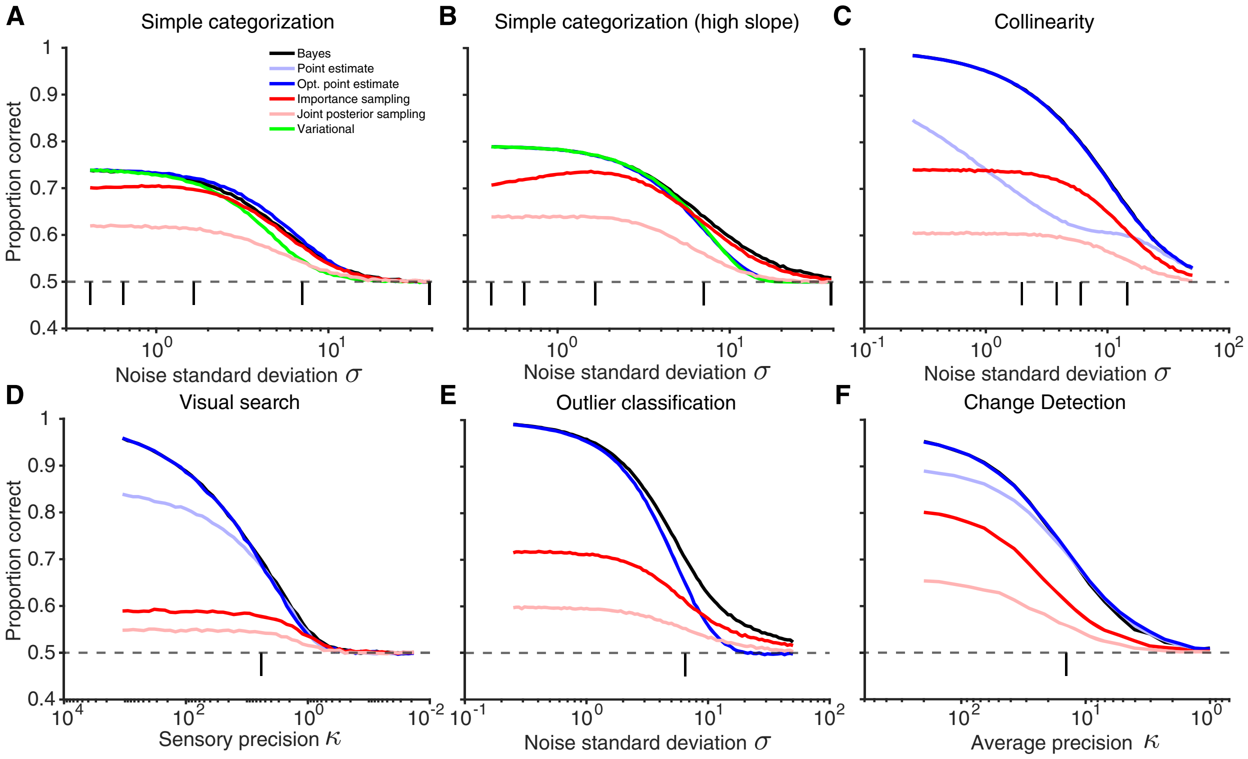
Model Fits for the Change Detection Task (Yoo et al., 2021)



Note. A: Illustration of the task: Participants indicated whether they believed an ellipse changed in orientation over the working memory delay. B: Log-likelihood differences of the two point estimate observer models and the two sampling observers models, relative to the Bayesian observer model (all models share the same number of parameters). Dots represent the individual participants. The bar represents the average. The variational observer fails for this experiment. C: Proportion report "change" as a function of number of high-reliability ellipses, conditioned on whether there was no actual change (false alarm, FA, blue), a change in a low-reliability ellipse (H_{low}, red), a change in a high-reliability ellipse (H_{high}, purple), or a change in any ellipse (hit, H, yellow). See the online article for the color version of this figure.

Figure 8

Performance Comparisons Between Observers, Plotted Against Level of Sensory Noise



Note. Unless noted otherwise, we used the mean fitted slope (β) for the Bayesian observer, no lapse rate ($\lambda = 0$), and no bias ($\beta_0 = 0$) and created 100,000 new random trials for evaluation at each noise level. Ticks mark the mean noise parameters across subjects for the conditions measured in the experiments. For the sampling observers, we took 10 samples for each trial. A: Simple categorization task, plotting against the noise standard deviation. B: Simple categorization task, as in A, but with a much higher slope (100). C: Collinearity task, plotting against the noise standard deviation. D: One condition in the visual search task (3 Gabors in the display, uniformly distributed distractors), plotting against the precision parameter of the von Mises noise distribution. E: Outlier classification task, plotting against the noise standard deviation. F: Change detection task, plotting against the average precision of the variable precision model, setting the bias to the arbitrary value of 10 for the biased point estimate observer model as 0 is outside the distribution of decision values. See the online article for the color version of this figure.

essentially no losses). In contrast, the sampling observers lead to an overall scaling of the response curves, which incurs the highest losses at low noise levels with otherwise high performance throughout (Figure 8B–F). Note, that we chose a very low number of samples here to make the differences between models clear. At higher sample numbers—as we fit to our human observers—the performance of the sampling observers converges to the Bayesian observer as expected. Finally, the nonoptimized criterion point estimate observer performs worse primarily at low noise values with high performance (Figure 8C, D and F). The comparison to the optimal point estimate observer shows the substantial performance drop observers can incur by being biased. Taken together, these results show that the different approximate observer models lead to different relationships between noise level and strength of deviation from optimal performance.

Interestingly, we observe that for the simple categorization task, the optimal-criterion point estimate observer performs even better than the nominally optimal Bayesian one (Figure 8A). This happens due to an interaction with the late noise we inject by mapping the decision variable to response probabilities, instead of applying a hard threshold. If we use a much higher slope to approximate a hard threshold, the Bayesian observer performs best as expected (Figure 8B). This serves as a reminder that subsequent decision noise can change what the optimal solution for the decision variable is, as has been observed earlier (Li et al., 2017; Tsetsos et al., 2016).

Discussion

There is a lack of task-general alternatives to the Bayesian observer model that avoid the computationally intensive marginalization operation. In response to this deficit, we introduced the point estimate observer model and evaluated its ability to account for a wide range of behavioral data. These point estimate observers evaluate the plausibility of a category based only on the world state that is most likely, if that category were indeed correct. This model directly applies to all perceptual categorization tasks, and slightly different read outs may apply to most other decision making tasks. The point estimate observer takes a statistical model of the world into account, but it is not equivalent to the full Bayesian solution.

Comparing the point estimate observer to the Bayesian observer in five tasks, we found that the point estimate observer model performs somewhat better in two tasks (simple categorization and collinearity judgements), ties in two tasks (working memory change detection task and visual search) and loses clearly in one task (outlier classification). Overall, the point estimate observer model performs competitively to the Bayesian observer model and cannot be outright rejected.

As the evidence in favor of the Bayesian observer model is mostly provided by a single task, the outlier classification task of Shen and Ma (2016), we wondered how this task might be different from the others. First, this task has the most complex statistical dependencies between stimuli, which might trigger more complex cognitive processing, although this was neither expected nor encouraged. Second, a prominent feature of this task is the detection of an outlier from a set of equal stimuli. There is some evidence that percepts corresponding to groups of stimuli are summarized, while individual stimuli with prominent differences to the group "pop out" and are preferentially processed, starting relatively early in visual processing (Müller et al., 1995; Treisman & Gormican, 1988; Whitney & Yamanashi Leib, 2018) Thus, it could be that this task is solved by specialized processes. Third, different tasks were done by different participants and it is possible that the participants in this experiment were particularly Bayesian or diligent.[3] Finally, our analysis of simulated behavioral performance for the various observer models (Figure 8) suggested that in the outlier classification task, observers may pay an especially heavy price for performing point estimate inference, compared to performing Bayesian inference. If observers could choose when to engage the Bayesian observer instead, this task would thus encourage them most to do so. These ideas are all speculative, and our data do not allow us to state any with certainty.

We also compare to two sampling observers, which generally perform well. In three of our tasks (collinearity judgements, change detection, and visual search), their fit is worse than or equal to the full Bayesian observer and the fitted sample sizes are high, which we interpret as evidence that the sampling assumption does not help to improve fits in these tasks beyond the full Bayesian observer. For the outlier classification task, the importance sampling observer beats both the Bayesian observer and the point estimate observer. For the simple classification task, the importance sampling observer fit better than the Bayesian observer and almost exactly equally well or even slightly better than the point estimate observer.

Interestingly, the importance sampling observer performed better in all tasks than the joint sampling observer. This highlights that "sampling observer" might not be a sufficiently precise description and future research should aim to understand which sampling observers are particularly good models of human decision making. A key difference here is that the importance sampling observer, which performs better throughout, takes samples independently. Thus, it does not display sequential dependencies, which earlier research used as evidence in favor of sampling based, resource constrained models (Lieder et al., 2018).

Compared to the point estimate observer, the importance sampling observer loses once while tying with the Bayesian observer (collinearity judgements), wins decisively once (outlier classification), and otherwise essentially ties with the point estimate observer. Thus, the importance sampling observer and the point estimate observer do not dominate each other either way. They are not equal though. They perform better than the Bayesian observer in different tasks (collinearity exclusively for the point estimate observer, outlier classification exclusively for the importance sampling observer) and in the simple categorization task where they perform similarly they still make different predictions.

Finally, we also tried to apply a variational inference observer which fits a factorized approximation to the posterior. This observer model was not applicable to most of our tasks, because the supports of the stimulus distributions under the two categories were often distinct.[4] As discussed, in this situation, all sensible distributions that factorize over category and inferred stimulus are considered infinitely bad approximations from the perspective of variational inference. This highlights a shortcoming of variational inference that is widely recognized in machine learning (Bishop, 2006; Minka, 2005) but has been largely overlooked when considering variational inference as a principle for human perception (Friston, 2010). This shortcoming applies broadly across different tasks, as most experiments use mutually exclusive categories of stimuli in some aspect of experimental design. In the real world, many combinations of variables are impossible as well. A potential solution would be to assume that the observer's internal world model differs from the true world model, even before the mean field approximation. The free energy principle (which is based on variational inference) makes predictions for how a parametrized internal world model should be adjusted toward fitting observations better. However, there is no general specification of what world models human observers use, and the predictions of the theory substantially depend on this choice. Thus, apart from assuming observers use the true world model, there is no uniquely defined theory that we could test. This highlights the importance of specifying the internal world model and its deviations from the true statistical structure (Rahnev & Denison, 2018). Making the only obvious choice, of requiring observers to use the true world model, is far less trivial than it sounds and dramatically constrains the range of tasks to which this approach can be applied. In comparison, other approximate inference schemes do not share this restriction.

We found that the expectation propagation observer (as defined here) has a different problem when used as a model of human perception. Namely, it converges to the same marginal distributions as the Bayesian observer. As a result, this model cannot account for observed differences between human behavior and Bayesian optimal behavior (Adler & Ma, 2018; Rahnev & Denison, 2018) as long as tasks only ask about a single feature. This is because behavior in these cases depends on the marginal distribution over that single feature. As most tasks in general, and all tasks we discuss here ask only about a single feature, we excluded expectation propagation observers from formal model comparison on these theoretical grounds. Extensions of this model which commit to other approximate posteriors, or experiments which can distinguish this type of model from others, may well provide evidence for this model though.

In many of our analyses, some of the observer models are indistinguishable from each other. This is true despite all original publications containing model recovery analyses showing that similar models to the ones we investigate here could be successfully distinguished with essentially identical model comparison techniques. In situations where models are indistinguishable, either, the

[3] Anecdotally, the participants in this task were mostly graduate student friends of the first author of the study, who might have been more motivated or knowledgeable than typical participants.

[4] "support" of a distribution simply refers to the set of all possible outcomes, that is, the set of all possibilities with nonzero probability.

two models make the same predictions matching human behavior, or the two models make different predictions, but are equally wrong in predicting human behavior. In the first case, the models are valid, but we need other experiments to distinguish them. In the second case, we need to develop better models. Both cases happen in our analyses. The visual search task (Calder-Travis & Ma, 2020) and the working memory task (Yoo et al., 2021) seem to fall under the first case of essentially indistinguishable model predictions. In the categorization task by Adler and Ma (2018), the point estimate observer model and the importance sampling observer perform very similarly well, but make distinguishable predictions. We thus have to conclude that our models are not yet perfect for some data we have, but at the same time many data sets are not helpful for contrasting different approximations to the Bayesian observer.

Having established the point estimate observer model as a viable alternative to the Bayesian observer, it provides a new starting point for model development. Here, we intentionally present the point estimate observer in a simple form. Future versions may extend the point estimate observer. For example, the point estimate observer could be extended to more complex situations like processing over space or over time where the general principle of finding the most likely state of the world instead of the full distribution of possible world states should still apply. Also, one could consider variants, which allow for multiple optimisations, for example, to explore subhypotheses, such as possible target locations in our search tasks. Finally, one could aim at unifying the point estimate observer model with other observer models. One such unification could be based on exponentiating and renormalizing the posterior with some exponent α. For $\alpha = 1$, this yields the original Bayesian observer, for large α, the posterior collapses around the maximum of the posterior and thus toward the point estimate observer. This rescaling does not simplify the marginalization but is compatible with any of the approximate solutions available for full Bayesian inference (e.g., sampling or variational inference). Alternatively, one could try observer models which marginalize over some variables and use a point estimate for others (Lee & Ma, 2021) or other combinations of intermediate complexity. These mixtures open the possibility that humans adaptively apply different inference algorithms depending on task demands (Tavoni et al., 2022). Comparisons between these extensions and to the Bayes-optimal solutions may be more informative than the global question whether our observer model is a complete model of human behavior, just as these more detailed questions are more informative when testing human behavior against optimal behavior (Rahnev & Denison, 2018).

