A neural basis of probabilistic computation in visual cortex

Edgar Y. Walker[1,2†], R. James Cotton[1,2,3†], Wei Ji Ma[4‡], Andreas S. Tolias[1,2,5*‡]

[1] *Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, TX, USA*

[2] *Department of Neuroscience, Baylor College of Medicine, TX, USA*

[3] *Now at: Shirley Ryan AbilityLab, IL, USA*

[4] *Center for Neural Science and Department of Psychology, New York University, NY, USA.*

[5] *Department of Electrical and Computer Engineering, Rice University, TX, USA.*

†‡ These authors contributed equally to this work

∗ Corresponding author

For more than a century, Bayesian-inspired models have been used to explain human and animal behavior, suggesting that organisms represent the uncertainty associated with sensory variables. Nevertheless, the neural code of uncertainty remains elusive. A central hypothesis is that uncertainty is encoded in the population activity of cortical neurons in the form of likelihood functions. Here, we studied the neural code of uncertainty by simultaneously recording population activity in the visual cortex in primates during a visual categorization task for which trial-to-trial uncertainty about stimulus orientation was relevant for the animal's decision. We decoded the likelihood function from the trial-to-trial population activity and found that it predicted the monkey's decisions better than using only a decoded point

estimate of the orientation. Critically, this remained true even when we conditioned on the stimulus including its contrast, suggesting that random fluctuations in neural activity firing rates drive behaviorally meaningful variations in the likelihood function. Our results establish the role of population-encoded likelihood functions in mediating behavior and offer potential neural underpinnings for Bayesian models of perception.

When making perceptual decisions, organisms often benefit from representing uncertainty about sensory variables. More specifically, the theory that the brain performs Bayesian inference—which has roots in the work of Laplace[1] and von Helmholtz[2]—has been widely used to explain human and animal perception[3–6]. At its center lies the assumption that the brain maintains a statistical model of the world and when confronted with incomplete and imperfect information, makes inferences by computing probability distributions over task-relevant variables. In spite of the prevalence of Bayesian theories in neuroscience, evidence to support them stems primarily from behavioral studies (e.g.[7,8]), and how probability distributions or uncertainty are encoded in the brain remains unclear.

According to the probabilistic population coding (PPC) hypothesis[9,10], inference in the brain is performed by inverting a generative model of neural population activity. Specifically, according to PPC, a neural population encodes sensory uncertainty in the form of the sensory likelihood function—the probability of observing a given pattern of neural activity across hypothesized stimulus values[9,11,12]. The form of the likelihood function is inherited from the probability distribution describing neural variability (“noise”) for a given stimulus. A sensory likelihood function is typically unimodal[13,14], and its width could in principle serve as a measure of the sensory uncertainty about the stimulus. Whether the brain uses this particular uncertainty quantity in its decisions is unknown.

Testing this hypothesis thoroughly is not straightforward. Experiments to test for any population-level neural code for trial-by-trial uncertainty must: (1) Use a behavioral task in which uncertainty information, and not just a point estimate, is relevant for the perceptual decision, (2) Record simultaneously from a population of neurons, and (3) Show that the stimulus-conditioned fluctuations in uncertainty decoded from population responses—for us, the sensory likelihood function—mediate the behavioral outcomes (perceptual decisions). Despite previous efforts[10,15,16], these three criteria have not yet been met simultaneously, and therefore, the population neural code of uncertainty still remains unknown.

To illustrate the importance of the third criterion, consider a typical perceptual decision-making task (Fig. 1a) where the subject views a stimulus s, which elicits a cortical population response $\mathbf{r}$, for example in V1. Here by a stimulus, we refer collectively to all aspects of a visual stimulus such as its contrast and orientation. Stimulus information is eventually relayed to decision-making areas (e.g. prefrontal cortex), leading the animal to make a classification decision $\hat{C}$. The likelihood function $\mathcal{L}$ is decoded from the recorded population activities $\mathbf{r}$. Because variation in the stimulus (e.g. orientation or contrast) across trials can drive variation both in the decoded likelihood function and in the animal's decision, one may find a significant correlation between $\mathcal{L}$ and $\hat{C}$, even if the likelihood estimated from the recorded population $\mathbf{r}$ does not mediate the decision (Fig. 1c). When the stimulus is fixed, random fluctuations in the population response $\mathbf{r}$ can still result in variations in $\mathcal{L}$. If the likelihood truly mediates the decision, we expect that such variation in $\mathcal{L}$ would account for variation in $\hat{C}$. Therefore, to demonstrate that the likelihood $\mathcal{L}$ mediates decision $\hat{C}$ (Fig. 1d), it is imperative to show a correlation between $\mathcal{L}$ and $\hat{C}$ conditioned on the stimulus s (Fig. 1d,e).

In this work, we meet these requirements and provide the first evidence that in perceptual decision-making, cortical populations encode and utilize trial-by-trial sensory uncertainty information in the form of likelihood functions. We performed simultaneous V1 cortical population recordings as monkeys performed a visual classification task in which the trial-by-trial uncertainty information is beneficial to the animal[17]. To decode the trial-by-trial likelihood functions from the V1 population responses, we developed a novel technique based on deep learning[18,19]. This method extends beyond previous approaches that used strong parametric assumptions about the probability distribution of population activity[10,14,16]. These assumptions were theoretically convenient[9,20–23] but limited the generality of those approaches in decoding the likelihood functions. Finally, we performed all analyses conditioned on the contrast—an overt driver of uncertainty—and performed further orientation-conditioned analyses to isolate the effect of random fluctuations in the decoded likelihood function on behavior.

Results

Behavioral task Two Rhesus macaques (*Macacca mulatta*) were trained on an orientation classification task designed such that the optimal performance required the use of trial-by-trial uncertainty.

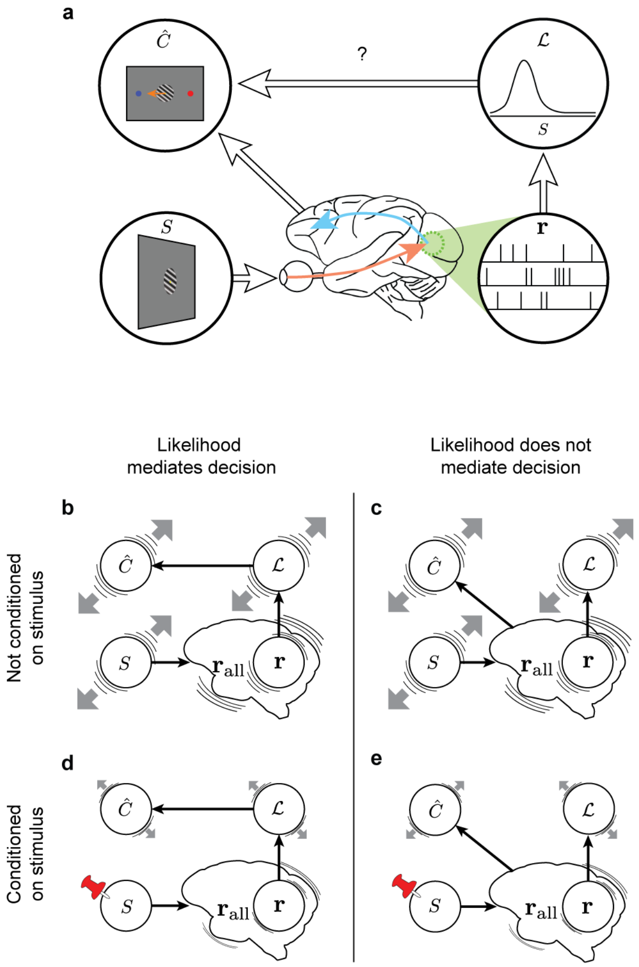

a
$\hat{C}$
?
$\mathcal{L}$
s
s
$\mathbf{r}$
Likelihood mediates decision
Likelihood does not mediate decision
b
c
d
e
Not conditioned on stimulus
Conditioned on stimulus
$\hat{C}$
$\mathcal{L}$
s
$\mathbf{r}_{\text{all}}$
$\mathbf{r}$

Figure 1: Conceptual overview of the decoding model. **a**, Information flow between stimulus s, recorded cortical population response $\mathbf{r}$, responses of all recorded and unrecorded neurons $\mathbf{r}_{\text{all}}$, decoded likelihood function $\mathcal{L}$, and subject's decision $\hat{C}$, as the subject performs a visual classification task. **b-e**, Possible relationship between variables in the model indicated by black arrows. We consider two scenarios: **b, d** the likelihood function mediates the decision, **c, e** the likelihood function does not mediate the decision. The gray arrow represents the trial-by-trial fluctuations in the subject's decisions $\hat{C}$ as predicted by the variable. **b, c**, When not conditioning on the stimulus, the stimulus can drive correlation among all variables, making it difficult to distinguish the two scenarios. **d, e**, When conditioning on the stimulus, we expect correlation between $\hat{C}$ and $\mathcal{L}$ only when $\mathcal{L}$ mediates the decision, allowing us to distinguish the two scenarios.

On each trial, one of two stimulus classes ($C = 1$ or $C = 2$) was chosen at random with equal probability. Each class was defined by a Gaussian probability distribution over the orientation. The two distributions shared the same mean but had different standard deviations (Fig. 2a). An orientation was drawn from the distribution belonging to the selected class, and a drifting grating stimulus with that orientation was then presented to the animal (Fig. 2b). In a given recording session, at least three distinct contrasts were selected at the beginning of the session, and on each trial, one of these values was randomly selected.