Response biases are a frequently observed type of deviation from optimality (Rahnev & Denison, 2018). For example, participants tend to favor one of the responses, favor repeating or switching responses, or have similar preferences, which do not improve task performance (Braun et al., 2018; de Gee et al., 2017; Urai et al., 2019). Such biases are usually modeled as shifts and miscallibrations of the decision criterion, or equivalently, as wrong prior expectations. This approach works for all observer models that produce a continuous decision variable, that is thresholded to make a decision. We have used this approach for all observer models here to model biases toward one of the response categories (see Methods section). For the Bayesian observer model, this modeling approach is a bit odd, as it implies that the participants actively distort an inherently unbiased estimate, for which the optimal criterion is always the same. In contrast, the point estimate observer has to adjust the criterion frequently anyway, as the optimal criterion varies between situations and the computations it performs do not yield the optimal criterion. The inclusion of a bias to correct for miscalibrated criterion to fit human behavior is therefore justified for the point estimate observer. Our results provide some evidence that observers adjust their criteria toward optimal criteria, because the point observer model with optimally set criteria consistently models human decisions better than the one with a fixed criterion on the decision variable.

One interesting alternative explanation for the question how humans may cope with the marginalization problem is that humans may reuse computations carried out for past inferences, to produce an approximate posterior (amortized inference; Dasgupta et al., 2018). Such approximations are expected to be more accurate for frequently experienced situations (Dasgupta et al., 2020). Alternatively, observers may stop their internal computations early, resulting in an imperfect approximation (Sanborn, Griffiths, & Navarro, 2010; Shi et al., 2010; Vul et al., 2014). Such approximations can be justified as resource rational (Gershman et al., 2015; Lieder & Griffiths, 2019), that is, as the best solution achievable with a given computational resource budget. These ideas have mostly been applied to sampling-based approximations, but they could also be applied easily to the point estimate observer. If the optimization to find the most probable stimulus is an iterative procedure, it could be stopped early to save computational resources. Additionally, choosing the starting values through an amortized procedure would improve convergence speed substantially. This would result in estimates being biased toward the initial or amortized estimates, as is the case for the sampling models, a feature that may account for anchoring effects (Dasgupta et al., 2018, 2020; Lieder et al., 2018). In contrast to the idea that observers may stop computation early, our fits of the sampling observer models yield high estimates for the number of samples. In many cases, the full Bayesian observer is a better fit than the sampling observers with 1,000 samples taken per category. While our results generally favor the sampling observers, different tasks lead to extremely different sample sizes or levels of sampling efficiency, which will eventually require explanation.

Future studies could also investigate how the point estimate observer might be implemented in the brain. For the Bayesian observer and the variational observer, some neuronal implementation solutions have been proposed and continue to be developed (e.g., Beck et al., 2008, 2011; Haefner et al., 2016; Ma et al., 2006; Parr et al., 2019; Zemel et al., 1998). Given that the variational observer model already requires optimization, it seems probable that a neural network could be designed that implements the necessary optimization for the point estimate observer (similar to Deneve et al., 1999).

It would also be useful to test the point estimate observer on more empirical data. As the point estimate observer requires no marginalization, comparing the point estimate observer to the full Bayesian observer allows us to judge whether humans do or do not use marginalization. To do so effectively, one should focus on experiments where the point estimate observer and the Bayesian observer produce different predictions.

The analogy of our point estimate observer model to frequentist statistics may provide some intuition for when the point estimate observer can make different predictions to the Bayesian observer model. In a frequentist context, the inference that our point estimate observer performs corresponds to the likelihood ratio test,[5] which is often the most powerful test (Neyman et al., 1933). If the model is correct and the sample is sufficiently large, the likelihood converges to a function with a narrow peak around the true state of the world under quite general conditions (Wald, 1949). In this case, Bayesian and frequentist analysis largely agree (Gelman et al., 2013, Chapter 4.2). Conversely, the point estimate observer and the Bayesian observer can only disagree if either the assumed model is wrong, or the sensory information is sufficiently weak that the estimates are far from convergence. Situations with substantial deviations between the true model and the model assumed by the subjects may exist, but are problematic for experiments, because it is almost impossible to experimentally fix or determine the model assumed by the subjects in this case. These observations suggest that the most promising situations to distinguish Bayesian and point estimate observers are situations in which the subject is particularly uncertain about some intermediate sensory variables, such as our s. In particular, experiments in which the width of the posterior can be manipulated independent of the overall probabilities of the categories should be informative. The recommendation to look at the high-uncertainty regime is further supported by our observation that the biggest performance differences between the Bayesian optimal observer and the point estimate observer occur at high noise levels.

If one wanted to know whether humans use factorized representations as the variational observer or the expectation propagation observer do, it might be particularly informative to run experiments in which participants are asked to simultaneously classify the stimulus on two different dimensions, C_1 and C_2. Building in interesting interactions between the two sets of categories might make the experiment even more powerful. When we ask about multiple features simultaneously, we may even detect a difference between expectation propagation and full Bayesian inference. However, we do not know of an experiment that allows this distinction. Indeed, the design of experiments that reliably distinguish the different observer models is further hampered by sensory noise, decision making noise and cognitive response distortions, which may all distort the outcomes of such experiments.

In sum, the point estimate observer demonstrates the feasibility of a general perceptual decision model that does not rely on marginalization. Although it does not explain all of human decision making, it outperforms the Bayesian observer and our sampling observer models in some tasks and is thus a competitive model in its own right. Also, our results highlight the importance of comparing Bayesian observer models to other observer models like the point estimate observer, since a data set which is well described by many observer models might provide little evidence in favor of either of them. Thus, even if the point estimate observer is not ultimately the correct model, it certainly extends our toolbox for modeling human behavior. Specifically, it may serve as a comparison model, helping us to determine which tasks provide evidence for humans using marginalization.

[5] By including the prior probabilities and treating the two response categories as the models to be compared.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1834–1844). Curran Associates.

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), Article e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Banks, M. S., Geisler, W. S., & Bennett, P. J. (1987). The physical limits of grating visibility. *Vision Research*, *27*(11), 1915–1924. https://doi.org/10.1016/0042-6989(87)90057-5

Beck, J. M., Latham, P. E., & Pouget, A. (2011). Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience*, *31*(43), 15310–15319. https://doi.org/10.1523/JNEUROSCI.1706-11.2011

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021

Berkes, P., Orban, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83–87. https://doi.org/10.1126/science.1195870

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. https://doi.org/10.1037/a0026450

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *JOSA A*, *14*(7), 1393–1411. https://doi.org/10.1364/JOSAA.14.001393

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive history biases result from confidence-weighted accumulation of past choices. *The Journal of Neuroscience*, *38*(10), 2418–2429. https://doi.org/10.1523/JNEUROSCI.2189-17.2017

Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, *108*(40), 16849–16854. https://doi.org/10.1073/pnas.1108491108

Calder-Travis, J., & Ma, W. J. (2020). Explaining the effects of distractor statistics in visual search. *Journal of Vision*, *20*(13), Article 11. https://doi.org/10.1167/jov.20.13.11

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Decision making under uncertain categorization. *Frontiers in Psychology*, *5*, Article 991. https://doi.org/10.3389/fpsyg.2014.00991

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2016). Eyetracking reveals multiple-category use in induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(7), 1050–1067. https://doi.org/10.1037/xlm0000222

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81. https://doi.org/10.1016/j.cognition.2018.04.017

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441. https://doi.org/10.1037/rev0000178

de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *ELife*, *6*, Article e23232. https://doi.org/10.7554/eLife.23232

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation*, *20*(1), 91–117. https://doi.org/10.1162/neco.2008.20.1.91

Deneve, S., Latham, P. E., & Pouget, A. (1999). Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience*, *2*(8), 740–745. https://doi.org/10.1038/11205

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, *115*(43), 11090–11095. https://doi.org/10.1073/pnas.1717720115

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), Article 14. https://doi.org/10.1167/11.5.14

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *Journal of Neuroscience*, *33*(49), 19060–19070. https://doi.org/10.1523/JNEUROSCI.1263-13.2013

Friston, K. (2008). Hierarchical models in the brain. *PLOS Computational Biology*, *4*(11), Article e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267–314. https://doi.org/10.1037/0033-295X.96.2.267

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

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

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. https://doi.org/10.1126/science.aac6076

Grabska-Barwińska, A., Barthelmé, S., Beck, J., Mainen, Z. F., Pouget, A., & Latham, P. E. (2017). A probabilistic approach to demixing odors. *Nature Neuroscience*, *20*(1), 98–106. https://doi.org/10.1038/nn.4444

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Haefner, R., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, *90*(3), 649–660. https://doi.org/10.1016/j.neuron.2016.03.020

Hinrichs, A., Novak, E., Ullrich, M., & Woźniakowski, H. (2014). The curse of dimensionality for numerical integration of smooth functions II. *Journal of Complexity*, *30*(2), 117–143. https://doi.org/10.1016/j.jco.2013.10.007

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(14), 1–16 (ECVP Abstract Supplement). https://doi.org/10.1177/03010066070360S101

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. https://doi.org/10.1038/nature02169

Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences*, *96*(1), 307–312. https://doi.org/10.1073/pnas.96.1.307

Lee, J. L., & Ma, W. J. (2021). Point-estimating observer models for latent cause detection. *PLOS Computational Biology*, *17*(10), Article e1009159. https://doi.org/10.1371/journal.pcbi.1009159

Li, V., Herce Castañón, S., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLOS Computational Biology*, *13*(8), Article e1005723. https://doi.org/10.1371/journal.pcbi.1005723

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1. https://doi.org/10.1017/S0140525X1900061X

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25* (1), 322–349. https://doi.org/10.3758/s13423-017-1286-8

Luce, R. D. (1959). *Individual choice behavior*. John Wiley.

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *ELife*, *7*, Article e33334. https://doi.org/10.7554/eLife.33334

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, *16*(10), 511–518. https://doi.org/10.1016/j.tics.2012.08.010

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*(1), 205–220. https://doi.org/10.1146/annurev-neuro-071013-014017

Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience*, *1*(6), 783–790. https://doi.org/10.1038/nn.2814

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155. https://doi.org/10.1017/S0952523808080905

Minka, T. (2005). *Divergence measures and message passing* (Tech. Rep. No. MSR-TR-2005-173). https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496. https://doi.org/10.1073/pnas.1101430108

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–494. https://doi.org/10.3758/PBR.15.3.465

Müller, H. J., Heller, D., & Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Perception & Psychophysics*, *57*(1), 1–17. https://doi.org/10.3758/BF03211845

Murphy, G. L., Chen, S. Y., & Ross, B. H. (2012). Reasoning with uncertain categories. *Thinking & Reasoning*, *18*(1), 81–117. https://doi.org/10.1080/13546783.2011.650506

Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When dopeople integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 263–276. https://doi.org/10.1037/a0018685

Murray, R. F., & Morgenstern, Y. (2010). Cue combination on the circle and the sphere. *Journal of Vision*, *10*(11), Article 15. https://doi.org/10.1167/10.11.15

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391. https://doi.org/10.1038/nature03390

Neal, R. (2011). MCMC using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain Monte Carlo* (Vol. 20116022, pp. 113–162). Chapman and Hall/CRC.

Neyman, J., Pearson, E. S., & Pearson, K. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of* *the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*(2), 530–543. https://doi.org/10.1016/j.neuron.2016.09.038

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, *40*(10), 1227–1268. https://doi.org/10.1016/S0042-6989(99)00244-8

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports*, *9*(1), Article 1889. https://doi.org/10.1038/s41598-018-38246-3

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Qiu, C., Luu, L., & Stocker, A. A. (2020). Benefits of commitment in hierarchical inference. *Psychological Review*, *127*(4), 622–639. https://doi.org/10.1037/rev0000193

Quarteroni, A., Sacco, R., & Saleri, F. (2000). *Numerical mathematics (No. 37)*. Springer.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, Article e223. https://doi.org/10.1017/S0140525X18000936

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*(3), 736–753. https://doi.org/10.1037/0278-7393.22.3.736

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, *112*, 98–101. https://doi.org/10.1016/j.bandc.2015.06.008

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001

Sanborn, A. N., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, *57*(3–4), 94–106. https://doi.org/10.1016/j.jmp.2013.05.002

Schütt, H. H., Yoo, A. H., Calder-Travis, J., & Ma, W. J. (2022). *Point estimate observers*. https://doi.org/10.17605/OSF.IO/X8Q6J

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review*, *123*(4), 452–480. https://doi.org/10.1037/rev0000028

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*(4), 443–464. https://doi.org/10.3758/PBR.17.4.443

Stengård, E., & van den Berg, R. (2019). Imperfect Bayesian inference in visual perception. *PLOS Computational Biology*, *15*(4), Article e1006465. https://doi.org/10.1371/journal.pcbi.1006465

Stocker, A. A., & Simoncelli, E. P. (2008). A Bayesian model of conditioned perception. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1409–1416). Curran Associates.

Tassinari, H., Hudson, T. E., & Landy, M. S. (2006). Combining priors and noisy visual cues in a rapid pointing task. *Journal of Neuroscience*, *26*(40), 10154–10163. https://doi.org/10.1523/JNEUROSCI.2779-06.2006

Tavoni, G., Doi, T., Pizzica, C., Balasubramanian, V., & Gold, J. I. (2022). Human inference reflects a normative balance of complexity and accuracy. *Nature Human Behaviour*, *6*(8), 1153–1168. https://doi.org/10.1038/s41562-022-01357-z

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/s0140525x01000061

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15–48. https://doi.org/10.1037/0033-295X.95.1.15

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. https://doi.org/10.1073/pnas.1519157113

Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *ELife*, *8*, Article e46331. https://doi.org/10.7554/eLife.46331

van Opheusden, B., Acerbi, L., & Ma, W. J. (2020). *Unbiased and efficient log-likelihood estimation with inverse binomial sampling*. arXiv:2001.03985 [cs, q-bio, stat].

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. https://doi.org/10.1111/cogs.12101

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118. https://doi.org/10.1111/1467-6419.00106

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, *20*(4), 595–601. https://doi.org/10.1214/aoms/1177729952

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wolpert, D., Ghahramani, Z., & Jordan, M. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882. https://doi.org/10.1126/science.7569931

Yoo, A. H., Acerbi, L., & Ma, W. J. (2021). Uncertainty is maintained and used in working memory. *Journal of Vision*, *21*(8), Article 13. https://doi.org/10.1167/jov.21.8.13

Zamboni, E., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1833), Article 20160263. https://doi.org/10.1098/rspb.2016.0263

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*(2), 403–430. https://doi.org/10.1162/089976698300017818

Zhou, Y., Acerbi, L., & Ma, W. J. (2019). *The role of sensory uncertainty in simple contour integration*. bioRxiv, 350082.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. https://doi.org/10.1037/rev0000190

(*Appendices follow*)

Appendix A

Experiment Details

Simple Categorization Task

A full description of the experiment can be found in Adler and Ma (2018). We only use choice data from task B here. The experiment additionally collected confidence data, which we will ignore here for consistency with the other tasks and with the modeling framework. Furthermore, we did not analyze Task A, in which the categories had the same standard deviation, but differed in mean. This Task A was generally found not to distinguish between models in the primary publication. Data are available at https://github.com/wtadler/confidence/tree/master/human_data.

In summary, 11 participants performed the task over five sessions each, which were all shared with Task A. Experiments were performed with a linearized 2013 Apple IPad display controlled using Psychtoolbox in MATLAB (Kleiner et al., 2007).

Stimuli

The background was mid-level gray (199 cd/m^2). The stimulus was either a drifting Gabor (Participants 3, 6, 8, 9, 10, and 11) or an ellipse (Participants 1, 2, 4, 5, and 7). The Gabor had a spatial frequency of 0.5 cycles per degrees of visual angle (dva), a speed of six cycles per second, a Gaussian envelope with a standard deviation of 1.2 dva, and a randomized starting phase. Each ellipse had a total area of 2.4 dva^2 and was black (0.01 cd/m^2). Contrast and aspect ratio of the stimuli were chosen uniformly randomly per trial to vary difficulty of the task. The six contrast levels for the Gabors were 0.4%, 0.8%, 1.7%, 3.3%, 6.7%, or 13.5% and ellipses had 0.15, 0.28, 0.41, 0.54, 0.67, or 0.8 eccentricity. In Task B, which we model here, stimulus orientations were drawn from normal distributions with mean 0, and standard deviations $\sigma_0 = 3$ for Category 1 and $\sigma_1 = 12°$ for Category 2.

Collinearity Judgements

A full description of the experiment can be found in Zhou et al. (2019) and data are available at https://github.com/yanlizhou/collinearity.

In summary, eight participants were asked to judge whether two line segments presented peripherally at the two sides of an occluder were collinear or not. The experiment took four 1 hr sessions and was performed on the same Apple IPad display as the first task. We only use the collinearity judgements here, that is, we ignore the additionally collected height judgements and confidence data.

Stimuli

A dark gray occluder (23 cd/m^2) with a width of 5.6 dva was displayed against a light-gray background (50 cd/m^2). A white (159 cd/m^2) fixation dot 0.24 dva in diameter was shown in the lower central part of the occluder. The stimuli consisted of two horizontal white line segments on both sides of the occluder. The line segments were 5.6 dva long and 0.16 dva wide. The mean of the normal distributions the line positions were drawn from was set to 0, 4.8, 9.6, or 16.8 dva above the fixation location. The standard deviation of the line positions was 0.48 dva. The occluder and the fixation dot were displayed for 850 ms, followed by the stimulus for 100 ms. In each session of 200 trials contained 100 collinear and 100 noncollinear trials. The participant pressed one of eight keys, corresponding to eight choice-confidence combinations, ranging from high-confident collinear to high-confident noncollinear. Response time was not constrained. No performance feedback was given.

Shen and Ma

A full description of this experiment can be found in Shen and Ma (2016) and data are available at https://github.com/shenshan/optimal_simple.

In summary, nine participants were asked whether the one of four stimuli with a different orientation than the others was tilted left or right. Stimuli were displayed on a 21 in. Liquid crystal display (LCD) monitor with a refresh rate of 60 Hz and a resolution of 1,280 × 1,024 pixels on a gray background with a luminance of 29.3 cd/m^2.

Stimuli

The four stimuli were placed 5 dva diagonaly away from fixation. Each stimulus was a Gabor with a peak luminance of 35.2 cd/m^2, a spatial frequency of 3.13 cycles per degree, a standard deviation of 0.254 degrees of visual angle, and a phase of 0. Target and distractor orientations were independently drawn from a Gaussian distribution with a standard deviation of 9.06° around vertical. Stimuli were shown for 50 ms after a 500 ms display of only the fixation cross. Response time was not limited. After the response, correctness feedback was given by coloring the fixation dot red or green for 500 ms. The experiment consisted of three sessions with 1,000 trials each, of which the first was excluded from the analysis as training.

Change Detection

A full description of this experiment can be found in Yoo et al. (2021) and data are the Ellipse condition available at https://github.com/aspenyoo/uncertaintyWM.

In summary, 13 participants indicated whether they believed the orientation of any of four oriented ellipses changed after a working memory delay. Stimuli were displayed on a 23 in. Light emitting diode monitor with a refresh rate of 60 Hz and a resolution of 1920 × 1,080 pixels.

Stimuli

Stimuli were four, light-gray, oriented ellipses on a medium-gray background. Each ellipse could be long or short, to provide respectively higher or lower reliability information regarding the orientation of the ellipses. All ellipses had an area of 1.19 dva. The high-reliability ellipse had an ellipse eccentricity of 0.9, such that the major axis and minor axes were 1.02 and 0.37 dva, respectively. The low-reliability ellipse eccentricity was determined separately for each participant to equate performance.

On every trial, a stimulus display consisted of four ellipses. The probability of each ellipse being high reliability was 0.5, independent of the reliability of the other ellipses. The location of the first

(*Appendices continue*)

ellipse was drawn from a uniform distribution between polar angles 0° and 90°. Each ellipse after that was placed such that all ellipses were 90° apart on an imaginary annulus that was 7 dva away from fixation. Afterward, the *x*- and *y*-location of the ellipses were independently jittered −0.3 to 0.3 dva.

Visual Search

A full description of the study can be found in the original article (Calder-Travis & Ma, 2020). Data are available at https://doi.org/10.17605/OSF.IO/NERZK.

In summary, 14 participants took part in the experiment. Only data from the 13 participants who completed all four 1-hr sessions of the study is analyzed. Participants viewed an LCD monitor with 60 Hz refresh rate and 1920 × 1,080 resolution at a distance of approximately 60 cm.

Stimuli

Stimuli were comprised of between 2 and 6 Gabor patches located on the circumference of an imaginary circle, and were all 4.99 dva from the imaginary line running from the participant to the center of the screen. The six possible Gabor locations were fixed throughout the experiment. The standard deviation of the Gaussian window for each Gabor was 0.25 dva. On trials in which the target was present, one of the Gabors was oriented at 45° clockwise from vertical. The target could be at any of the six possible Gabor locations with equal probability. The orientation of all the other Gabors (i.e., the distractors) was drawn from one of two distributions that depended on the current block. In "uniform" distractor blocks distractors took any orientation with equal probability. In "concentrated" distractor blocks distractors were more likely to have an orientation similar to that of the target orientation. Every time the block type switched participants were informed and were provided with examples of the distractors in the upcoming block.

Trials began with the presentation of a fixation cross for 500 ms, followed by the presentation of the stimulus for 100 ms. Participants had unlimited time to report "target present" or "target absent" using a key press. Participants received trial-by-trial feedback on the accuracy of their responses.

Plotting

For plotting the data and model fits, we quantile binned the distractor statistics separately for each participant and data series. We computed the mean value in each bin, and the mean of this mean value across participants then determined the *x*-location of the bin. The *y*-values were determined by the mean (and SEM) across participants, of the variable plotted on the *y*-axis.

Appendix B

Observer Model Details

Here, we derive formulas for the responses of the different observer models to the individual tasks.

Simple Categorization Task

There is a single oriented stimulus. When $C = 0$, the stimulus is drawn from a Gaussian with mean 0 and standard deviation σ_0. When $C = 1$, the stimulus is drawn from a Gaussian with mean 0 and standard deviation $\sigma_1 > \sigma_0$. We assume Gaussian measurement noise. Stimulus reliability was varied in six levels, which means that the variance of the measurement noise σ_n^2 varies. For each individual trial, the distribution $P(x|C)$ thus follows the following distribution:

$$\begin{aligned}P(x|C=0) &= \int N(x;s,\sigma_n^2)N(s;0,\sigma_0^2)ds = N(x;0,\sigma_0^2+\sigma_n^2),\\ P(x|C=1) &= \int N(x;s,\sigma_n^2)N(s;0,\sigma_1^2)ds = N(x;0,\sigma_1^2+\sigma_n^2). \qquad \text{(B1)}\end{aligned}$$

Bayesian Model

Inserting these distributions into the formulas for the Bayesian model yields:

$$\begin{aligned}d_B &= \log\frac{N(x;0,\sigma_0^2+\sigma_n^2)}{N(x;0,\sigma_1^2+\sigma_n^2)},\\ &= \log\frac{\sqrt{2\pi(\sigma_1^2+\sigma_n^2)}}{\sqrt{2\pi(\sigma_0^2+\sigma_n^2)}} - \frac{x^2}{2(\sigma_0^2+\sigma_n^2)} + \frac{x^2}{2(\sigma_1^2+\sigma_n^2)},\\ &= \frac{1}{2}\log\left(\frac{\sigma_1^2+\sigma_n^2}{\sigma_0^2+\sigma_n^2}\right) - \frac{x^2}{2}\frac{\sigma_1^2-\sigma_0^2}{(\sigma_0^2+\sigma_n^2)(\sigma_1^2+\sigma_n^2)}. \qquad \text{(B2)}\end{aligned}$$

Point Estimate Model

For the point estimate observer, we derive the following decision variable:

(*Appendices continue*)

$$
\begin{aligned}
\log(q(s)) &= \langle \log p(x, C, s)\rangle_{q(C)} + \text{Const}_0,\\
&= q(C=0)\log[N(x; s, \sigma_n^2)N(s|0, \sigma_0^2)] + q(C=1)\log[N(x; s, \sigma_n^2)N(s; 0, \sigma_1^2)] + \text{Const}_0,\\
&= -q(C=0)\frac{s^2}{2\sigma_0^2} - q(C=1)\frac{s^2}{2\sigma_1^2} - \frac{(s-x)^2}{2\sigma_n^2} + \text{Const}_0,\\
&= -\frac{(q(C=0)\sigma_1^2\sigma_n^2 + q(C=1)\sigma_0^2\sigma_n^2 + \sigma_0^2\sigma_1^2)s^2 - 2\sigma_0^2\sigma_1^2 xs + \sigma_0^2\sigma_1^2 x^2}{2\sigma_0^2\sigma_1^2\sigma_n^2} + \text{Const}_0.
\end{aligned}
\tag{B7}
$$

$$
\begin{aligned}
d_P &= \log\frac{\max_s p(x|s, C=0)p(s|C=0)}{\max_s p(x|s, C=1)p(s|C=1)},\\
&= \log\frac{\max_s N(s; x, \sigma_n^2)N(s; 0, \sigma_0^2)}{\max_s N(s; x, \sigma_n^2)N(s; 0, \sigma_1^2)},\\
&= \log\frac{\max_s N\left(s; \frac{\sigma_0^2 x}{\sigma_0^2+\sigma_n^2}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{1}{\sigma_n^2}}\right)N(x; 0, \sigma_0^2+\sigma_n^2)}{\max_s N\left(s; \frac{\sigma_1^2 x}{\sigma_1^2+\sigma_n^2}, \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_n^2}}\right)N(x; 0, \sigma_1^2+\sigma_n^2)},\\
&= \log\frac{N\left(\mu_1; \mu_1, \frac{\sigma_0^2\sigma_n^2}{\sigma_n^2+\sigma_0^2}\right)N(x; 0, \sigma_0^2+\sigma_n^2)}{N\left(\mu_2; \mu_2, \frac{\sigma_1^2\sigma_n^2}{\sigma_n^2+\sigma_1^2}\right)N(x; 0, \sigma_1^2+\sigma_n^2)},\\
&= \frac{1}{2}\log\frac{\frac{\sigma_1^2\sigma_n^2}{\sigma_n^2+\sigma_1^2}}{\frac{\sigma_0^2\sigma_n^2}{\sigma_n^2+\sigma_0^2}} - \frac{x^2}{2(\sigma_0^2+\sigma_n^2)} + \frac{x^2}{2(\sigma_1^2+\sigma_n^2)} + \frac{1}{2}\log\frac{\sigma_1^2+\sigma_n^2}{\sigma_0^2+\sigma_n^2},\\
&= \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_0^2} - \frac{x^2}{2}\frac{\sigma_1^2-\sigma_0^2}{(\sigma_0^2+\sigma_n^2)(\sigma_1^2+\sigma_n^2)}.
\end{aligned}
\tag{B3}
$$

This has a different constant term than the Bayesian observer, but shows the same dependance on x. Thus, the main difference created is a different distribution of bias for the different noise strengths σ_n.