In our previous study[17], we designed this task so that an optimal Bayesian observer would incorporate the trial-by-trial sensory uncertainty about stimulus orientation in making classification decisions. Indeed, both humans and monkeys decisions seemed to utilize trial-by-trial uncertainty about the stimulus orientation. In the current study, one of the two monkeys (Monkey L) was the same monkey that participated in the previous study and thus has been shown to have learned the task well. The second monkey (Monkey T) was naïve to the task, but learned to perform equally well, closely matching the performance of Monkey L (Fig. 2c), with psychometric curves displaying a strong dependence on both contrast and orientation (Fig. 2d,e).

In our analyses, we grouped the trials with the same contrast within the same session and refer to such a group as a “contrast-session”.

Decoding a cortical population representation of uncertainty Each monkey was implanted with a chronic multi-electrode (Utah) array in the parafoveal primary visual cortex (V1) to record the simultaneous cortical population activity as they performed the orientation classification task

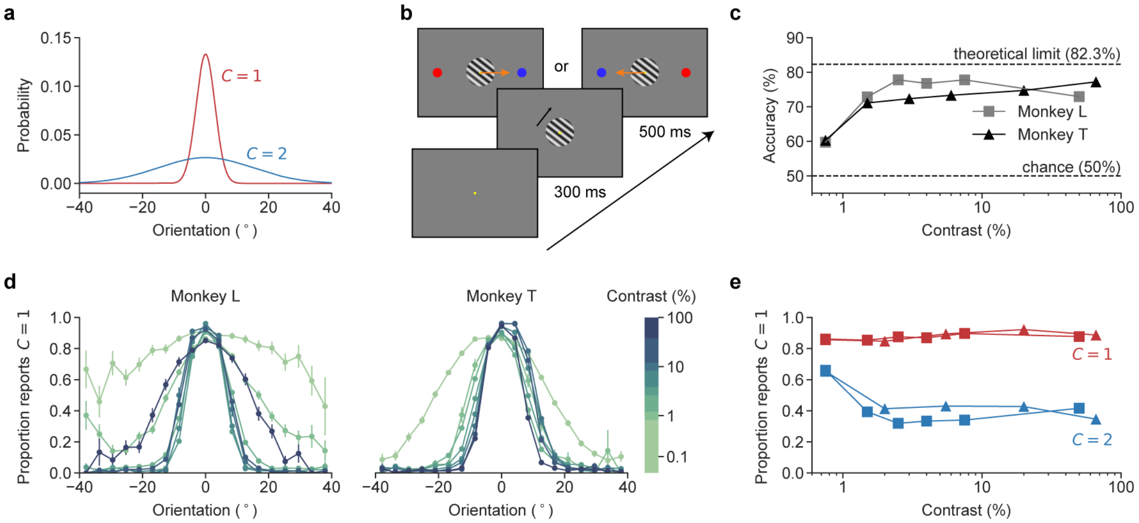


Figure 2: Behavioral task. **a**, The stimulus orientation distributions for the two classes. The two distributions shared the same mean ($\mu = 0°$) but differed in their standard deviations ($\sigma_1 = 3°$ and $\sigma_2 = 15°$). **b**, Time course of a single trial. The subject fixated onto the fixation target for 300 ms before a drifting grating stimulus was shown. After 500 ms of stimulus presentation, the subject broke fixation and saccaded to one of the two colored targets to indicate their class decision (color matches class color in **a**). The left-right configuration of the colored targets were chosen at random for each trial. **c**, Performance of the two monkeys on the task across stimulus contrast. "Theoretical limit" corresponds to the performance of an ideal observer with no observation noise. **d**, Psychometric curves. Each curve shows the proportion of trials on which the monkey reported $C = 1$ as a function of stimulus orientation, computed from all trials within a single contrast bin. All data points are means and error bars indicate standard error of the means. **e**, Class-conditioned responses. For each subject, the proportions of $C = 1$ reports is shown across contrasts, conditioned on the ground-truth class: $C = 1$ (red) and $C = 2$ (blue). The symbols have the same meaning as in **c**.

(Fig. 3a). A total of 61 and 71 sessions were analyzed from Monkey L and Monkey T for a total of 110,695 and 192,631 trials, respectively. In each recording session, up to 96 channels were recorded. On each trial and for each channel, we computed the total number of spikes that occurred during the 500 ms of stimulus presentation preceding the decision-making cue (Fig. 3a), yielding a vector of population responses $\mathbf{r}$ used in the subsequent analyses (Fig. 3b).

Existing computational methods for decoding the trial-by-trial likelihood function from the cortical population activities are typically based on making strong parametric assumptions about the stimulus conditioned distribution of the population response (i.e. generative model of the population response). For example, population responses to a stimulus can be modeled as an independent Poisson distribution, allowing each recorded unit to be characterized by a simple tuning curve (which may be further parameterized)[14,16,20–23]. While this simplifying assumption allows the trial-by-trial likelihood function to be computed straightforwardly, it disregards any potential correlations among the units in population responses (i.e. noise correlations and internal brain state fluctuations[24–29]), and can lead to biased estimates of the likelihood encoded by the cortical population. While more generic parametric models—such as Poisson-like distributions—of population distribution have been proposed[9,10,17,30,31], they still impose restrictive assumptions.

We devised a technique based on deep learning to decode the trial-by-trial likelihood function from the V1 population response. We trained a fully connected deep neural network (DNN)[19] to predict the per-trial likelihood function $\mathcal{L}(\theta) \equiv p(\mathbf{r}|\theta)$ over stimulus orientation θ from the vectorized population response $\mathbf{r}$ (Fig. 3c; for details on the network architecture and training objective, refer

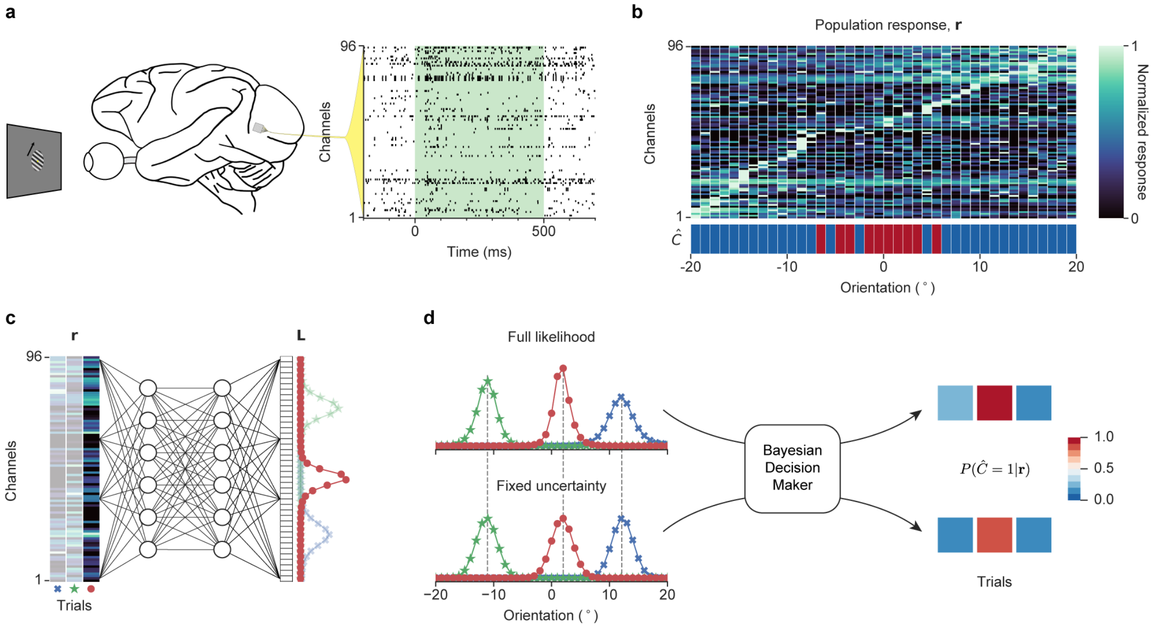


Figure 3: Encoding and decoding of the stimulus orientation. **a**, An example of 96 channels spike traces from a single trial from Monkey T. The vector of spike counts, **r**, was accumulated over the pre-saccade stimulus presentation period (time 0-500 ms, green shade). **b**, The population response for the selected trials from a single contrast-session (Monkey T, 64% contrast). Each column is a population response **r** on a trial randomly drawn from the trials falling into a specific orientation bin. Each row is a response from a single channel. For visibility, the channel's responses are normalized to the maximum response across all trials. The channels were sorted by the preferred orientation of the channel. The subject's class decision $\hat{C}$ for each trial is depicted by the colored patch: red and blue for $\hat{C} = 1$ and $\hat{C} = 2$, respectively. **c**, A schematic of a deep neural network trained to decode the vectorized likelihood function **L** from **r**. All likelihood functions are area-normalized. **d**, Two decision models M. In the *Full-Likelihood Model*, the decoded likelihood function **L** was used without modification to predict the probability that the monkey will report $\hat{C} = 1$. In the *Fixed-Uncertainty Model*, the likelihood function was approximated by a fixed width Gaussian centered at the mean of the normalized likelihood (dashed line), where the fixed width was fitted separately for each contrast-session. For both models, their likelihood functions were fed into a parameterized Bayesian decision maker to yield the decision prediction $p(\hat{C} = 1|\mathbf{r})$, where parameters of the decision maker were fitted to the subject's classification decisions, separately for each contrast-session.

to Methods). A separate network was trained for each contrast-session and no behavioral data was utilized in training the DNN. By using a DNN to directly learn the likelihood decoder, we avoid making any parametric assumption on the population response distribution. DNNs can also learn to decode likelihood functions for neural populations with known parametric generative models such as Poisson-like distributions, and therefore constitute a strictly more flexible likelihood decoding method. On simulated population responses, trained DNNs could well recover the ground-truth likelihood functions (Extended Data Fig. 1; refer to Methods for the simulation details).