Optimal Decision Rule Point Estimate

As the dependance of d on x is the same for the Bayesian observer and the point estimate observer, we can equate their behavioral predictions by shifting the criterion of one of them by an amount equal to:

$$
\frac{1}{2}\log\frac{\sigma_1^2+\sigma_n^2}{\sigma_0^2+\sigma_n^2} - \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_0^2},
\tag{B4}
$$

As this makes the model equivalent to the Bayes-optimal observer model, this is the optimal setting of the criterion and the optimal decision rule point estimate observer behaves the same way as the Bayesian observer.

Variational Inference

This task is the only one where we can find a sensible factorized approximation to the true posterior using variational inference. To do so, we search for an approximate distribution $q(C, s) = q(C)q(s) \approx p(C, s|x)$ which minimizes the KL divergence

$$
\text{KL}(p||q) = \sum_C \int p(C, s|x)\log\frac{p(C, s|x)}{q(C)q(s)}ds,
\tag{B5}
$$

To solve this, we can use a general inference scheme for variational inference to iterate between updating $q(C)$ and $q(s)$. If all other factors are fixed, the best solution for the logarithm of factor $q(y)$ is the expectation of the logarithm of the true posterior distribution, based on the product of all other factors $q^{\setminus y}$ (Bishop, 2006):

$$
\log(q(y)) = \langle \log p\rangle_{q^{\setminus y}} + \text{Const}.
\tag{B6}
$$

As an update for our factor $q(s)$, this results in:
(See above)

Now define $\sigma_{\text{new}}^2 = \frac{\sigma_0^2\sigma_1^2}{q(C=0)\sigma_1^2\sigma_n^2 + q(C=1)\sigma_0^2\sigma_n^2 + \sigma_0^2\sigma_1^2}$

$$
\begin{aligned}
\log(q(s)) &= -\frac{s^2 - 2\sigma_{\text{new}}^2 xs + \sigma_{\text{new}}^2 x^2}{2\sigma_{\text{new}}^2\sigma_n^2} + \text{Const}_0,\\
&= -\frac{(s - \sigma_{\text{new}}^2 x)^2}{2\sigma_{\text{new}}^2\sigma_n^2} + \text{Const}_1(x).
\end{aligned}
\tag{B8}
$$

Thus $q(s)$ is normally distributed:

$$
q(s) = N(s; \sigma_{\text{new}}^2 x, \sigma_{\text{new}}^2\sigma_n^2),
\tag{B9}
$$

For $q(C)$, we can use the fact that $q(s)$ is a normal distribution to derive the following update equation based on the mean μ_s and standard deviation σ_s of the current $q(s)$:

$$
\begin{aligned}
\log(q(C=1)) &= \langle \log p(x, C=1, s)\rangle_{q(s)} + \text{Const}_0,\\
&= \int N(s; \mu_s, \sigma_s^2)\log[N(s; x, \sigma_n^2)N(s; 0, \sigma_0^2)]ds + \text{Const}_0,\\
&= \int N(s; \mu_s, \sigma_s^2)\log[N(s; x, \sigma_n^2)]ds\\
&\quad + \int N(s; \mu_s, \sigma_s^2)\log[N(s; 0, \sigma_0^2)]ds + \text{Const}_0.
\end{aligned}
\tag{B10}
$$

The first integral does not depend on C, that is, it will cancel once we get $q(C)$. Thus:

$$
\log(q(C=1)) = \int N(s; \mu_s, \sigma_s^2)\log[N(s; 0, \sigma_0^2)]ds + \text{Const}_1,
\tag{B11}
$$

Using a formula for the cross-entropy of Gaussians:

$$
\log(q(C=1)) = -\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{\sigma_s^2+\mu_s^2}{2\sigma_0^2} + \text{Const}_1,
\tag{B12}
$$

(*Appendices continue*)

The derivation for $q(C = 2)$ is analogous. Thus, we get the following formula for $q(C = 1)$:

$$q(C=1) = \frac{\exp\left[-\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{\sigma_s^2 + \mu_s^2}{2\sigma_0^2}\right]}{\exp\left[-\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{\sigma_s^2 + \mu_s^2}{2\sigma_0^2}\right] + \exp\left[-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{\sigma_s^2 + \mu_s^2}{2\sigma_1^2}\right]},$$
$$= \frac{\frac{1}{\sigma_0}\exp\left(-\frac{\sigma_s^2 + \mu_s^2}{2\sigma_0^2}\right)}{\frac{1}{\sigma_0}\exp\left(-\frac{\sigma_s^2 + \mu_s^2}{2\sigma_0^2}\right) + \frac{1}{\sigma_1}\exp\left(-\frac{\sigma_s^2 + \mu_s^2}{2\sigma_1^2}\right)}, \tag{B13}$$

To evaluate this formula, we compute $\frac{1}{\sigma_0}\exp\left[-\frac{\sigma_s^2 + \mu_s^2}{2\sigma_0^2}\right]$ and $\frac{1}{\sigma_1}\exp\left[-\frac{\sigma_s^2 + \mu_s^2}{2\sigma_1^2}\right]$ and normalize to get $q(C = 1)$ and $q(C = 2)$. We iterate this scheme for up to 50 iterations or until the change in $q(C)$ becomes smaller than 10^{-5}. After convergence, we compute the decision variable from $q(C)$ as:

$$d_V = \log\frac{q(C = 1)}{q(C = 2)} \tag{B14}$$

Importance Sampling

As described in the main text this observer approximates the integrals in the equations for the Bayesian observer by sampling from the prior. For this particular situation, the formula for the decision variable based on N_s samples $s_{1,\ i}$ and $s_{0,\ i}$ from the priors under the two categories $p(s|C = 1)$ and $p(s|C = 0)$ respectively becomes:

$$d_s = \log\frac{\frac{1}{N_s}\sum_{i=1}^{N_s} p(x|s_{1,i})}{\frac{1}{N_s}\sum_{i=1}^{N_s} p(x|s_{2,i})},$$
$$= \log\sum_{i=1}^{N_s}\exp\left(\frac{-(x - s_{1,i})^2}{2\sigma_n}\right) - \log\sum_{i=1}^{N_s}\exp\left(\frac{-(x - s_{2,i})^2}{2\sigma_n}\right). \tag{B15}$$

Here, the normalization constant of the Gaussian cancels in the ratio, as the noise distribution has the same variance in both categories.

Joint Posterior Sampling

This observer type performs Metropolis–Hastings sampling based on proposal samples from the prior over C and s. To add a new sample C_{i+1}, s_{i+1}, we thus draw a proposal sample $\tilde{C}, \tilde{s}$ and accept it with the following probability, otherwise setting $C_{i+1} = C_i$, $s_{i+1} = s_i$:

$$p(\text{accept}) = \min\left(1, \frac{p(\tilde{s}, \tilde{C}|x)p(s_i, C_i)}{p(s_i, C_i|x)p(\tilde{s}, \tilde{C})}\right),$$
$$= \min\left(1, \frac{p(x|\tilde{s})p(\tilde{s}, \tilde{C})p(s_i, C_i)}{p(x|s_i)p(s_i, C_i)p(\tilde{s}, \tilde{C})}\right),$$
$$= \min\left(1, \frac{p(x|\tilde{s})}{p(x|s_i)}\right) \tag{B16}$$

This derivation holds for all our experiments. For this particular experiment, the likelihood ratio $\frac{p(x|\tilde{s})}{p(x|s_i)}$ becomes:

$$\frac{p(x|\tilde{s})}{p(x|s_i)} = \exp\left(-\frac{(x - \tilde{s})^2}{2\sigma_n} + \frac{(x - s_i)^2}{2\sigma_n}\right), \tag{B17}$$

Collinearity Judgements

The task was to detect whether two lines are colinear or not. Stimuli were presented at different peripheral locations, which in our models changes the standard deviation of the observations. For convenience, we here set $C = 1$ to mean separate s, as this is the more flexible category, and express s relative to the nominal eccentricity set to the mean of the s distribution(s).

$$P(s|C = 0) = \mathbb{1}_{s=s_1=s_2} N(s; 0, \sigma_0^2),$$
$$P(s|C = 1) = N(s_1; 0, \sigma_0^2)N(s_1; 0, \sigma_0^2),$$
$$P(x_i|s_i) = N(x_i; s_i, \sigma_n^2). \tag{B18}$$

(*Appendices continue*)

$$
\begin{aligned}
P(x)P(C=1|x) &= \int P(x_1|s_1)P(x_2|s_2)P(s_1,s_2|C=1)ds_1ds_2,\\
&= \int N(s_1;x_1,\sigma_n^2)N(s_1;0,\sigma_0^2)ds_1 \int N(s_2;x_2,\sigma_n^2)N(s_2;0,\sigma_0^2)ds_2,\\
&= \frac{1}{(\sigma_n\sigma_0 2\pi)^2}\int \exp -\frac{1}{2}\left[\frac{(s_1-x_1)^2}{\sigma_n^2}+\frac{s_1^2}{\sigma_0^2}\right]ds_1 \int \exp -\frac{1}{2}\left[\frac{(s_2-x_2)^2}{\sigma_n^2}+\frac{s_2^2}{\sigma_0^2}\right]ds_2,\\
&= \frac{1}{2\pi(\sigma_n^2+\sigma_0^2)}\exp -\left[\frac{1}{2(\sigma_0^2+\sigma_n^2)}(x_1^2+x_2^2)\right],\\
P(x)P(C=0|x) &= \int P(x_1|s)P(x_2|s)P(s|C=0)ds,\\
&= \int N(x_1;s,\sigma_n^2)N(x_2;s,\sigma_n^2)N(s;0,\sigma_0^2)ds,\\
&= \frac{1}{(2\pi)^{3/2}\sigma_n^2\sigma_0}\int \exp\left(-\frac{1}{2}\left[\frac{(x_1-s)^2}{\sigma_n^2}+\frac{(x_2-s)^2}{\sigma_n^2}+\frac{s^2}{\sigma_0^2}\right]\right)ds,\\
&= \frac{1}{(2\pi)^{3/2}\sigma_n^2\sigma_0}\int \exp\left(-\frac{1}{2}\frac{s^2-2\frac{\sigma_0^2}{2\sigma_0^2+\sigma_n^2}(x_1+x_2)s+\frac{\sigma_0^2}{2\sigma_0^2+\sigma_n^2}(x_1^2+x_2^2)}{\frac{\sigma_0^2\sigma_n^2}{2\sigma_0^2+\sigma_n^2}}\right)ds,\\
&= \frac{1}{(2\pi)\sigma_n\sqrt{2\sigma_0^2+\sigma_n^2}}\exp\left(\frac{1}{2\sigma_n^2}\frac{\sigma_0^2}{2\sigma_0^2+\sigma_n^2}(x_1+x_2)^2-\frac{1}{2\sigma_n^2}(x_1^2+x_2^2)\right),
\end{aligned}
\tag{B19}
$$

Bayesian Model

The Bayesian model was already derived in the original article. The posterior probability for the two categories are as follows:

(See above)

From the log-ratio of these two probabilities, we get the decision variable d_B:

$$
\begin{aligned}
d_B &= \log\frac{P(C=1|x)}{P(C=0|x)}\\
&= \log\frac{\sigma_n\sqrt{2\sigma_0^2+\sigma_n^2}}{\sigma_n^2+\sigma_0^2}-\frac{1}{2(\sigma_0^2+\sigma_n^2)}(x_1^2+x_2^2)\\
&\quad -\frac{1}{2\sigma_n^2}\frac{\sigma_0^2}{2\sigma_0^2+\sigma_n^2}(x_1+x_2)^2+\frac{1}{2\sigma_n^2}(x_1^2+x_2^2).
\end{aligned}
\tag{B20}
$$

Point Estimate Model

For the point estimate observer, we first have to find the maximum a posteriori estimates for $\hat{s}$. From standard formulas for the maximum of the product of two normal distributions, we get the following estimates for s under $C=1$:

$$
\hat{s}_1=\frac{\sigma_0^2}{\sigma_0^2+\sigma_n^2}x_1 \qquad \hat{s}_2=\frac{\sigma_0^2}{\sigma_0^2+\sigma_n^2}x_2 \tag{B21}
$$

Under $C=0$, s_1 and s_2 have to be equal. Again using standard formulas for the mean of the product of Gaussians, we can calculate an estimate for s as:

$$
\hat{s}=\hat{s}_1=\hat{s}_2=\frac{\sigma_0^2}{2\sigma_0^2+\sigma_n^2}(x_1+x_2). \tag{B22}
$$

Based on these expressions, we can now compute d_P:

$$
d_P=\log\frac{N(\hat{s}_1;0,\sigma_0)N(\hat{s}_2;0,\sigma_0)N(\hat{s}_1;x_1,\sigma_n)N(\hat{s}_2;x_2,\sigma_n)}{N(\hat{s};0,\sigma_0)N(\hat{s};x_1,\sigma_n)N(\hat{s};x_2,\sigma_n)}. \tag{B23}
$$

As usual, we compute the optimal criterion based on optimization over a sample.

Variational Observer

In this experiment, the distribution over s under Category 0 is a one dimensional subset of the two dimensional distribution under Category 1. One consequence of that is that the two cannot be expressed as densities relative to the same base measure. Informally speaking, the distribution under Category 0 would have to be infinitely high to compensate for the infinitely small area. it covers in the original space. Thus, the KL divergences the variational observer is meant to optimize are not defined and there are no sensible factorized approximations to the posterior.

Importance Sampling

As described in the main text, this observer approximates the integrals in the equations for the Bayesian observer by sampling from the prior. For this particular situation, the formula for the decision variable based on N_s samples $s_{c0,\ i}=(s_{c0,\ i,\ 1}, s_{c0,\ i,\ 2})$ and $s_{c1,\ i}=(s_{c1,\ i,\ 1}, s_{c1,\ i,\ 2})$ from the priors under the two categories $p(s|C=0)$ and $p(s|C=1)$ respectively becomes:

$$
\begin{aligned}
d_s &= \log\sum_{i=1}^{N_s}\exp\left(\frac{-(x_1-s_{c1,i,1})^2-(x_2-s_{c1,i,2})^2}{2\sigma_n}\right)\\
&\quad -\log\sum_{i=1}^{N_s}\exp\left(\frac{-(x_1-s_{c0,i,1})^2-(x_2-s_{c0,i,2})^2}{2\sigma_n}\right)
\end{aligned}
\tag{B24}
$$

(*Appendices continue*)

$$
\begin{aligned}
p(x)p(C=1|x) &= p(C=1)\int p(s_L|C)p(s_{\backslash L}p(L)p(x|L, s_D, s_T)ds_T ds_D dL,\\
&= \frac{1}{2N}\sum_{L=1}^{N}\int p(s_L|C)p(s_{\backslash L})p(x|L, s_{\backslash L}, s_L)ds_L ds_{\backslash L},\\
&= \frac{1}{N}\sum_{L=1}^{N}\int \mathbb{1}_{s_T>0}N(s_T; 0, \sigma^2)N(x_L; s_T, \sigma_N^2)ds_T \int N(s_D; 0, \sigma^2)\prod_{i\neq L}N(x_i; s_D, \sigma_N^2)ds_D,\\
&= \frac{1}{N}\sum_{L=1}^{N}\int \mathbb{1}_{s_T>0}\frac{1}{Z_T}N\left(s_T; \frac{\sigma^2 x_L}{\sigma^2+\sigma_N^2}; \frac{\sigma^2\sigma_N^2}{\sigma^2+\sigma_N^2}\right)ds_T \int \frac{1}{Z_D}N\left(s_D; \frac{\sigma^2\sum_{i\neq L}x_i}{3\sigma^2+\sigma_N^2}; \frac{\sigma^2\sigma_N^2}{3\sigma^2+\sigma_N^2}\right)ds_D,\\
&= \frac{1}{N}\sum_{L=1}^{N}\frac{1}{Z_D}\frac{1}{Z_T}\int \mathbb{1}_{s_T>0}N\left(s_T; \frac{\sigma^2 x_L}{\sigma^2+\sigma_N^2}; \frac{\sigma^2\sigma_N^2}{\sigma^2+\sigma_N^2}\right)ds_T,\\
&= \frac{1}{N}\sum_{L=1}^{N}\frac{1}{Z_D}\frac{1}{Z_T}\Phi\left(\frac{\sigma x_L}{\sqrt{(\sigma^2+\sigma_N^2)\sigma_N^2}}\right), \qquad \text{(B27)}
\end{aligned}
$$

Again, the normalization constant of the Gaussian cancels in the ratio, as the noise distribution has the same variance in both categories.

Joint Posterior Sampling

The acceptance probability for the Metropolis–Hastings sampler depends only on the likelihood ratio as we showed for the first task above. For this experiment, the likelihood ratio for accepting a new sample $\tilde{s} = (\tilde{s}_1, \tilde{s}_2)$ compared to a current sample $s = (s_1, s_2)$ becomes:

$$
\begin{aligned}
\frac{p(x|\tilde{s})}{p(x|s)} = \exp\bigg(&\frac{1}{2\sigma_n}(-(x_1-\tilde{s}_1)^2 - (x_2-\tilde{s}_2)^2 + (x_1-s_1)^2\\
&+ (x_2-s_2)^2)\bigg), \qquad \text{(B25)}
\end{aligned}
$$

Here, C does not directly enter the acceptance probability, but changes the distribution we sample $\tilde{s}$ from.

Shen and Ma

Participants were presented with four Gabor targets s_1, s_2, s_3, s_4. Three of these shared the same orientation. The participants were asked to report whether the fourth target at location L was tilted left or right.

The generative model for this experiment is as follows:

$$
\begin{aligned}
P(x, s, C, L) &= P(x|s)P(s|C, L)P(C)P(L),\\
P(x_i|s_i) &= N(x; s, \sigma_n^2),\\
P(s_{\backslash L}) &= N(s_{\backslash L}, 0, \sigma_0^2),\\
P(s_L|C{=}\,0) &= \mathbb{1}_{s_L<0}N(s; 0, \sigma_0^2),\\
P(s_L|C{=}\,1) &= \mathbb{1}_{s_L>0}N(s; 0, \sigma_0^2),\\
P(L) &= \frac{1}{4},\\
P(C) &= \frac{1}{2}. \qquad \text{(B26)}
\end{aligned}
$$

with a single noise variance σ_n^2 and variance of the prior σ_n^2. The notation $s_{\backslash L}$ indicates the common orientation of the distractors, and s_L denotes the orientation of the target at location L.

Bayesian Model

The Bayesian model was already described in the original publication. The derivation is as follows:

(See above)

which is consistent with the article with

$$
\begin{aligned}
Z_T &= \frac{N\left(s_T; \frac{\sigma^2 x_L}{\sigma^2+\sigma_N^2}; \frac{\sigma^2\sigma_N^2}{\sigma^2+\sigma_N^2}\right)}{N(s_T; 0, \sigma^2)N(s_T; x_L, \sigma_N^2)},\\
&= \frac{\sqrt{2\pi\sigma^2\sigma_N^2}}{\sqrt{\frac{\sigma^2\sigma_N^2}{\sigma^2+\sigma_N^2}}}\exp\left[-\frac{1}{2}\frac{\left(-\frac{\sigma^2 x_L}{\sigma^2+\sigma_N^2}\right)^2}{\frac{\sigma^2\sigma_N^2}{\sigma^2+\sigma_N^2}}\right]\exp\left[\frac{1}{2}\frac{(-x_L)^2}{\sigma_N^2}\right],\\
&= \sqrt{2\pi(\sigma^2+\sigma_N^2)}\exp\left[\frac{1}{2}\frac{(x_L)^2}{(\sigma^2+\sigma_N^2)}\right]. \qquad \text{(B28)}
\end{aligned}
$$

and analogous for Z_D:

$$
\begin{aligned}
Z_D &= \frac{N\left(s_D; \frac{\sigma^2\sum_{i\neq L}x_L}{3\sigma^2+\sigma_N^2}; \frac{\sigma^2\sigma_N^2}{3\sigma^2+\sigma_N^2}\right)}{N(s_D; 0, \sigma^2)\prod_{i\neq L}N(s_D; x_i, \sigma_N^2)},\\
&= \frac{\sqrt{(2\pi\sigma_N^2)^3\sigma^2}}{\sqrt{\frac{\sigma^2\sigma_N^2}{3\sigma^2+\sigma_N^2}}}\exp\left(-\frac{1}{2}\frac{\left(-\frac{\sigma^2\sum_{i\neq L}x_i}{3\sigma^2+\sigma_N^2}\right)^2}{\frac{\sigma^2\sigma_N^2}{3\sigma^2+\sigma_N^2}}\right)\exp\left(\frac{1}{2}\frac{\sum_{i\neq L}(-x_i)^2}{\sigma_N^2}\right),\\
&= \sqrt{(2\pi)^3\sigma_N^2\sigma_N^2(3\sigma^2+\sigma_N^2)}\exp\left(\frac{\left(\frac{1}{3}\sum_{i\neq L}x_i\right)^2}{2(\sigma^2+\frac{\sigma_N^2}{3})} + \frac{\sum_{i\neq L}(x_i-\bar{x}_{\backslash L})^2}{2\sigma_N^2}\right), \qquad \text{(B29)}
\end{aligned}
$$

(*Appendices continue*)

$$\begin{aligned} p(C=1|x) &= \frac{1}{p(x)}\sum_{L=1}^{N}\exp\left(-\frac{1}{2}\frac{\sigma^2(\sum_{i\neq L}x_i)^2}{(3\sigma^2+\sigma_N^2)\sigma_N^2}+\frac{1}{2}\frac{\sum_{i\neq L}(x_i)^2}{\sigma_N^2}-\frac{1}{2}\frac{(x_L)^2}{(\sigma^2+\sigma_N^2)}\right)\Phi\left(\frac{\sigma x_L}{\sqrt{(\sigma^2+\sigma_N^2)\sigma_N^2}}\right), \\ &= \frac{1}{p(x)}\sum_{L=1}^{N}\exp\left(-\frac{(\frac{1}{3}\sum_{i\neq L}x_i)^2}{2(\sigma^2+\frac{\sigma_N^2}{3})}-\frac{\sum_{i\neq L}(x_i-\bar{x}_{\setminus L})^2}{2\sigma_N^2}-\frac{(x_L)^2}{2(\sigma^2+\sigma_N^2)}\right)\Phi\left(\frac{\sigma x_L}{\sqrt{(\sigma^2+\sigma_N^2)\sigma_N^2}}\right), \end{aligned} \quad \text{(B30)}$$

Dropping all parts which do not depend on x we obtain:

(See above)

which is equivalent to the original article. In the original article, the authors the authors modeled a Bayesian observer who reports the category for which $P(C|x)$ was larger. Here, we use these values to calculate a posterior ratio and map it through the same decision noise and lapse rate machinery as all other observer models.

Point Estimate Model

In this case, we cannot easily estimate s independent of L, because L determines which orientations have to be the same. We can, however, estimate the stimulus values for each category and base our decision on the likelihood of this interpretation: We thus evaluate each combination of C and L as a category:

$$\begin{aligned} P(s_T, s_D|L, C, x) &\propto p(s_T|C)p(s_D)p(x|L, s_D, s_T), \\ &= p(s_T|C)p(s_D)\prod_i p(x_i|L, s_D, s_T), \end{aligned} \quad \text{(B31)}$$

As all distributions here are normal, we can calculate the point estimates of s analytically:

$$\hat{s}_T = \begin{cases} 0 & \text{if } C \neq \text{sign}\, x_L \\ \frac{\sigma^2}{\sigma^2+\sigma_N^2}x_L & \text{if } C = \text{sign}\, x_L \end{cases} \quad \text{(B32)}$$

$$\hat{s}_D = \frac{3\sigma^2}{\sigma_N^2+3\sigma^2}\left(\frac{1}{3}\sum_{i\neq L}x_i\right), \quad \text{(B33)}$$

With this point estimate, we can calculate the maximum posterior density for each L, C combination as:

$$\begin{aligned} \hat{p}(L,C) &= p(\hat{s}_D)p(\hat{s}_T)p(x_L|\hat{s}_T)\prod_{i\neq L}p(x_i|\hat{s}_D), \\ &= N(\hat{s}_D|0,\sigma)N(\hat{s}_T|0,\sigma)N(x_L|\hat{s}_T,\sigma_N)\prod_{i\neq L}N(x_i|\hat{s}_D,\sigma_N), \\ &= \frac{1}{\sqrt{2\pi}^6\sigma^2\sigma_N^4}\exp\left(-\frac{\hat{s}_D^2}{2\sigma^2}-\frac{\hat{s}_T^2}{2\sigma^2}-\frac{(x_L-\hat{s}_T)^2}{2\sigma_N^2}-\sum_{i\neq L}\frac{(x_i-\hat{s}_D)^2}{2\sigma_N^2}\right). \end{aligned} \quad \text{(B34)}$$

Taking the logarithm and dropping parts that do not depend on x:

$$\log \hat{p}(L, C) = -\frac{1}{2}\left(\frac{\hat{s}_D^2}{\sigma^2}+\frac{\hat{s}_T^2}{\sigma^2}+\frac{(x_L-\hat{s}_T)^2}{\sigma_N^2}+\sum_{i\neq L}\frac{(x_i-\hat{s}_D)^2}{\sigma_N^2}\right), \quad \text{(B35)}$$

Inserting the formulas for $\hat{s}_D$ and $\hat{s}_T$ for $C = signx_L$:

$$\begin{aligned} \log \hat{p}(L, C) = &-\frac{1}{2}\left(\frac{9\sigma^2}{(\sigma_N^2+3\sigma^2)^2}\left(\frac{1}{3}\sum_{i\neq L}x_i\right)^2+\frac{\sigma^2}{(\sigma_N^2+\sigma^2)^2}x_L^2\right. \\ &\left.+\frac{\sigma_N^2}{(\sigma_N^2+\sigma^2)^2}x_L^2\right), \\ &-\frac{1}{2}\sum_{i\neq L}\frac{\left(x_i-\frac{3\sigma^2}{\sigma_N^2+3\sigma^2}\left(\frac{1}{3}\sum_{j\neq L}x_j\right)\right)^2}{\sigma_N^2}, \end{aligned} \quad \text{(B36)}$$

As a result we get eight numbers for the four locations times two directions. To combine these, the most compatible version for the inference scheme is to take the maximum for each C effectively treating L as another nuisance parameter to maximize over.

Using the maximum over locations, we get the following equation for d_P:

$$\begin{aligned} d_P &= \max_{s,L}\log[p(x|s, L, C=1)p(s|L, C=1)] \\ &\quad -\max_{s,L}\log[p(x|s, C=0, L)p(s|L, C=0)], \\ &= \max_L \log \hat{p}(L, C=1) - \max_L \log \hat{p}(L, C=0). \end{aligned} \quad \text{(B37)}$$

Optimal-Criterion Point Estimate Observer

This case is symmetric in the two categories and has equal complexity of the compared alternatives. Thus, in this case, the optimal bias is 0!