The likelihood functions decoded by the DNNs exhibit the expected dependencies on the overt drivers of uncertainty, contrast (Fig. 4a-c): the width of the likelihood function is higher at lower contrast (Fig. 4d).

Decision-making model As we varied the stimulus contrast from trial to trial, the expected uncertainty about the stimulus orientation varied, and one would expect the monkeys to represent and make use of their trial-by-trial sensory uncertainty in making decisions. However, we make a much stronger claim here: even at a fixed contrast, because of random fluctuations in the population response[32,33], we predict (1) the uncertainty encoded in the population, that is, the likelihood function, to still fluctuate from trial to trial, and (2) the effect of such fluctuations to manifest in the monkey's decisions (Fig. 1d) on a trial-by-trial basis. We tested this prediction by fitting, separately for each contrast-session, the following two models and comparing their performance in predicting the monkey's decision (Fig. 3d): (1) a Full-Likelihood Model, which utilizes the trial-by-trial uncertainty information decoded from the population response in the form of the likelihood function,

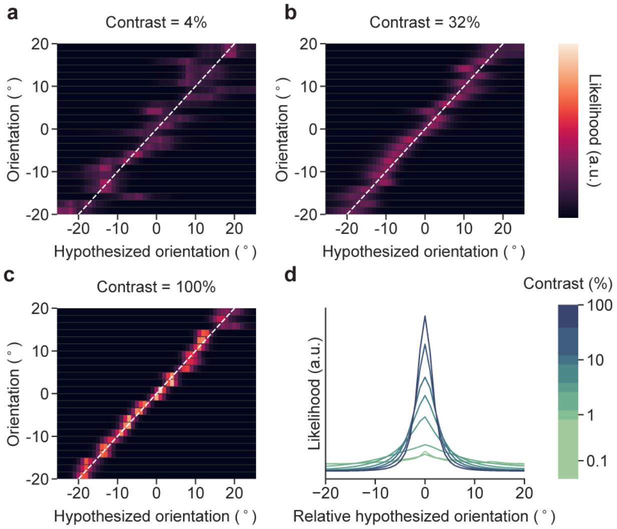


Figure 4: Likelihood functions decoded by the trained neural networks. **a-c**, Example decoded likelihood functions from three contrast-sessions from Monkey T. Each row represents the decoded likelihood function over the hypothesized orientation for a randomly selected trial within the specific orientation bin. All likelihood functions are area-normalized. Brighter colors correspond to higher values of the likelihood function. **d**, Average likelihood function by contrast. On each trial, the likelihood function was shifted such that the mean orientation of the normalized likelihood function occured at 0°. The centered likelihood functions were then averaged across all trials within the same contrast bin.

and (2) a Fixed-Uncertainty Model, which approximates the trial-by-trial likelihood function by a Gaussian function centered at the mean of the normalized likelihood function, with a fixed width that is fitted separately for each contrast-session.

In both models, the (original or Gaussian approximated) likelihood functions were fed into the Bayesian decision maker to yield trial-by-trial prediction of the subject's decision in the form of $p(\hat{C}|\mathbf{r}, M)$, or the likelihood of subject's decisions $\hat{C}$ conditioned on the population response $\mathbf{r}$ and the model M. The Bayesian decision maker computed the posterior probability of each class and used these to produce a stochastic decision. The means of the class distributions assumed by the observer, the class priors, the lapse rate, and a parameter to adjust the exact decision-making strategy were used as free parameters (refer to Methods for details). The model parameters were fitted by maximizing the total log likelihood over all trials in a contrast-session $\sum_i \log p(\hat{C}_i|\mathbf{r}_i, M)$. The fitness of the models was assessed through cross-validation, and we reported mean and total log likelihood of the models across all trials in the test set.

Both models incorporated a trial-by-trial point estimate of the stimulus orientation (the mean of the normalized likelihood function) and only differed in whether they contain additional uncertainty information about the stimulus orientation carried by the trial-by-trial fluctuations in the shape of the likelihood function. We use the term "shape" to refer to all aspects of the likelihood function besides its mean, including its width. If the fluctuations in the shape of the likelihood function truly captured the fluctuations in the sensory uncertainty as represented and utilized by the animal, one would expect the Full-Likelihood Model to yield better trial-by-trial predictions of the monkey's decisions than the Fixed-Uncertainty Model.

We observed that the Full-Likelihood Model performed well above chance in predicting the monkey's behavior across all contrasts (Extended Data Fig. 2), reaching up to 90% accuracy. In mean log likelihood, the Full-Likelihood Model consistently outperformed the Fixed-Uncertainty Model across contrasts and for both monkeys (Fig. 5a,b; trial log likelihood differences between the Full-Likelihood and Fixed-Uncertainty Model: Monkey L: paired t-test, $t(110694) = 25.32$, $p < 0.001$, $\delta_{\text{total}} = 1975.90$ and Monkey T: $t(192610) = 19.23$, $p < 0.001$, $\delta_{\text{total}} = 1611.50$; δ_{total} is the total log likelihood difference across all trials). This shows that the trial-by-trial fluctuations in the shape of the likelihood function are informative about the monkey's trial-by-trial decisions, demonstrating that decision-relevant sensory uncertainty information is contained in population responses that can be captured in the shape of the likelihood function. This in turn strongly supports the hypothesis that visual cortex encodes stimulus uncertainty in the form of a likelihood function on the trial-by-trial basis.

We next asked how meaningful our effect sizes (model performance differences) are. To answer this question, we simulated the monkey's responses across all trials and contrast-sessions taking the trained Full-Likelihood Model to be the ground truth, and then retrained the Full-Likelihood Model and the Fixed Uncertainty Model from scratch on the simulated data. This approach yields a theoretical upper bound on the observable difference between the two models if the Full-Likelihood Model was the true model of the monkeys' decision-making process. The model performance differences between the Full-Likelihood Model and the Fixed-Uncertainty Model (1975.90 and

1611.50 total log likelihood differences across all trials for Monkey L and T, respectively) were very close in magnitude to the expected total upper bound log likelihood differences of 2368.25 ± 131.45 and 2355.22 ± 144.48 based on the simulations (representing mean ± standard deviation across 5 repetitions of simulation for Monkey L and T, respectively) (Extended Data Fig. 3). This suggests that our effect sizes are meaningful and that the Full-Likelihood Model is a reasonable approximate description of the monkey's true decision-making process.

We next isolated the contribution to the perceptual decision of trial-by-trial fluctuations in the shape of the likelihood function from the contribution of the stimulus orientation, by conditioning the analysis on the stimulus orientation (Fig. 1d). Specifically, we shuffled the shapes of the decoded likelihood functions across trials within the same orientation bin, separately for each contrast-session. This shuffling preserved the expected uncertainty at each stimulus orientation—the dependence of the likelihood shape on the stimulus orientation—and the trial-by-trial correlation between the mean of the likelihood function and the subject's perceptual decision (Fig. 5c), while removing the trial-by-trial correlation between the shape of the likelihood function and the behavioral decision conditioned on the stimulus orientation.

By design, the Fixed-Uncertainty Model should perform identically on the original and the shuffled data. If the Full-Likelihood Model outperformed the Fixed-Uncertainty Model simply because it captured spurious correlations between the stimulus orientation and the shape (including the width) of the likelihood function, the performance difference between the two models should remain the same when trained and tested on the shuffled data. However, if the difference in performance

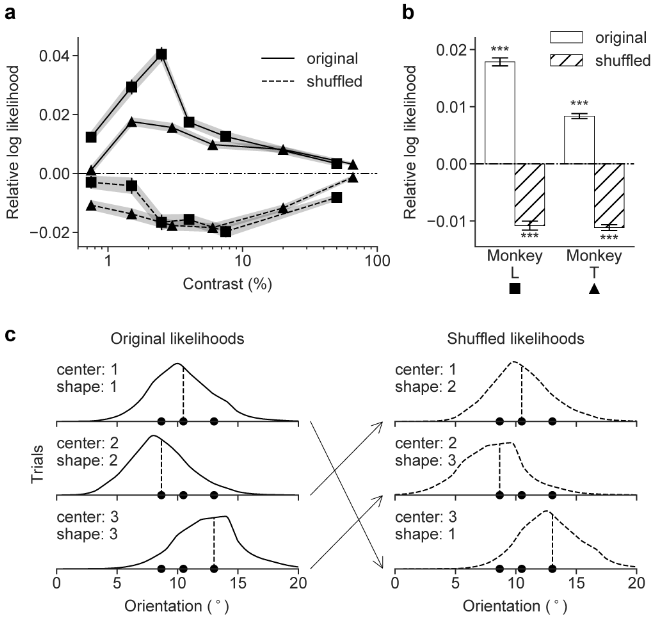


Figure 5: Model performance. **a**, Average trial-by-trial performance of the Full-Likelihood Model relative to the Fixed-Uncertainty Model across contrasts, measured as the average trial difference in the log likelihood. The results for the original (unshuffled) and the shuffled data are shown in solid and dashed lines, respectively. The squares and triangles mark Monkey L and T, respectively. **b**, Relative model performance summarized across all contrasts. Results are shown for each monkey and for original (solid) and shuffled (hatched) data. The difference between the Full-Likelihood and Fixed-Uncertainty Models was significant with $p < 0.005$ (***) for both monkeys, and on both the original and the shuffled data. Furthermore, the difference between the Full-Likelihood Model on the original and the shuffled data was significant ($p < 0.005$ for both monkeys). For **a** and **b**, all data points are means, and error bar/shaded area indicate standard error of the means. **c**, Shuffling scheme for three example trials drawn from the same stimulus orientation bin. Shuffling maintains the means but swaps the shapes of the likelihood functions.

existed because the trial-by-trial fluctuations in the likelihood shape captured the fluctuations in the decision-relevant sensory uncertainty as we hypothesized, one would expect this difference to disappear on the shuffled data.

Indeed, the shuffling of the likelihood function shapes abolished the improvement in prediction performance that the Full-Likelihood Model had over the Fixed-Uncertainty Model. In fact, the Full-Likelihood Model consistently underperformed the Fixed-Uncertainty Model on the shuffled data (Fig. 5a,b; trial log likelihood difference between the Full-Likelihood Model and the Fixed-Uncertainty Model on the shuffled data: Monkey L: paired t-test $t(110694) = -13.33$, $p < 0.001$, $\delta_{\text{total}} = -11667.74$ and Monkey T: $t(192610) = -21.19$, $p < 0.001$, $\delta_{\text{total}} = -2097.06$; δ_{total} is the total log likelihood difference across all trials). Therefore, there were significant performance differences in Full-Likelihood Model between the unshuffled and shuffled data (trial log likelihood difference: Monkey L: paired t-test $t(110694) = 32.68$, $p < 0.001$, $\delta_{\text{total}} = 3142.64$ and Monkey T: $t(192610) = 34.81$, $p < 0.001$, $\delta_{\text{total}} = 3708.56$).

To confirm our effect sizes were appropriate, we again compared these values to those obtained from simulations in which we took the Full-Likelihood Model to be the ground truth (Extended Data Fig. 3). The simulations yielded total log likelihood differences of the Full-Likelihood Model between the unshuffled and shuffled data of 3675.30 ± 248.42 (Monkey L) and 4068.98 ± 173.71 (Monkey T) (mean ± standard deviation across 5 repetitions), similar in magnitude to the observed values.

Taken together, the shuffling analyses show that the better prediction performance of the Full-

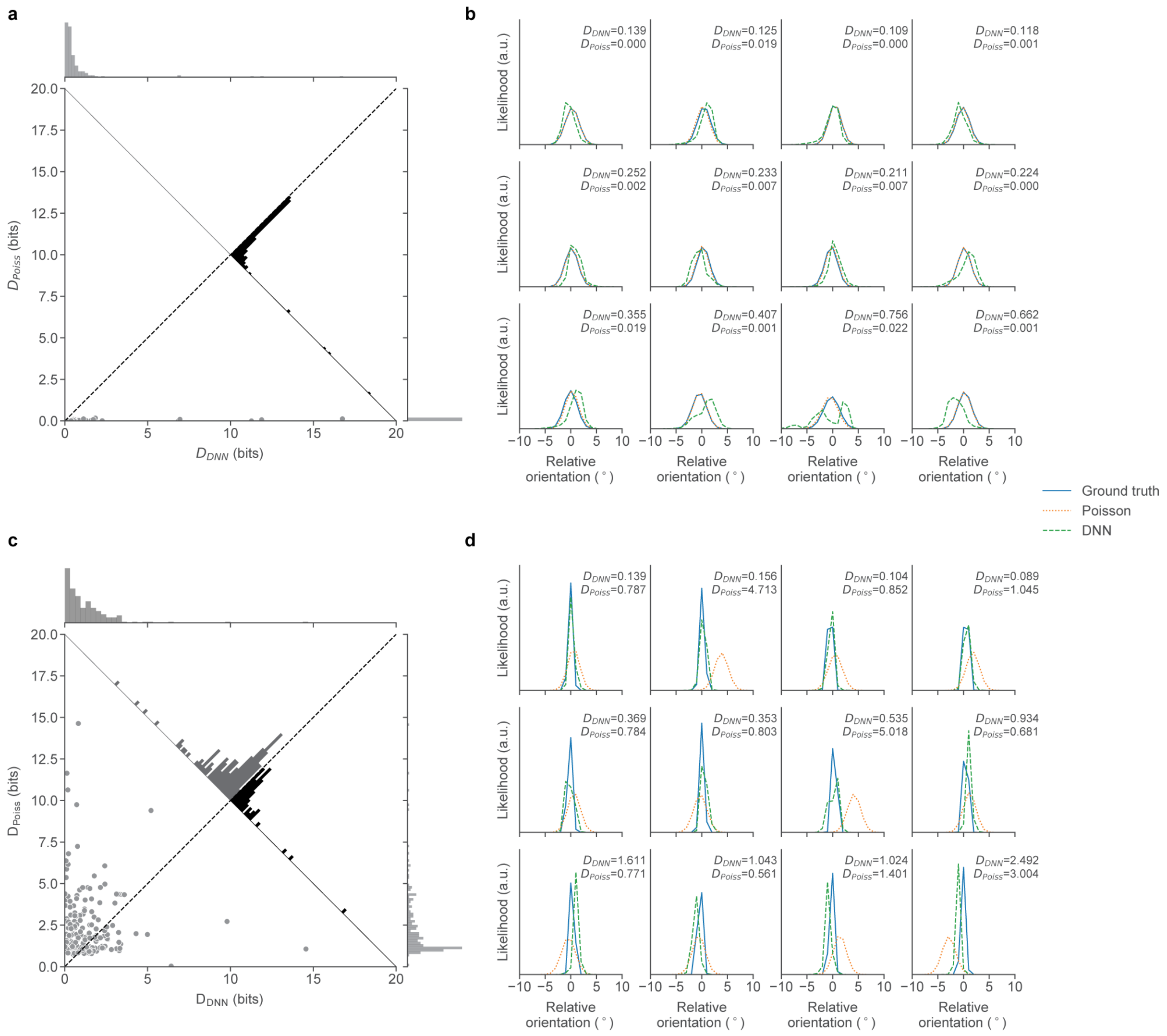
Likelihood Model is not due to the confound between the stimulus and the likelihood shape. We conclude that the trial-by-trial likelihood function decoded from the population represents behaviorally relevant stimulus uncertainty information, even when conditioned on the stimulus.

Discussion

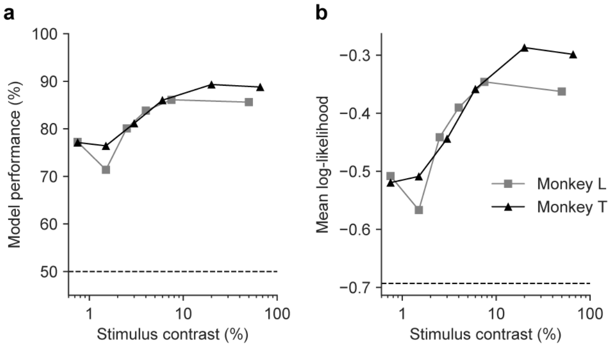
Given the stochastic nature of the brain, repeated presentations of even identical stimuli elicit variable responses. This noise creates an implicit code where multiple hypothetical stimuli are consistent with a given response but with differing likelihoods. Here, we decoded trial-to-trial likelihood functions from the population activity in visual cortex, and found that a model utilizing the full likelihood function predicted the monkeys' choices better than alternative models that ignore variations in the shape of the likelihood function. Our results provide the first population-level evidence in support of the theoretical framework of probabilistic population coding, i.e. that the brain performs Bayesian inference under a generative model of the neural activity.

Our findings were made possible by recording from a large population simultaneously and by using a task in which uncertainty is relevant to the animal. In addition, we developed a novel method to decode the likelihood functions, which is based on a deep neural network and does not rely on the strong parametric assumptions that have dominated previous work. Importantly, we conditioned our analyses on the stimulus to rule out a confounding effect of the stimulus on the observed relationship between the decoded likelihood function and the subject's decision. This is critical because previous behavioral studies on cue combination and Bayesian integration, for instance, always relied on varying stimulus features (e.g. contrast, blur, motion coherence) to manipulate uncertainty[7,8,16,34]. As a result, these studies cannot rule out that correlation between a proposed encoding of uncertainty and the behavior may be confounded by the stimulus (Fig. 1b,c), and therefore fail to provide rigorous assessment on the representation of uncertainty. Carefully controlling for the effect of stimulus fluctuations allowed us to present rigorous evidence that the trial-by-trial fluctuations in the likelihood functions carry behaviorally relevant stimulus uncertainty information.