Importance Sampling

For importance sampling, we again sample from the priors to estimate the integrals required for solving the Bayesian observer. In this case, this implies first sampling a location for the outlier and then

(*Appendices continue*)

sampling the four stimulus orientations based on the location and the category. The formula for the decision variable based on N_s samples $s_{c0,\,i} = (s_{c0,\,i,\,1}, s_{c0,\,i,\,2}, s_{c0,\,i,\,3}, s_{c0,\,i,\,4})$ and $s_{c1,\,i} = (s_{c1,\,i,\,1}, s_{c1,\,i,\,2}, s_{c1,\,i,\,3}, s_{c1,\,i,\,4})$ from the priors under the two categories $p(s|C=0)$ and $p(s|C=1)$ respectively becomes:

$$\begin{aligned} d_s = {} & \log \sum_{i=1}^{N_s} \exp\left(\frac{-\sum_{j=1}^{4}(x_j - s_{c1,i,j})^2}{2\sigma_n}\right), \\ & - \log \sum_{i=1}^{N_s} \exp\left(\frac{-\sum_{j=1}^{4}(x_j - s_{c0,i,j})^2}{2\sigma_n}\right). \end{aligned} \tag{B38}$$

Again, the normalization constant of the Gaussian cancels in the ratio, as the noise distribution has the same variance in both categories.

Joint Posterior Sampling

The acceptance probability for the Metropolis–Hastings sampler depends only on the likelihood ratio as we showed for the first task above. For this experiment, the likelihood ratio for accepting a new sample $\tilde{s} = (\tilde{s}_1, \tilde{s}_2, \tilde{s}_3, \tilde{s}_4)$ compared to a current sample $s = (s_1, s_2, s_3, s_4)$ becomes:

$$\frac{p(x|\tilde{s})}{p(x|s)} = \exp\left(\frac{1}{2\sigma_n}\left(-\sum_{j=1}^{4}(x_j - \tilde{s}_j)^2 + \sum_{j=1}^{4}(x_j - s_j)^2\right)\right), \tag{B39}$$

Both C and L do not directly enter the acceptance probability, but change the distribution $\tilde{s}$ we sample from.

Change Detection

The task was to detect whether or not one of four ellipses changed in orientation. Using *VM* () to denote the von Mises distribution, the generative model for this experiment is:

$$\begin{aligned} p(C) &= \frac{1}{2}, \\ p(\Delta) &= \frac{1}{2\pi}, \\ p(L) &= \frac{1}{N}, \\ p(\mathrm{D}|\Delta, C, L) &= \delta(\mathrm{D} - C\mathbf{1}_L(\Delta)), \\ p(\xi) &= \left(\frac{1}{2\pi}\right)^N, \\ p(\phi|\xi, \mathrm{D}) &= \delta(\phi - \xi - \mathrm{D}), \\ p(\mathbf{x}|\xi) &= \prod_{i=1}^{N} \mathrm{VM}(x_i; \xi_i, \kappa_{x,i}), \\ p(\mathbf{y}|\phi) &= \prod_{i=1}^{N} \mathrm{VM}(y_i; \phi_i, \kappa_{y,i}), \end{aligned} \tag{B40}$$

where $\Delta \in [0, 2\pi]$ is the size of the change, L is the location of the change, D is the vector of changes, ξ is the vector of true orientations in the first display, each in $[0, 2\pi]$, ϕ is the vector of true orientations in the second display and X, Y are the observations of the orientations in the first and second display respectively.

Bayesian Model

The Bayesian model was also derived in the original article.

$$d_B = \frac{p(C=1|\mathbf{x},\mathbf{y})}{p(C=0|\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{x},\mathbf{y}|C=1)p(C=1)}{p(\mathbf{x},\mathbf{y}|C=0)p(C=0)}. \tag{B41}$$

In order to compute this value, we must marginalize over the unknown variables ξ, ϕ, Δ, and Δ.

(See below)

Then, plugging this equation into the likelihood ratio:

$$\begin{aligned} \frac{p(x,y|C=1)p(C=1)}{p(x,y|C=0)p(C=0)} &= \frac{p(C=1)}{p(C=0)} \frac{\sum_{i=1}^{N} \iint p(\mathbf{x}|\xi)p(\mathbf{y}|\xi + C\Delta 1_i)d\xi d\Delta}{\sum_{i=1}^{N} \iint p(\mathbf{x}|\xi)p(\mathbf{y}|\xi)d\xi d\Delta}, \\ &= \frac{p(C=1)}{p(C=0)} \frac{\sum_{i=1}^{N} \iint p(\mathbf{x}|\xi)p(\mathbf{y}|\xi + C\Delta 1_i)d\xi d\Delta}{2\pi N \int p(\mathbf{x}|\xi)p(\mathbf{y}|\xi)d\xi}. \end{aligned} \tag{B43}$$

$$\begin{aligned} p(x,y|C)p(C) &= \iiiint p(\mathbf{x}|\xi)p(\xi)p(\mathbf{y}|\phi)p(\phi|\mathrm{D})p(\mathrm{D}|C,\Delta)p(\Delta)p(C)d\xi d\phi d\mathrm{D}d\Delta, \\ &= p(C)\left(\frac{1}{2\pi}\right)^{N+1} \iiiint p(\mathbf{x}|\xi)p(\mathbf{y}|\phi)\delta(\phi - (\xi + C\mathrm{D})), \\ & \quad \left(\frac{1}{N}\sum_{i=1}^{N}\delta(\mathrm{D}_i - C1_i)\right)d\xi d\phi d\mathrm{D}d\Delta, \\ &= p(C)\left(\frac{1}{2\pi}\right)^{N+1} \frac{1}{N}\sum_{i=1}^{N} \iint p(\mathbf{x}|\xi)p(\mathbf{y}|\xi + C\Delta 1_i)d\xi d\Delta. \end{aligned} \tag{B42}$$

(*Appendices continue*)

$$
\begin{aligned}
d_B &= \frac{p(C=1)}{p(C=0)} \frac{\sum_{i=1}^{N} \int (\prod_{j\neq i} \int p(x_j|\xi_j)p(y_j|\xi_j)d\xi_j)(\int p(x_i|\xi_i)p(y_i|\xi_i+\Delta)d\xi_i)d\Delta}{2\pi N \prod_{i=1}^{N} \int p(x_i|\xi_i)p(y_i|\xi_i)d\xi_i},\\
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{\int (\prod_{j\neq i} \int p(x_j|\xi_j)p(y_j|\xi_j)d\xi_j)(\int p(x_i|\xi_i)p(y_i|\xi_i+\Delta)d\xi_i)d\Delta}{2\pi N(\prod_{j\neq i} \int p(x_j|\xi_j)p(y_j|\xi_j)d\xi_j)(\int p(x_i|\xi_i)p(y_i|\xi_i)d\xi_i)},\\
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{\iint p(x_i|\xi_i)p(y_i|\xi_i+\Delta)d\xi_i d\Delta}{2\pi N \int p(x_i|\xi_i)p(y_i|\xi_i)d\xi_i},\\
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{\iint \mathrm{VM}(x_i;\xi_i,\kappa_{x,i})\mathrm{VM}(y_i;\xi_i+\Delta,\kappa_{y,i})d\xi_i d\Delta}{2\pi N \int \mathrm{VM}(x_i;\xi_i,\kappa_i)\mathrm{VM}(y_i;\xi_i,\kappa_i)d\xi_i},\\
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{1}{2\pi N \int \mathrm{VM}(x_i;\xi_i,\kappa_{x,i})\mathrm{VM}(y_i;\xi_i,\kappa_{y,i})d\xi_i},\\
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{1}{2\pi N \int \frac{I_0(\kappa)}{2\pi I_0(\kappa_{x,i})I_0(\kappa_{y,i})} \mathrm{VM}(\xi_i;\mu,\kappa)d\xi_i}, \qquad \text{(B44)}
\end{aligned}
$$

Because the N items are conditionally independent, we can break the expression up into a product of each item and further simplify.

(See above)

where $\mu = x_i + \arctan(\sin(y_i - x_i), (\kappa_{x,i}/\kappa_{y_i}) + \cos(y_i - x_i))$ and $\kappa = \sqrt{\kappa_{x,i}^2 + \kappa_{y,i}^2 + 2\kappa_{x,i}\kappa_{y,i}\cos(x_i - y_i)}$.

$$
\begin{aligned}
&= \frac{p(C=1)}{p(C=0)} \sum_{i=1}^{N} \frac{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}{N I_0(\kappa) \int \mathrm{VM}(\xi_i;\mu,\kappa)d\xi_i},\\
&= \frac{p(C=1)}{p(C=0)} \frac{1}{N} \sum_{i=1}^{N} \frac{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}{I_0(\kappa)}, \qquad \text{(B45)}
\end{aligned}
$$

This produces the final expression of the decision variable,

$$
d_B = \frac{p(C=1)}{p(C=0)} \frac{1}{N} \sum_{i=1}^{N} d_i. \qquad \text{(B46)}
$$

where

$$
d_i = \frac{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}{I_0(\sqrt{\kappa_{x,i}^2 + \kappa_{y,i}^2 + 2\kappa_{x,i}\kappa_{y,i}\cos(x_i - y_i)})}. \qquad \text{(B47)}
$$

Point Estimate Model

$$
d_P = \log \frac{\max_{\xi, L, \Delta}[p(\mathbf{x},\mathbf{y}|\xi, L, \Delta, C=1)p(\xi)p(L)p(\Delta)]}{\max_{\xi}[p(\mathbf{x},\mathbf{y}|\xi, C=0)p(\xi)p(L)p(\Delta)]}, \qquad \text{(B48)}
$$

Note that we do not need to additionally maximize over variables D and ϕ, because they are deterministically related to the variables we are maximizing, namely ξ, Δ, and L. Expanding the denominator first,

(See below)

where $\mu_{\xi,i} = x_i + \arctan(\sin(y_i - x_i), (\kappa_{x,i}/\kappa_{y,i}) + \cos(y_i - x_i))$ and $\kappa_{\xi,i} = \sqrt{\kappa_{x,i}^2 + \kappa_{y,i}^2 + 2\kappa_{x,i}\kappa_{y,i}\cos(y_i - x_i)}$ (Murray & Morgenstern, 2010).

We choose $\hat{\xi}_i = \mu_{\xi,i}$, to maximize:

$$
\begin{aligned}
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{N+1} \prod_{i=1}^{N} \frac{I_0(\kappa_{\xi,i})}{2\pi I_0(\kappa_{x,i})I_0(\kappa_{y,i})} \frac{e^{\kappa_{\xi,i}}}{2\pi I_0(\kappa_{\xi,i})},\\
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{3N+1} \prod_{i=1}^{N} \frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}, \qquad \text{(B50)}
\end{aligned}
$$

$$
\begin{aligned}
\text{denominator} &= \max_{\xi}[p(\mathbf{x},\mathbf{y}|\xi, C=0)p(\xi)p(L)p(\Delta)],\\
&= \max_{\xi}[p(\mathbf{y}|\phi=\xi)p(\mathbf{x}|\xi)p(\xi)p(L)p(\Delta)],\\
&= \max_{\xi_1,\ldots,\xi_N}\left[\prod_{i=1}^{N}\mathrm{VM}(y_i;\xi_i,\kappa_{y,i})\prod_{i=1}^{N}\mathrm{VM}(x_i;\xi_i,\kappa_{x,i})\left(\frac{1}{2\pi}\right)^{N}\frac{1}{2\pi}\frac{1}{N}\right],\\
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{N+1}\prod_{i=1}^{N}\max_{\xi_i}[\mathrm{VM}(y_i;\xi_i,\kappa_{y,i})\mathrm{VM}(x_i;\xi_i,\kappa_{x,i})],\\
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{N+1}\prod_{i=1}^{N}\max_{\xi_i}\left[\frac{I_0(\kappa_{\xi,i})}{2\pi I_0(\kappa_{x,i})I_0(\kappa_{y,i})}\mathrm{VM}(\xi_i;\mu_{\xi,i},\kappa_{\xi,i})\right], \qquad \text{(B49)}
\end{aligned}
$$

(*Appendices continue*)

$$\begin{aligned}
\text{numerator} &= \max_{\xi, L, \Delta}[p(\mathbf{x}, \mathbf{y}|\xi, L, \Delta, C = 1)p(\xi)p(L)p(\Delta)]\\
&= \max_{\xi, L, \Delta}[p(\mathbf{y}|\xi, \Delta, L, C = 1)p(\mathbf{x}|\xi)p(\xi)p(\Delta)p(L)],\\
&= \max_{\xi_1, \ldots, \xi_N, L, \Delta}\left[\prod_{i=1}^{N}\mathrm{VM}(y_i; \xi_i + I_L\Delta, \kappa_{y,i})\prod_{i=1}^{N}\mathrm{VM}(x_i; \xi_i, \kappa_{x,i})\left(\frac{1}{2\pi}\right)^N\frac{1}{2\pi}\frac{1}{N}\right],\\
&\text{where } I_L \text{ is 1 if } i = L \text{ and 0 otherwise,}\\
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{N+1}\prod_{i=1}^{N}\max_{\xi_i, L, \Delta}\mathrm{VM}(y_i; \xi_i + I_L\Delta, \kappa_{y,i})\mathrm{VM}(x_i; \xi_i, \kappa_{x,i}),
\end{aligned} \tag{B51}$$

Expanding the numerator next,

(See above)

For the Lth location, this expression is maximized if we choose $\hat{\xi}_L = x_L$, and $\Delta = y_L - x_L$. At the remaining $N - 1$ locations, the stimulus did not change, and thus the same derivation applies as for items in the denominator. Further simplifying,

(See below)

Combining and simplifying the numerator and denominator:

$$\begin{aligned}
\exp(d_P) &= \frac{\max_L\left[\frac{1}{N}\left(\frac{1}{2\pi}\right)^{3N+1}\frac{e^{\kappa_{x,L}}e^{\kappa_{y,L}}}{I_0(\kappa_{x,L})I_0(\kappa_{y,L})}\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}\right]}{\frac{1}{N}\left(\frac{1}{2\pi}\right)^{3N+1}\prod_{i=1}^{N}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}},\\
&= \max_L\frac{\frac{e^{\kappa_{x,L}}e^{\kappa_{y,L}}}{I_0(\kappa_{x,L})I_0(\kappa_{y,L})}\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}}{\prod_{i=1}^{N}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}},\\
&= \max_L\frac{\frac{e^{\kappa_{x,L}}e^{\kappa_{y,L}}}{I_0(\kappa_{x,L})I_0(\kappa_{y,L})}\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}}{\frac{e^{\kappa_{\xi,L}}}{I_0(\kappa_{x,L})I_0(\kappa_{y,L})}\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}},\\
&= \max_L\frac{e^{\kappa_{x,L}}e^{\kappa_{y,L}}}{e^{\kappa_{\xi,L}}}.
\end{aligned} \tag{B53}$$

Keeping $p(L)p(\Delta)$ only for $C = 1$ would yield an additional factor $\frac{1}{2\pi N}$. We assume the model observer calculates this expression for each possible value of L, and chooses the one which provides the largest value. Writing out d_p explicitly yields:

$$\begin{aligned}
d_P &= \max_L[\kappa_{x,L} + \kappa_{y,L} - \kappa_{\xi,L}],\\
&= \max_L\left[\kappa_{x,L} + \kappa_{y,L} - \sqrt{\kappa_{x,L}^2 + \kappa_{y,L}^2 + 2\kappa_{x,L}\kappa_{y,L}\cos(y_L - x_L)}\right],
\end{aligned} \tag{B54}$$

which has a minimum of 0 at $y_L = x_L$ and a maximum $\leq\kappa_{x,L} + \kappa_{y,L}$ for $x_L - y_L = \pi$.