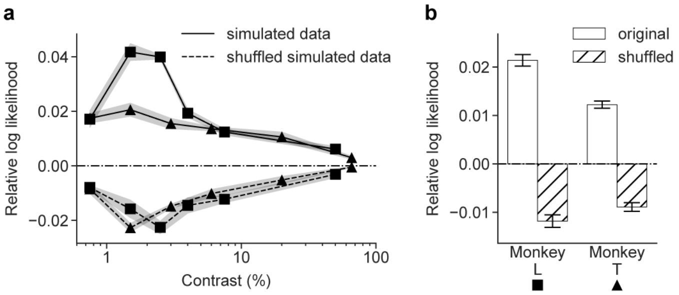
While the sensory likelihood function is a crucial building block for probabilistic computation in the brain, fundamental questions remain regarding the nature of such computation. First, how do downstream areas process the information contained in sensory likelihood functions to make better decisions? Previous work has manually constructed neural networks for downstream computation that relied heavily on the assumption of Poisson-like variability[9,10,17,35–37]. However, more recent work has demonstrated that training generic shallow networks accomplishes the same goal without the need for task-specific manual construction[38]. Second, does each area in a feedforward chain of computation encode a likelihood function over its own variable, with the computation propagating the uncertainty information from one variable to the next? For example, in our task, it is conceivable that prefrontal cortex encodes a likelihood function over class that is derived from a likelihood function over orientation coming in from V1. While answering these questions will require major efforts, we expect that our findings regarding the neural basis of sensory uncertainty will help put those efforts on a more solid footing.



Extended Data Figure 1: Performance of the likelihood functions decoded by DNN-based decoders. **a-b**, Results on independent Poisson population responses. **a**, KL divergence between the ground truth likelihood function and likelihood function decoded with: a trained DNN D_{DNN} vs. independent Poisson distribution assumption D_{Poiss}. Each point is a single trial in the test set. The distributions of D_{DNN} and D_{Poiss} are shown at the top and right margins, respectively. The distribution of pair-wise difference between D_{DNN} and D_{Poiss} is shown on the diagonal. **b**, Example likelihood functions. The ground truth (solid blue), independent-Poisson based (dotted orange), and DNN-based (dashed green) likelihood functions are shown for selected trials from the test set. Four random samples (columns) were drawn from the top, middle and bottom 1/3 of trials sorted by the D_{DNN} (rows). **c-d**, Same as in **a-b** but for simulated population responses with correlated Gaussian distribution where variance is scaled by the mean.



Extended Data Figure 2: Full-Likelihood Model performance. **a**, Model performance measured in proportions of trials correctly predicted by the model as a function of contrast for two monkeys (squares and triangles marks for Monkey L and T, respectively). On each trial, the class decision that would maximize the posterior $p(\hat{C}|\mathbf{r})$ was chosen to yield a concrete classification prediction. **b**, Same as in **a** but with performance measured as the trial-averaged log likelihood of the model. For **a** and **b**, dashed lines indicate the performance at chance (50% and $\ln(0.5)$, respectively).



Extended Data Figure 3: Expected model performance on simulated data using the trained Full-Likelihood Model as the ground truth. **a**, Average trial-by-trial performance of the Full-Likelihood Model relative to the Fixed-Uncertainty Model across contrasts on the simulated data, measured as the trial-averaged difference in the log likelihood. The results for the unshuffled and the shuffled simulated data are shown in solid and dashed lines, respectively. The squares and triangles mark Monkey L and T, respectively. **b**, Relative model performance summarized across all contrasts. Results are shown for each monkey and for unshuffled (solid) and shuffled (hatched) simulated data. For **a** and **b**, all data points are the means and error bar/shaded area indicate the standard deviation across the 5 simulation repetitions.

Methods

Experimental model and subject details All behavioral and electrophysiological data were obtained from two healthy, male rhesus macaque (*Macaca mulatta*) monkeys (L and T) aged 10 and 7 years and weighting 9.5 and 15.1 kg, respectively. All experimental procedures complied with guidelines of the NIH and were approved by the Baylor College of Medicine Institutional Animal Care and Use Committee (permit number: AN-4367). Animals were housed individually in a room located adjacent to the training facility on a 12h light/dark cycle, along with around ten other monkeys permitting rich visual, olfactory, and auditory social interactions. Regular veterinary care and monitoring, balanced nutrition and environmental enrichment were provided by the Center for Comparative Medicine of Baylor College of Medicine. Surgical procedures on monkeys were conducted under general anesthesia following standard aseptic techniques.

Stimulus presentation Each visual stimulus was a single drifting oriented sinusoidal grating (spatial frequency: 2.79 cycles/degree visual angle, drifting speed: 0.028 cycles/s) presented through a circular aperture situated at the center of the screen. The size of the aperture was adjusted to cover receptive fields of the recorded populations, extending 2.14° and 2.86° of visual angle for Monkey L and Monkey T, respectively. The orientation and contrast of the stimulus were adjusted on a trial-by-trial basis as will be described later. The stimulus was presented on a CRT monitor (at a distance of 100 cm; resolution: 1600 × 1200 pixels; refresh rate: 100 Hz) using Psychophysics Toolbox[39]. The monitor was gamma-corrected to have a linear luminance response profile. Video cameras (DALSA genie HM640; frame rate 200Hz) with custom video eye tracking software developed in LabVIEW were used to monitor eye movements.

Behavioral paradigm On a given trial, the monkey viewed a drifting oriented grating with orientation θ, drawn from one of two classes, each defined by a Gaussian probability distribution. Both distributions have a mean of 0° (grating drifting horizontally rightward, positive orientation corresponding to counter-clockwise rotation), but their standard deviations differed: $\sigma_1 = 3°$ for class 1 ($C = 1$) and $\sigma_2 = 15°$ for class 2 ($C = 2$). On each trial, the class was chosen randomly with equal probability, with the orientation of the stimulus then drawn from the corresponding distribution, $p(\theta|C)$. At the beginning of each recording session, at least three distinct values of contrasts were selected, and one of these values was chosen at random on each trial. Unlike more typical two-category tasks using distributions with identical variances but different means, optimal decision-making in our task requires the use of sensory uncertainty on a trial-by-trial basis[17].

Each trial proceeded as follows. A trial was initiated by a beeping sound and the appearance of a fixation target (0.15° visual angle) in the center of the screen. The monkey fixated on the fixation target for 300 ms within 0.5°–1° visual angle. The stimulus then appeared at the center of the screen. After 500 ms, two colored targets (red and green) appeared to the left and the right of the grating stimulus (horizontal offset of 4.29° from the center with the target diameter of 0.71° visual angle), at which point the monkey saccaded to one of the targets to indicate their choice of class. For Monkey L, the grating stimulus was removed from the screen when the saccade target appeared, while for Monkey T, the grating stimulus remained on the screen until the subject completed the task by saccading to the target. The left-right configuration of the colored targets were varied randomly for each trial. Through training, the monkey learned to associate the red and the green targets with the narrow ($C = 1$) and the wide ($C = 2$) class distributions, respectively.

For illustrative clarity, we used blue to indicate $C = 2$ throughout this document. The monkey received a juice reward for each correct response (0.10–0.15 mL).

During the training, the monkeys were first trained to perform the colored version of the task, where the grating stimulus was colored to match the correct class C for that trial (red for $C = 1$ and green for $C = 2$). Under this arrangement, the monkey simply learned to saccade to the target matching the color of the grating stimulus, although the grating stimulus orientation information was always present. As the training proceeded, we gradually removed the color from the stimulus, encouraging the monkey to make use of the orientation information in the stimulus to perform the task. Eventually, the color was completely removed, and at that point the monkey was performing the full version of the task.

Surgical Methods Our surgical procedures followed a previously established approach[29,40,41]. Briefly, a custom-built titanium cranial headpost was first implanted for head stabilization under general anesthesia using aseptic conditions in a dedicated operating room. After premedication with Dexamethasone (0.25-0.5 mg/kg; 48 h, 24 h and on the day of the procedure) and atropine (0.05 mg/kg prior to sedation), animals were sedated with a mixture of ketamine (10 mg/kg) and xylazine (0.5 mg/kg). During the surgery, anesthesia was maintained using isoflurane (0.5-2%). After the monkey was fully trained, we implanted a 96-electrode microelectrode array (Utah array, Blackrock Microsystems, Salt Lake City, UT, USA) with a shaft length of 1 mm over parafoveal area V1 on the right hemisphere. This surgery was performed under identical conditions as described for headpost implantation. To ameliorate pain, analgesics were given for 7 days following a surgery.

Electrophysiological recording and data processing The neural signals were pre-amplified at the head stage by unity gain preamplifiers (HS-27, Neuralynx, Bozeman MT, USA). These signals were then digitized by 24-bit analog data acquisition cards with 30 dB onboard gain (PXI-4498, National Instruments, Austin, TX) and sampled at 32 kHz. Broadband signals (0.5 Hz to 16 kHz) were continuously recorded using custom-built LabVIEW software for the duration of the experiment. Eye positions were tracked at 200 Hz using video cameras (DALSA genie HM640) with custom video eye tracking software developed in LabVIEW. The spike detection was performed offline according to a previously described method[27,29,40]. Briefly, a spike was detected when the signal on a given electrode crossed a threshold of five times the standard deviation of the corresponding electrode. To avoid artificial inflation of the threshold in the presence of a large number of high-amplitude spikes, we used a robust estimator of the standard deviation[42], given by median$(|x|)/0.6745$. Spikes were aligned to the center of mass of the continuous waveform segment above half the peak amplitude. Code for spike detection is available online at `https://github.com/atlab/spikedetection`. In this study, the term “multiunit” refers to the set of all spikes detected from a single channel (i.e. electrode) of the Utah array, and all analyses in the main text were performed on multiunits. For each multiunit, the total number of spikes during the 500 ms of pre-target stimulus presentation, r_i for the i^{th} unit, was used as the measure of the multiunit's response for a single trial. The population response $\mathbf{r}$ is the vector of spike counts for all 96 multiunits.