Optimal-Criterion Point Estimate Observer

To compute the Optimal-Criterion Point Estimate Observer, we need to find the offsets, $c(k_x, k_y)$, to maximize task performance of the decision boundary $d_H = d_P + c(k_x, k_y)$. This is difficult to do analytically, so we compute $c(k_x, k_y)$ numerically.

Variational Observer

The variational observer is not defined in this task. A variational observer would not be defined in a one-item change detection task for the same reasons laid out in the collinearity judgment task (B); it is even less clear how this observer would be defined in a four-item change detection task.

Importance Sampling

As described in the main text and earlier derivations, this observer approximates the integrals in the equations for the Bayesian observer by sampling from the prior.

$$\begin{aligned}
p(\mathbf{x}, \mathbf{y}|C) &= \int p(\mathbf{x}, \mathbf{y}|\xi, \phi)p(\xi, \phi|C)d\xi d\phi,\\
&= \int p(\mathbf{x}|\xi)p(\mathbf{y}|\phi)p(\xi, \phi|C)d\xi d\phi,\\
&= \int p(\mathbf{x}|\xi)p(\mathbf{y}|\xi + \mathbf{1}_C\mathrm{D})p(\xi)p(\mathrm{D})d\xi d\mathrm{D},\\
&= \frac{1}{p(\mathbf{x})}\int p(\xi|\mathbf{x})p(\mathbf{y}|\xi + \mathbf{1}_C\mathrm{D})p(\mathrm{D})d\xi d\mathrm{D},\\
&\approx \frac{1}{N_s p(\mathbf{x})}\sum_{i=1}^{N_s}p(\mathbf{y}|\xi_i + \mathbf{1}_C\mathrm{D}_i),
\end{aligned} \tag{B55}$$

where we used Bayes rule $p(\xi|\mathbf{x}) = \frac{p(\mathbf{x}|\xi)p(\xi)}{p(\mathbf{x})}$ for the third step.

Now, we take the ratio of the two hypotheses:

$$\begin{aligned}
d_s &= \frac{p(\mathbf{x}, \mathbf{y}|C = 1)}{p(\mathbf{x}, \mathbf{y}|C = 0)},\\
&= \frac{\sum_{i=1}^{N_s}p(\mathbf{y}|\xi_i + \mathrm{D}_i)}{\sum_{i=1}^{N_s}p(\mathbf{y}|\xi_i)}.
\end{aligned} \tag{B56}$$

$$\begin{aligned}
&= \max_L\left[\frac{1}{N}\left(\frac{1}{2\pi}\right)^{N+1}\mathrm{VM}(y_L; y_L, \kappa_{y,L})\mathrm{VM}(x_L; x_L, \kappa_{x,L})\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{4\pi^2 I_0(\kappa_{x,i})I_0(\kappa_{y,i})}\right],\\
&= \frac{1}{N}\left(\frac{1}{2\pi}\right)^{3N+1}\max_L\left[\frac{e^{\kappa_{x,L}}e^{\kappa_{y,L}}}{I_0(\kappa_{x,L})I_0(\kappa_{y,L})}\prod_{i\neq L}\frac{e^{\kappa_{\xi,i}}}{I_0(\kappa_{x,i})I_0(\kappa_{y,i})}\right].
\end{aligned} \tag{B52}$$

(*Appendices continue*)

Joint Posterior Sampling

This observer uses the same Metropolis–Hastings sampling method explained in text in earlier derivations, accepting proposal samples from the prior over C, ξ, and ϕ with the following probability:

$$\min\left(1, \frac{p(x|\tilde{\xi})p(y|\tilde{\phi})}{p(x|\xi)p(y|\phi)}\right), \tag{B57}$$

We calculate $p(x|\xi)p(y|\phi)$:

$$\begin{aligned} p(x|\xi)p(y|\phi) &= \prod_{j=1}^{N} \mathrm{VM}(x_j; \xi_j, \kappa_j)\mathrm{VM}(y_j; \phi_j, \kappa_{y,j}), \\ &= \prod_{j=1}^{N} \mathrm{VM}(x_j; \xi_j, \kappa_j)\mathrm{VM}(y_j; \xi_j + 1_j\Delta, \kappa_{y,j}), \\ &\propto \prod_{j=1}^{N} \mathrm{VM}(\xi_j; x_j, \kappa_j)\mathrm{VM}(y_j; \xi_j + 1_j\Delta, \kappa_{y,j}), \end{aligned} \tag{B58}$$

where $1_j = 1$ in location of change and 0 otherwise. We can directly sample from these distributions, to calculate the acceptance probability.

Visual Search

In this task, participants are presented with N items, each possessing an orientation s_i. The task is to determine whether one of the items has been drawn from the "target" distribution (this case is denoted $C = 1$), with the remainder of items drawn from a "distractor" distribution, or whether the items are all drawn from the distractor distribution ($C = 0$). We use $L = i$ to denote that item location i is designated as the "target" location. If item location i is the designated target location and $C = 1$, then the item at location i is drawn from the target distribution. If the item location i is not the designated target location, or if $C = 0$, then the item at location i is drawn from the distractor distribution.

A subtlety here is that we could frame the relationship between C and L in two different ways. One option is that we could conceptualize L as being independent of C. In this case,

$$p(L = i|C) = \frac{1}{N}, \tag{B59}$$

where N is the number of items in a trial. When $C = 0$, a location is still designated as the target location, but this designation has no effect. However, we could also conceptualize L as dependent on C so that,

$$p(L = i|C) = \begin{cases} 1 & \text{if } C = 0 \text{ and } i = -1 \\ 0 & \text{if } C = 0 \text{ and } i \neq -1 \\ \frac{1}{N} & \text{if } C = 1 \text{ and} 1 \leq i \geq N \end{cases}, \tag{B60}$$

where we have used $i = -1$ to indicate that none of the possible item locations have been selected as the target location. While this distinction may appear inconsequential from a normative perspective, it affects the point estimate observer, as we will see below.

For the item at location i, the model observer receives a noisy measurement x_i of the true orientation s_i. By assumption or design,

$$p(x_i|s_i) = \mathrm{VM}(x_i; s_i, \kappa), \tag{B61}$$

$$p(s_i|L \neq i, C = 1) = \mathrm{VM}(s_i; 0, \kappa_s), \tag{B62}$$

$$p(s_i|C = 0) = \mathrm{VM}(s_i; 0, \kappa_s), \tag{B63}$$

$$p(s_i|L = i, C = 1) = \delta(s_i), \tag{B64}$$

$$p(C) = \frac{1}{2}. \tag{B65}$$

$\delta()$ is the Dirac delta function. *VM*() is the von Mises distribution, and its two parameters (following the semicolon) are the mean and concentration parameter. s_i codes the orientation in a very specific way: It represents twice the difference between the orientation of an item, and the target orientation, in radians. This ensures all Gabor orientations are between −90° and 90°, avoiding any issues with the fact that a Gabor rotated 180° is identical to the original Gabor. We assume the measurement x_i is conditionally independent of the other measurements, given s_i. This setup almost exactly matches the one used in the original article (Calder-Travis & Ma, 2020). We will use x and s without subscripts to denote vectors containing every x_i and every s_i, respectively.

An important detail (both here and in the original article) is that we do not assume that κ, which is related to the precision of observers' noisy stimulus measurements, is independent of the number of items in the display. Instead, we allow the possibility that more items in a display may lead to noisier measurements of those items. κ therefore corresponds to one of four values $\kappa_{N=2}$, $\kappa_{N=3}$, $\kappa_{N=4}$, or $\kappa_{N=6}$, depending on the number of items in a trial, N.

Bayesian Model

The derivations in this case are similar to those in the original article (Calder-Travis & Ma, 2020). Note that L is defined differently here, for consistency with the other studies presented in the present article (see description above). The decision variable used by the Bayesian observer is given by Equation 2,

$$\begin{aligned} d_B &= \log\frac{P(C = 1|x)}{P(C = 0|x)}, \\ &= \log\frac{P(x|C = 1)}{P(x|C = 0)} + \log\frac{P(C = 1)}{P(C = 0)}, \\ &= \log\frac{P(x|C = 1)}{P(x|C = 0)}. \end{aligned} \tag{B66}$$

We see d_B is just the log-likelihood ratio. In the case we consider, the log-likelihood ratio can be written in terms of "local log-likelihood ratios" (Ma et al., 2011; Palmer et al., 2000),

$$d_B = \log\left(\frac{1}{N}\sum_{i=1}^{N} e^{d_i}\right). \tag{B67}$$

(*Appendices continue*)

Each local log-likelihood ratio is given by,

$$\begin{aligned} d_i &= \log\frac{p(x_i|L=i,C=1)}{p(x_i|C=0)},\\ &= \log\frac{\int p(x_i|s_i)p(s_i|L=i,C=1)ds}{\int p(x_i|s_i)p(s_i|C=0)ds},\\ &= \log\frac{\mathrm{VM}(x_i;0,\kappa)}{\int \mathrm{VM}(x_i;s_i,\kappa)\mathrm{VM}(s_i;0,\kappa_s)ds}, \end{aligned} \tag{B68}$$

where we have used the properties of the generative model set out above.

Using an expression given by Murray and Morgenstern (2010) for the product of von Mises distributions, in Calder-Travis and Ma (2020), we noted,

$$d_i = \kappa\cos(x_i) + \log\frac{I_0(\kappa_s)}{I_0(\sqrt{\kappa^2+\kappa_s^2+2\kappa\kappa_s\cos(x_i)})}, \tag{B69}$$

where I_0 are modified Bessel functions of the first kind and order zero. We can then find the Bayesian observer's decision variable by using these local log-likelihoods in Equation B67.

Point Estimate Observer Model

The point estimate observer uses the decision variable given by Equation 3. In our case, in addition to s, we also have the location variable L that the Bayesian observer marginalizes out, and which the point estimate observer will maximize over. This gives us, for the decision variable,

$$d_P = \log\frac{P(C=1)\max_{s,L}[P(x|s)P(s|L,C=1)p(L|C=1)]}{P(C=0)\max_{s,L}[P(x|s)P(s|L,C=0)p(L|C=0)]}. \tag{B70}$$

We will see that the distinction noted above, between treating L as independent of C, versus treating L as dependent on C, has an effect here. First consider the case in which we conceptualize L as independent of C. In this conceptualization of the problem, $p(L=i|C)=1/N$ for both $C=1$ and $C=0$ and for all $L=i$. Hence, the decision variable in this case is,

$$d_P = \log\frac{\max_{s,L}[P(x|s)P(s|L,C=1)]}{\max_s[P(x|s)P(s|C=0)]}. \tag{B71}$$

We refer to this model as the "point estimate Ind. L model."

Consider now the second possibility in which we treat L as dependent on C, and equal to a constant value when $C=0$. In this case, Equation B70 becomes,

$$d_P = \log\frac{\max_{s,L}[P(x|s)P(s|L,C=1)p(L|C=1)]}{\max_s[P(x|s)P(s|C=0)]}, \tag{B72}$$

and using Equation B60,

$$d_P = \log\frac{\max_{s,L}\left[\frac{1}{N}P(x|s)P(s|L,C=1)\right]}{\max_s[P(x|s)P(s|C=0)]}. \tag{B73}$$

We refer to this model as the "point estimate Dep. L model." This is the model that we report as the "point estimate" model in the main text (in the sections for this visual search task).