Dataset and inclusion criteria. We recorded a total of 61 and 71 sessions from Monkey L and T, for a total of 112,072 and 193,629 trials, respectively. We removed any trials with electrophysiology recordings contaminated by noise in the recording devices (e.g. poor grounding connector resulting in movement noise) or equipment failures. To do so, we established the following trial inclusion criteria:

1. The total spike counts $r_t = \sum_i r_i$ across all channels should fall within the $\pm 4\sigma_{\text{adj}}$ from the median total spike counts across all trials from a single session. σ_{adj} is the standard deviation of the total spike count distribution robustly approximated using the interquartile range IQR as follows: $\sigma_{\text{adj}} = \frac{\text{IQR}}{1.35}$.

2. For at least 50% of all units, the observed i^{th} unit spike count r_i for the trial should fall within a range defined as: $|r_i - \text{MED}_i| \leq 1.5 \cdot \text{IQR}_i$, where MED_i and IQR_i are the median and interquartile ranges of the i^{th} unit spike counts distribution throughout the session, respectively.

We only included trials that satisfied both of the above criteria in our analysis. Empirically, we found the above criteria to be effective in catching obvious anomalies in the spike data while introducing minimal bias into the data. After the application of the criteria, we were left with 110,695 and 192,631 trials for Monkey L and T, thus retaining 98.77% and 99.48% of the total trials, respectively. While this selection criteria allowed us to remove apparent anomaly in the data, we found that the main findings described in this paper were not sensitive to the precise definition of the inclusion criteria.

During each recording session, stimuli were presented under three or more contrast values. In all analyses to follow, we studied the trials from distinct contrast separately for each recording session, and we refer to this grouping as a “contrast-session”.

Receptive field mapping On the first recording session for each monkey, the receptive field was mapped using spike-triggered averaging of the multiunit responses to a white noise random dot stimulus. The white noise stimulus consisted of square dots of size 0.29° of visual angle presented on a uniform gray background, with randomly varying location and color (black or white) every 30 ms for 1 second. We adjusted the size of the grating stimulus as necessary to ensure that the stimulus covers the population receptive field entirely.

Likelihood decoding Given the population activity $\mathbf{r}$ in response to an orientation θ, we aimed to decode uncertainty information in the form of a likelihood function $\mathcal{L}(\theta) \equiv p(\mathbf{r}|\theta)$, as a function of θ. This may be computed through the knowledge of the generative relation leading from θ to $\mathbf{r}$—that is, by describing the underlying orientation conditioned probability distribution over $\mathbf{r}$, $p(\mathbf{r}|\theta)$. This procedure is typically approximated by making rather strong assumptions about the form of the density function, for example by assuming that neurons fire independently and each neuron fires according to the Poisson distribution[21]. Under this approach, the expected firing rates (i.e. tuning curves) of the i^{th} neuron $\mathbb{E}[r_i|\theta] = f_i(\theta)$ must be approximated as well, for example by fitting a parametric function (e.g. von Mises tuning curves[43]) or employing kernel regression[21]. While these approaches have proven useful, the effect of the strong and likely inaccurate assumptions on the decoded likelihood function remains unclear. Ideally, we can more directly estimate the likelihood function $\mathcal{L}(\theta)$ without having to make strong assumptions about the underlying conditional probability distribution over $\mathbf{r}$.

To this end, we employed a deep neural network (DNN)[18] to directly approximate the likelihood function over the stimulus orientation, θ, from the recorded population response $\mathbf{r}$. Here we present a brief derivation that serves as the basis of the network design and training objective. Let us assume that m multiunits were recorded simultaneously in a single recording session, so that $\mathbf{r} \in \mathbb{R}^m$. To make the problem tractable, we bin the stimulus orientation θ into n distinct values, θ_1 to θ_n (the derivation holds in general for arbitrarily fine binning of the orientation). With this, the likelihood function can be captured by a vector $\mathbf{L} \in \mathbb{R}^n$ where $\mathbf{L}_i = \mathcal{L}(\theta_i)$. Let us assume that we can train some DNN to learn a mapping f from the population response $\mathbf{r}$ to the log of the likelihood function $\mathbf{L}$ up to a constant offset b. That is, $f : \mathbb{R}^m \mapsto \mathbb{R}^n$,

$$\mathbf{r} \mapsto f(\mathbf{r}) = \log \mathbf{L} + b(\mathbf{r}) = \log p(\mathbf{r}|\theta) + b(\mathbf{r}) \tag{1}$$

for some scalar function $b \in \mathbb{R}$. As the experimenter, we know the distribution of the stimulus orientation, $\mathbf{p}_\theta \in \mathbb{R}^n$, where $\mathbf{p}_{\theta,i} = p(\theta_i)$. We combine $f(\mathbf{r})$ and $\mathbf{p}_\theta$ to compute the log posterior over stimulus orientation θ up to some scalar value $b'(\mathbf{r})$,

$$\mathbf{z}(\mathbf{r}) \equiv \log \mathbf{p}_\theta + f(\mathbf{r}) = \log p(\theta|\mathbf{r}) + b'(\mathbf{r}) \tag{2}$$

We finally take the softmax of $\mathbf{z}(\mathbf{r})$, and recover the normalized posterior function $\mathbf{q}(\mathbf{r}) \equiv \text{softmax}(\mathbf{z}(\mathbf{r}))$ where,

$$\mathbf{q}_i(\mathbf{r}) = \frac{e^{\mathbf{z}_i(\mathbf{r})}}{\sum_j e^{\mathbf{z}_j(\mathbf{r})}} \tag{3}$$

$$= \frac{e^{b'(\mathbf{r})} p(\theta = \theta_i|\mathbf{r})}{e^{b'(\mathbf{r})} \sum_j p(\theta = \theta_j|\mathbf{r})} \tag{4}$$

$$= p(\theta = \theta_i|\mathbf{r}) \tag{5}$$

Overall, $\mathbf{q}(\mathbf{r}) = \text{softmax}(\log \mathbf{p} + f(\mathbf{r}))$.

The goal then is to train the DNN $f(\mathbf{r})$ such that the overall function $\mathbf{q}(\mathbf{r})$ matches the posterior over the stimulus, $\mathbf{p}(\mathbf{r})$ where $\mathbf{p}_i(\mathbf{r}) = p(\theta = \theta_i|\mathbf{r})$ based on the available data. This in turn allows the network output $f(\mathbf{r})$ to approach the log of the likelihood function $\mathbf{L}$, up to a constant $b(\mathbf{r})$. For 1-out-of-n classification problems, minimizing the cross-entropy between $\mathbf{q}(\mathbf{r})$ and the stimulus orientation θ for a given $\mathbf{r}$ lets the overall function $\mathbf{q}(\mathbf{r})$ approach the true posterior $\mathbf{p}(\mathbf{r})$, as desired[44,45]. To show this, let us start by minimizing the difference between the model estimated posterior $\mathbf{q}(\mathbf{r})$ and the true posterior $\mathbf{p}(\mathbf{r})$ over the distribution of $\mathbf{r}$. We do this by minimizing the loss L defined as the expected value of the Kullback-Leibler divergence[46] between the two posteriors:

$$\begin{aligned}
L(W) &= \mathbb{E}_{\mathbf{r}}\left[D_{KL}(\mathbf{p}||\mathbf{q})\right] && (6)\\
&= \mathbb{E}_{\mathbf{r}}\left[\mathbb{E}_{\theta|\mathbf{r}}\left[\log\frac{p(\theta|\mathbf{r})}{q(\theta|\mathbf{r},W)}\right]\right] && (7)\\
&= \mathbb{E}_{\mathbf{r},\theta}\left[\log\frac{p(\theta|\mathbf{r})}{q(\theta|\mathbf{r},W)}\right] && (8)\\
&= -\mathbb{E}_{\mathbf{r},\theta}\left[\log q(\theta|\mathbf{r},W)\right] - H(\theta|\mathbf{r}) && (9)
\end{aligned}$$

where $p(\theta = \theta_i|\mathbf{r}) \equiv \mathbf{p}_i(\mathbf{r})$, $q(\theta = \theta_i|\mathbf{r}, W) \equiv \mathbf{q}_i(\mathbf{r}, W)$, W is a collection of all trainable parameters in the network, and $H(\theta|\mathbf{r})$ is the conditional entropy of θ conditioned on $\mathbf{r}$, which is an unknown but a fixed quantity with respect to W and the data distribution. Here we used the notation $\mathbf{q}(\mathbf{r}, W)$ to highlight the dependence of the network estimated posterior $\mathbf{q}(\mathbf{r})$ on the network parameters W. We can redefine the loss, L^*, only leaving the terms that depends on the trainable parameters W, and then apply a Monte Carlo method[47] to approximate the loss from samples:

$$\begin{aligned}
L^*(W) &= -\mathbb{E}_{\mathbf{r},\theta}\left[\log q(\theta|\mathbf{r},W)\right] && (10)\\
&\approx -\frac{1}{N}\sum_i \log q(\theta^{(i)}|\mathbf{r}^{(i)},W) && (11)
\end{aligned}$$

where $\left(\theta^{(i)}, \mathbf{r}^{(i)}\right)$ are samples drawn from a training set for the network. Eq. 11 is precisely the definition of the cross-entropy as we set out to show.