We see that the point estimate observer uses a very similar decision variable, whichever formalization of the generative model they assume. Comparing Equations B71 and B73, we see that the two decision variables differ only in that one features $1/N$ in one of the maximizations. Regardless of the point estimate variant used, we need to find expressions for the following two distributions:

$$\begin{aligned} p(x,s|L,C=1) &= p(x|s)p(s|L,C=1),\\ p(x,s|C=0) &= p(x|s)p(s|C=0). \end{aligned} \tag{B74}$$

First consider $p(x,s|C=0)$. Using the assumption that the noise corrupting the measurement of s_i is independent of the noise corrupting s_j, and that the item orientations are independent of each other, conditional on none of them being the target, we can write,

$$\begin{aligned} p(x,s|C=0) &= \prod_i p(x_i|s_i)p(s_i|C=0),\\ &= \prod_i \mathrm{VM}(x_i;s_i,\kappa)\mathrm{VM}(s_i;0,\kappa_s),\\ &= \prod_i \mathrm{VM}(s_i;x_i,\kappa)\mathrm{VM}(s_i;0,\kappa_s),\\ &= \prod_i \frac{I_0(\kappa_d(x_i))}{2\pi I_0(\kappa)I_0(\kappa_s)}\mathrm{VM}(s_i;\mu_d(x_i),\kappa_d(x_i)),\\ &= \prod_i \rho(x_i)\mathrm{VM}(s_i;\mu_d(x_i),\kappa_d(x_i)). \end{aligned} \tag{B75}$$

We have again used the result in Murray and Morgenstern (2010) for the multiplication of two von Mises distributions such that,

$$\mu_d(x_i) = x_i + \arctan(-\sin(x_i), \frac{\kappa}{\kappa_s} + \cos(x_i)), \tag{B76}$$

$$\kappa_d(x_i) = \sqrt{\kappa^2+\kappa_s^2+2\kappa\kappa_s\cos(x_i)}, \tag{B77}$$

and we have used the following abbreviation,

$$\rho(x_i) = \frac{I_0(\kappa_d(x_i))}{2\pi I_0(\kappa)I_0(\kappa_s)}. \tag{B78}$$

Note the "(x_i)" in front of μ_d, κ_d, and ρ is used to indicate that these are three functions of x_i. Looking at the result in Equation B75, we see that the s_i are decoupled in the sense that they all appear in separate factors that are multiplied together. Therefore, we can find the value of s that maximizes B by maximizing each of these factors individually. Each factor in the product is maximal when,

$$s_i = \mu_d(x_i), \tag{B79}$$

and we have

$$\max_s p(x,s|C=0) = \prod_i \rho(x_i)\mathrm{VM}(0;0,\kappa_d(x_i)). \tag{B80}$$

(*Appendices continue*)

$$\begin{aligned} p(x, s|L = i, C = 1) &= p(x|s)p(s|L = i, C = 1), \\ &= p(x_i|s_i)p(s_i|L = i, C = 1)\prod_{j \neq i} p(x_j|s_j)p(s_j|L \neq j, C = 1), \\ &= \mathrm{VM}(x_i; s_i, \kappa)\delta(s_i)\prod_{j \neq i}\mathrm{VM}(x_j; s_j, \kappa)\mathrm{VM}(s_j; 0, \kappa_s), \\ &= \mathrm{VM}(x_i; s_i, \kappa)\delta(s_i)\prod_{j \neq i}\rho(x_j)\mathrm{VM}(s_j; \mu_d(x_j), \kappa_d(x_j)). \end{aligned} \tag{B81}$$

Now, consider $p(x, s|L, C = 1)$ when $L = i$,

(See above)

We see that the different s_k again appear in separate factors that are multiplied together. Hence, we can again maximize this expression over s by maximizing over each s_k individually. This gives, $s_i = 0$ where $L = i$, and $s_j = \mu_d(x_j)$ where $L \neq j$. Therefore,

$$\max_s p(x, s|L = i, C = 1) = \mathrm{VM}(x_i; 0, \kappa)\prod_{j \neq i}\rho(x_j)\mathrm{VM}(0; 0, \kappa_d(x_j)). \tag{B82}$$

This is an expression for the maximal value of $p(x, s|L = i, C = 1)$, assuming a specific target location ($L = i$). Returning to Equation B71 and Equation B73, we also have a maximization over L. Using our expressions for $p(x, s|L = i, C = 1)$ and $p(x, s|C = 0)$, we have in Equation B73, that is, for the point estimate Dep. L model,

$$\begin{aligned} d_P &= \log\frac{\frac{1}{N}\max\limits_i\left[\mathrm{VM}(x_i; 0, \kappa)\prod_{j \neq i}\rho(x_j)\mathrm{VM}(0; 0, \kappa_d(x_j))\right]}{\prod\limits_j\rho(x_j)\mathrm{VM}(0; 0, \kappa_d(x_j))}, \\ &= \log\left(\frac{1}{N}\max_i\left[\frac{\mathrm{VM}(x_i; 0, \kappa)}{\rho(x_i)\mathrm{VM}(0; 0, \kappa_d(x_i))}\right]\frac{\prod_j\rho(x_j)\mathrm{VM}(0; 0, \kappa_d(x_j))}{\prod\limits_j\rho(x_j)\mathrm{VM}(0; 0, \kappa_d(x_j))}\right), \\ &= \max_i\log(2\pi I_0(\kappa_s)e^{\kappa\cos(x_i) - \kappa_d(x_i)}) - \log N. \end{aligned} \tag{B83}$$

The result for the decision variable described by Equation B71, and used in the point estimate Ind. L model, is almost the same and only differs in not having a factor of $1/N$ in the argument of the logarithm, which only results in the removal of the log N shift.

$$d_P = \max_i\log(2\pi I_0(\kappa_s)e^{\kappa\cos(x_i) - \kappa_d(x_i)}). \tag{B84}$$

As mentioned, for the result reported in the main text, we exclusively used the Dep. L model in which the model observer treats the location variable as dependent on the category. This model is arguably more theoretically sound: In the alternative Ind. L model, where the model observer assumes that the location variable and category variable are independent, the model observer maximizes over target location when evaluating the merit of the hypothesis that no target is present. Nevertheless, we fitted both models and report the model comparison below.

Optimal-Criterion Point Estimate Observer

On each trial there will be some value of κ (which varies with the number of items in the display), and some value of κ_s (which varies depending on the distribution from which distractors are drawn; see Appendix A). For each value of κ and κ_s, we evaluated, via simulation, the optimal offset to apply to the standard point estimate observer's decision variable, as described in the section "Method."

Variational Inference

Using variational inference to find an approximate posterior distribution $q(C, s, L) = q(C)q(s)q(L)$ that factorizes over C, s, and L does not work in this case, because of the nature of the joint probability distribution over these variables. In the task considered here if $C = 1$ and $L = i$ then $s_i = 0$, and no other values for s_i are possible. For the reasons discussed in the main text (see "Theoretical analysis"), the approximate posterior must therefore assign zero probability to $q(C = 1, L = i, s_i \neq 0) = q(C = 1)q(L = i)q(s_i \neq 0)$. This is only satisfied if $q(C = 1) = 0$, $q(L = i) = 0$ or $q(s_i \neq 0) = 0$. This line of reasoning applies to all item locations i. Hence, either $q(C = 1) = 0$, or one of $q(L = i) = 0$ and $q(s_i \neq 0) = 0$ is satisfied at each location i. Both of these options lead us to an implausible approximation of the posterior distribution.

Importance Sampling

The importance sampling observer approximates the ratio used by the Bayesian observer in by computing approximate integrals as follows,

$$\begin{aligned} d &= \log\frac{p(x|C = 1)}{p(x|C = 0)}, \\ &= \log\frac{\int p(x|s)p(s|C = 1)ds}{\int p(x|s)p(s|C = 0)ds}, \\ &\approx \log\frac{\frac{1}{N_s}\sum_{i=1}^{N_s}p(x|s^{(i, C=1)})}{\frac{1}{N_s}\sum_{i=1}^{N_s}p(x|s^{(i, C=0)})}, \end{aligned} \tag{B85}$$

(*Appendices continue*)

Where N_s is the number of samples that this observer uses, and $s^{(i,\,C\,=\,\psi)}$ denotes the ith sample from the distribution $p(s|C = \psi)$.

$p(s|C = 0)$ can be computed from Equation B63, using that the stimuli at each location, s_i, are conditionally independent of the other stimuli. To find $p(s|C = 1)$, we have to additionally take the location variable, L, into account,

$$p(s|C = 1) = \sum_{l=1}^{N} p(s|L = l, C = 1)p(L = l|C = 1), \tag{B86}$$

and then, we can use Equations B60, B62 and B64. As before, N is the number of items in a trial.

Equation B85 can be rewritten in a way that is easier to evaluate. To do so, we will first find an expression for $\log p(x|s)$,

$$\begin{aligned}
\log p(x|s) &= \log \prod_{j=1}^{N} p(x_j|s_j), \\
&= \sum_{j=1}^{N} \log \mathrm{VM}(x_j; s_j, \kappa), \\
&= \sum_{j=1}^{N} \log \frac{\exp(\kappa \cos(x_j - s_j))}{2\pi I_0(\kappa)}, \\
&= -N \log(2\pi I_0(\kappa)) + \sum_{j=1}^{N} \kappa \cos(x_j - s_j). \qquad \text{(B87)}
\end{aligned}$$

With result Equation B87 in hand, we can now rewrite Equation B85 as follows,

$$\begin{aligned}
d &\approx \log \sum_{i=1}^{N_s} e^{\log p(x|s^{(i,\,C=1)})} - \log \sum_{i=1}^{N_s} e^{\log p(x|s^{(i,\,C=0)})}, \\
&= \log \sum_{i=1}^{N_s} e^{-N\log(2\pi I_0(\kappa)) + \sum_{j=1}^{N} \kappa \cos(x_j - s_j^{(i,\,C=1)})} \\
&\quad - \log \sum_{i=1}^{N_s} e^{-N\log(2\pi I_0(\kappa)) + \sum_{j=1}^{N} \kappa \cos(x_j - s_j^{(i,\,C=0)})}, \\
&= \log \sum_{i=1}^{N_s} e^{\left[\kappa \sum_{j=1}^{N} \cos\left(x_j - s_j^{(i,\,C=1)}\right)\right]} \\
&\quad - \log \sum_{i=1}^{N_s} e^{\left[\kappa \sum_{j=1}^{N} \cos\left(x_j - s_j^{(i,\,C=0)}\right)\right]}. \qquad \text{(B88)}
\end{aligned}$$

Joint Posterior Sampling

The joint posterior sampling observer uses the decision variable (Equation 5), and draws samples from the prior over C and s. For the visual search experiment we also have the location variable, L, and so the observer draws samples from the prior over C, s, and L,

$$p(s, L, C) = p(s|L, C)p(L|C)p(C). \tag{B89}$$

Following Metropolis–Hastings sampling, if we draw a proposal sample $\tilde{C}, \tilde{L}, \tilde{s}$, the probability of accepting it and transitioning away from our current state, $C^{(i)}$, $L^{(i)}$, $s^{(i)}$ (superscript in brackets denotes sample number) is

$$\begin{aligned}
p(\text{accept}) &= \min\left(1, \frac{p(\tilde{s}, \tilde{L}, \tilde{C}|x)p(s^{(i)}, L^{(i)}, C^{(i)})}{p(s^{(i)}, L^{(i)}, C^{(i)}|x)p(\tilde{s}, \tilde{L}, \tilde{C})}\right), \\
&= \min\left(1, \frac{p(x|\tilde{s})}{p(x|s^{(i)})}\right). \qquad \text{(B90)}
\end{aligned}$$

Using the previous result (Equation B87) for $\log p(x|s)$, we have that,

$$\log \frac{p(x|\tilde{s})}{p(x|s^{(i)})} = \kappa\left(\sum_{j=1}^{N} \cos(x_j - \tilde{s}_j) - \cos(x_j - s_j^{(i)})\right), \tag{B91}$$

giving,

$$p(\text{accept}) = \min\left(1, e^{\kappa(\sum_{j=1}^{N} \cos(x_j - \tilde{s}_j) - \cos(x_j - s_j^{(i)}))}\right). \tag{B92}$$

Von Mises Approximate Sampling

To make repeated sampling from the von Mises distribution feasible—something required to implement the above visual search observer models—we employed various approximations to sampling from this distribution.

(*Appendices continue*)

Usually, we used an approach inspired by sampling-importance-resampling, as described in Bishop (2006). At evenly spaced angles (6,284 angles), we evaluated the von Mises probability density function. We then sampled from the evenly spaced angles, using the values of the probability density function as weights, to determine the probability each angle was sampled.

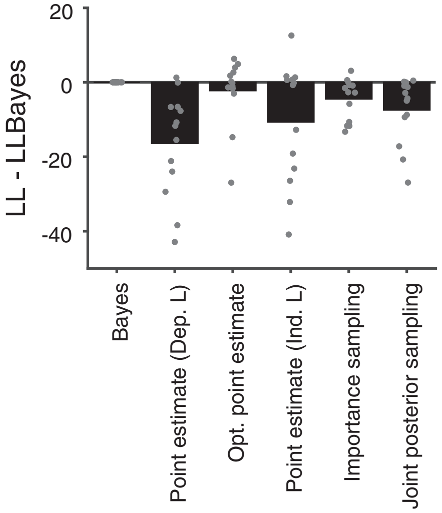
We used other approaches in specific cases. For mean $\mu = 0$, and concentration parameter $\kappa = 1.5$, we sampled with replacement from a predrawn pool of 50,000 values, that were themselves sampled from the von Mises. For $\kappa = 0$, we did not need to use an approximation, and simply sampled from the uniform distribution. For the computation of the optimal offset used by the optimal-criterion point estimate observer, we sampled directly from the von Mises without approximation.

Model Comparison With Additional Point Estimate Observer

Here, we compare the models from the main text with the additional point estimate model, point estimate Ind. *L* (see above). The point estimate observer from the main text is here referred to as the point estimate Dep. *L* model. We fitted the models in the same way as in the main text (except this time only repeating the fitting from 20 different starting positions for each model and participant). As before, we can evaluate the models using the maximum log-likelihood obtained, because all models have the same number of parameters (Figure B1). The point estimate Ind. *L* model appeared to perform slightly better than the point estimate Dep. *L* model. However, the mean maximum log-likelihood for the point estimate Ind. *L* model was lower than the mean for both the Bayesian and optimal point estimate models, leading to no change in the ordering of the models compared to the main text.

Figure B1

Model Comparison for the Visual Search Task, Including the Additional Point Estimate Model



Note. In this additional model, *L* is considered independent of *C*, which slightly changes the dependance of the bias on the number of targets. As in the main text, the maximum log-likelihood is compared to the Bayesian observer results. Dots represent individual subjects. The bars represent the mean.

Received May 14, 2021
Revision received September 25, 2022
Accepted October 4, 2022 ■