Therefore, by optimizing the overall function $\mathbf{q}(\mathbf{r})$ to match the posterior distribution through the use of cross-entropy loss, the network output $f(\mathbf{r})$ can approximate the log of the likelihood function $\mathcal{L}(\theta)$ for each $\mathbf{r}$ up to an unknown constant $b(\mathbf{r})$. Because we do not know the value of $b(\mathbf{r})$, the network will not learn to recover the underlying generative function linking from θ to $\mathbf{r}$, $p(\mathbf{r}|\theta)$.

As an example, consider a neural population with responses that follows a Poisson-like distribution (i.e. a version of the exponential distribution with linear sufficient statistics[9,10]). Learning a decoder for such population responses occurs as a special case of training a DNN-based likelihood decoder. For Poisson-like variability, the stimulus-conditioned distribution over $\mathbf{r}$ is $p(\mathbf{r}|\theta) = \phi(\mathbf{r})e^{\mathbf{h}^\top(\theta)\mathbf{r}}$. The log likelihood function is then $\log \mathbf{L} = \log \phi(\mathbf{r}) + \mathbf{H}^\top \mathbf{r}$, where $\mathbf{H}$ is a matrix whose i^{th} column is $\mathbf{h}(\theta_i)$. If we let $f(\mathbf{r}) = \mathbf{H}^\top \mathbf{r}$, then $f(\mathbf{r}) = \log \mathbf{L} + b(\mathbf{r})$ as desired, for $b(\mathbf{r}) = -\log \phi(\mathbf{r})$. Hence, if we used a simple fully connected network, training the network is equivalent to fitting the kernel function $\mathbf{h}(\theta)$ of the Poisson-like distribution.

In this work, we modeled the mapping $f(\mathbf{r})$ as a DNN with two hidden layers[19], consisting of two repeating blocks of a fully connected layer followed by a rectified linear unit (ReLU)[18] and a drop-out layer[48], and a fully connected readout layer with no output nonlinearity (Figure 3c). To encourage smoother likelihood functions, we added an L_2 regularizer on $\log \mathbf{L}$ filtered with a Laplacian filter of the form $\mathbf{h} = [-0.25, 0.5, -0.25]$. Therefore, the training loss included the term:

$$R = \gamma \sum_i \mathbf{u}_i^2 \tag{12}$$

for $\mathbf{u} = (\log \mathbf{L}) * \mathbf{h}$, where $*$ denotes convolution operation, $\mathbf{u}_i$ is the i^{th} element of the filtered log likelihood function $\mathbf{u}$, and γ is the weight on the smoothness regularizer.

We trained a separate instance of the network for each contrast-session. During the training, each contrast-session was randomly split in proportions of 80% / 20% to yield the training set and the validation set, respectively. The stimulus orientation θ was binned into integers in the range $[-45°, 45°]$, and we excluded trials with orientations outside this range. This led to the exclusion of 157 out of 110,695 trials (0.14%) and 254 out of 192,631 trials (0.13%) for Monkey L and T data, respectively. The network was trained on the training set, and its performance on the validation set was monitored to perform early stopping[49], and subsequently hyperparameter selection. For early stopping, we computed the mean squared error (MSE) between the maximum-a-posteriori (MAP) readout of the network output posterior $\mathbf{q}$ and the stimulus orientation θ on the validation set, and the training was terminated (early-stopped) if MSE failed to improve over 400 consecutive epochs, where each epoch is defined as one full pass through the training set. Upon early stopping, the parameter set that yielded the best validation set MSE during the course of the training was restored. Once the training was complete, the trained network was evaluated on the validation set to yield the final score, which served as the basis for our hyperparameter selections. The values of hyperparameters for the networks, including the size of hidden layers, the weight on the likelihood function smoothness regularizer γ, and drop-out rates during the training were selected by performing a grid search over candidate values to find the combination that yielded the best validation set score for each contrast-session instance of the network.

Decoding likelihood functions from known response distributions To assess the effectiveness of the DNN-based likelihood decoding method described above, we simulated neural population responses with known noise distributions, trained DNN decoders on the simulated population responses, and compared the decoded likelihood functions to the ground-truth likelihood functions obtained by inverting the known generative model for the responses. We also compared the quality of the DNN-decoded likelihood functions to those decoded by assuming independent Poisson distribution on the population responses, as done in previous work[14,16,20,21,23].

We simulated the activities of a population of 96 multiunits $\mathbf{r}_{\text{sim}}$ responding to the stimulus orientation θ drawn from the the distribution defined for our task such that:

$$p(\theta) = \frac{1}{2}\mathcal{N}(\theta; 0, \sigma_1^2) + \frac{1}{2}\mathcal{N}(\theta; 0, \sigma_2^2) \tag{13}$$

where $\sigma_1 = 3°$ and $\sigma_2 = 15°$.

We modeled the expected response of i^{th} unit to θ—that is, the tuning function $f_i(\theta)$—with a Gaussian function:

$$f_i(\theta) = Ae^{-\frac{(\theta-\mu_{\text{sim},i})^2}{2\sigma_{\text{sim}}^2}} \tag{14}$$

For the simulation, we have set $A = 6$ and $\sigma_{\text{sim}} = 21°$. We let the mean of the Gaussian tuning curves for the 96 units to uniformly tile the stimulus orientation between $-40°$ and $40°$. In other words,

$$\mu_{\text{sim},i} = -40 + \frac{16}{19}(i-1) \tag{15}$$

for $i \in [1, 96]$.

For any given trial with a drawn orientation θ, the population response $\mathbf{r}_{\text{sim}}$ was then generated under two distinct models of distributions. In the first case, the population responses were drawn from an independent Poisson distribution as is commonly assumed in many works:

$$p(\mathbf{r}_{\text{sim}}|\theta) = \prod_i \text{Poiss}(r_{\text{sim},i}; f_i(\theta)) \tag{16}$$

$$= \prod_i \frac{f_i(\theta)^{r_{\text{sim},i}} e^{-f_i(\theta)}}{r_{\text{sim},i}!} \tag{17}$$

In the second case, the population responses were drawn from a multivariate Gaussian distribution with covariance matrix $\Sigma \in \mathbb{R}^{96\times 96}$ that scales with the mean response of the population. That is:

$$p(\mathbf{r}_{\text{sim}}|\theta) = \mathcal{N}(\mathbf{r}_{\text{sim}}; \mathbf{f}(\theta), \Sigma(\theta)) \tag{18}$$

for

$$\Sigma(\theta) = (\mathbf{f}^{1/2}(\theta))^{\top} C(\mathbf{f}^{1/2}(\theta)) \tag{19}$$

where $\mathbf{f}^{1/2}(\theta) \in \mathbb{R}^{96}$ such that $\mathbf{f}_i^{1/2}(\theta) = \sqrt{f_i(\theta)}$, and $C \in \mathbb{R}^{96\times 96}$ is a correlation matrix. Under this distribution, the variance of any unit's response scales linearly with its mean just as in the case of the Poisson distribution, but the population responses can be highly correlated depending on the choice of the correlation matrix C. For the simulation, we randomly generated a correlation matrix with the average units correlation of 0.227.

For each case of the distribution, we simulated population responses for the total of 1200 trials. Among these, 200 trials were set aside as the test set. We trained the DNN-based likelihood decoder on the remaining 1000 trials, splitting them further into 800 and 200 trials as the training and validation set, respectively. We followed the exact DNN training and hyperparameter selection procedure as described earlier.

For comparison, we also decoded the likelihood function from the population response $\mathbf{r}_{\text{sim}}$ under the assumption of independent Poisson variability, regardless of the "true" distribution. Each unit's responses over the 1000 trials were fitted separately with a Gaussian tuning curve (Eq. 14). The parameters of the tuning curve A_i, μ_i and $\sigma_{\text{sim, i}}$ were obtained by minimizing the least square difference between the Gaussian tuning curve and the observed i^{th} unit's responses $(\theta, r_{\text{sim},i})$ using `least_squares` function from Python SciPy optimization library.

The ground-truth likelihood function $p(\mathbf{r}_{\text{sim}}|\theta)$ was computed for each simulated trial according to the definition of the distribution as found in Eq. 16 for the independent Poisson population or Eq. 18 for the mean scaled correlated Gaussian population.

We then assessed the quality of the decoded likelihood functions under the independent Poisson model $\mathcal{L}_{\text{Poiss}}(\theta)$ and under the DNN model $\mathbf{L}_{\text{DNN}}$ by computing their Kullback-Leibler (KL) divergence[46] from the ground-truth likelihood function $\mathcal{L}_{\text{gt}}(\theta)$, giving rise to D_{Poiss} and D_{DNN}, respectively. All continuous likelihood functions ($\mathcal{L}_{\text{gt}}$ and $\mathcal{L}_{\text{Poiss}}$) were sampled at orientation θ where $\theta \in \mathbb{Z}$ and $\theta \in [-45°, 45°]$, giving rise to discretized likelihood functions $\mathbf{L}_{\text{gt}}$ and $\mathbf{L}_{\text{Poiss}}$ matching the dimensionality of the discretized likelihood function $\mathbf{L}_{\text{DNN}}$ computed by the DNN. We then computed the KL divergence as:

$$D_{\text{Poiss}} = \sum_i \log \frac{\mathbf{L}_{\text{gt},i}}{\mathbf{L}_{\text{Poiss},i}} \mathbf{L}_{\text{gt},i} \tag{20}$$

and

$$D_{\text{DNN}} = \sum_i \log \frac{\mathbf{L}_{\text{gt},i}}{\mathbf{L}_{\text{DNN},i}} \mathbf{L}_{\text{gt},i} \tag{21}$$

We computed the KL divergence for both models across all 200 trials in the test set for both simulated population distributions (Extended Data Fig. 1a,c). When the simulated population distribution was independent Poisson, then $D_{\text{Poiss}} < D_{\text{DNN}}$ for all test set trials (Extended Data Fig. 1a), indicating that $\mathbf{L}_{\text{Poiss}}$ better approximated $\mathbf{L}_{\text{gt}}$ overall than $\mathbf{L}_{\text{DNN}}$. However, $\mathbf{L}_{\text{DNN}}$ still closely approximated $\mathbf{L}_{\text{gt}}$ (Extended Data Fig. 1b).

When the simulated population distribution was mean scaled correlated Gaussian, $\mathbf{L}_{\text{DNN}}$ better approximated $\mathbf{L}_{\text{gt}}$ than $\mathbf{L}_{\text{Poiss}}$ on the majority of the trials (Extended Data Fig. 1c). Furthermore, $\mathbf{L}_{\text{Poiss}}$ provided qualitatively worse fit to the $\mathbf{L}_{\text{gt}}$ for the simulated correlated Gaussian distribution compared to the fit of $\mathbf{L}_{\text{DNN}}$ to $\mathbf{L}_{\text{gt}}$ for the simulated independent Poisson distribution (Extended Data Fig. 1b,d).

Overall, the simulation results suggest that (1) when the form of the underlying population distribution is known (such as in the case of independent Poisson population), more accurate likelihood functions can be decoded by directly using the knowledge of the population distribution than through the DNN-based likelihood decoder, but (2) when the form of the underlying distribution is unknown (such as in the case of the mean scaled correlated Gaussian distribution), then a DNN-based likelihood decoder can yield much more accurate likelihood functions than if one was to employ a wrong assumption about the underlying distribution in decoding likelihood functions, and (3) a DNN-based likelihood decoder can provide reasonable estimate of the likelihood function across wide range of underlying distributions. Because the true underlying population distribution is hardly ever known to the experimenter, we believe that our DNN-based likelihood decoder stands as the most flexible method in decoding likelihood functions from the population responses to stimuli.

Mean and standard deviation of likelihood function For uses in the subsequent analyses, we computed the mean and the standard deviation of the likelihood function by treating the likelihood function as an unnormalized probability distribution:

$$\mu_L = \frac{\int \theta \mathcal{L}(\theta)\, d\theta}{\int \mathcal{L}(\theta)\, d\theta} \tag{22}$$

$$\sigma_L = \sqrt{\frac{\int (\theta - \mu_L)^2 \mathcal{L}(\theta)\, d\theta}{\int \mathcal{L}(\theta)\, d\theta}} \tag{23}$$

We took the μ_L and σ_L to be the point estimate of the stimulus orientation and the measure of the spread of the likelihood function, respectively, used in all subsequent analyses. Although not presented here, we performed the following analyses with other point estimates of the stimulus orientation such as the orientation at the maximum of the likelihood function and the median of the likelihood functions, and observed that models with mean of the likelihood function as the point estimate performed the best.

Decision-making models Given the hypothesized representation of the stimulus and its uncertainty in the form of the likelihood function $\mathcal{L}(\theta) \equiv p(\mathbf{r}|\theta)$, the monkey's trial-by-trial decisions were modeled based on the assumption that the monkey computes the posterior probability over the two classes $C = 1$ and $C = 2$, and utilizes this information in making decisions—that is, in accordance to a model of a Bayesian decision maker. The orientation distributions for the two classes are $p(\theta|C = 1) = \mathcal{N}(\theta; \mu, \sigma_1^2)$ and $p(\theta|C = 2) = \mathcal{N}(\theta; \mu, \sigma_2^2)$ with $\mu = 0$ and $\sigma_1 = 3^\circ$ and $\sigma_2 = 15^\circ$ where $\mathcal{N}(\theta; \mu, \sigma^2)$ denotes a Gaussian distribution over θ with mean μ and variance σ^2. The posterior ratio ρ for the two classes is:

$$\begin{aligned}\rho &= \frac{p(C=2|\mathbf{r})}{p(C=1|\mathbf{r})} && (24)\\ &= \frac{p(C=2)\int p(\mathbf{r}|\theta)p(\theta|C=2)\,d\theta}{p(C=1)\int p(\mathbf{r}|\theta)p(\theta|C=1)\,d\theta} && (25)\\ &= \frac{p(C=2)\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_2^2)\,d\theta}{p(C=1)\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_1^2)\,d\theta} && (26)\end{aligned}$$

A Bayes-optimal observer should select the class with the higher probability—a strategy known as maximum-a-posteriori (MAP) decision-making:

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, p(C|\mathbf{r}) \tag{27}$$

where $\hat{C}$ is the subject's decision. However, according to the posterior probability matching strategy[50,51], the decision of subjects on certain tasks are better modeled as sampling from the posterior probability:

$$p(\hat{C}) = p(C=\hat{C}|\mathbf{r}) \tag{28}$$

To capture either decision-making strategy, we modeled the subject's classification decision probability ratio as follows:

$$\frac{p(\hat{C}=2)}{p(\hat{C}=1)} = \left(\frac{p(C=2|\mathbf{r})}{p(C=1|\mathbf{r})}\right)^{\alpha} = \rho^{\alpha} \tag{29}$$

where $\alpha \in \mathbb{R}^+$. When $\alpha = 1$, the decision-making strategy corresponds to the posterior probability matching while $\alpha = \infty$ corresponds to the MAP strategy[51]. We fitted the value of α for each contrast-session during the model fitting to capture any variation of the strategy. Furthermore, we incorporated a lapse rate λ, a fraction of trials on which the subject does not pay attention and makes a random decision. Hence, the final probability that the subject selects the class $C = 1$ was modeled as:

$$
\begin{aligned}
p(\hat{C} = 1) &= (1-\lambda)\frac{1}{1+\rho^{\alpha}} + 0.5\lambda && (30)\\
&= (1-\lambda)\left[1 + \left(\frac{p(C=2)\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_2^2)\,d\theta}{p(C=1)\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_1^2)\,d\theta}\right)^{\alpha}\right]^{-1} + 0.5\lambda && (31)\\
&= (1-\lambda)\left[1 + \left(\frac{(1-p(C=1))\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_2^2)\,d\theta}{p(C=1)\int \mathcal{L}(\theta)\mathcal{N}(\theta;\mu,\sigma_1^2)\,d\theta}\right)^{\alpha}\right]^{-1} + 0.5\lambda && (32)
\end{aligned}
$$

For each contrast-session, we fitted the above Bayesian decision model to the monkey's decisions by fitting the four parameters: μ, $p(C = 1)$, α, and λ. Fitting μ (the center of stimulus orientation distributions) and $p(C = 1)$ (prior over class) allowed us to capture the bias in the stimulation distribution that the subject may have acquired errorneously during the training, and fitting α and λ allowed for the model to match the decision-making strategy employed by the subject.

Utilizing the likelihood function $\mathcal{L}(\theta)$ decoded from the V1 population response via the DNN in Eq. 32 gave rise to the Full-Likelihood Model that made use of all information including the trial-by-trial uncertainty information as captured by the shape of the likelihood function. Alternatively, we approximated the trial-by-trial decoded likelihood function for a contrast-session with a (fitted) fixed width Gaussian function whose mean matched that of the normalized decoded likelihood function. This Fixed-Uncertainty Model effectively discarded all trial-by-trial fluctuations in the uncertainty as captured by the shape of the likelihood function but preserved the point estimate of the stimulus orientation $\hat{\theta}$ (i.e. mean of the likelihood function). That is:

$$\hat{\mathcal{L}}(\theta) = \mathcal{N}(\theta; \hat{\theta}, \sigma_c^2) \tag{33}$$

where $\hat{\theta} = \mu_L$ (Eq. 22). For each contrast-session, different values were fitted for the width of the Gaussian likelihood function, σ_c. Therefore, the Fixed-Uncertainty Model had 5 parameters to be fitted in total: μ, $p(C = 1)$, α, λ, and σ_c.

Model fitting and model comparison We used 10-fold cross-validation to fit and evaluate both decision models, separately for each contrast-session. We divided all trials from a given contrast-session randomly into 10 equally sized subsets, $B_1, B_2, \ldots, B_i, \ldots, B_{10}$ where B_i is the i^{th} subset. We then held out a single subset B_i as the test set, and trained the decision-making model on the remaining 9 subsets combined together, serving as the training set. The predictions and the performance of the trained model on the held out test set B_i was then reported. We repeated this 10 times, iterating through each subset as the test set, training on the remaining subsets.

The decision models were trained to minimize the negative log likelihood on the subject's decision across all trials in the training set:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left(-\log \prod_i p(\hat{C} = \hat{C}_i | M, \Theta) \right) \tag{34}$$

$$= \underset{\Theta}{\operatorname{argmin}} \left(-\sum_i \log p(\hat{C} = \hat{C}_i | M, \Theta) \right) \tag{35}$$

where Θ is the collection of the parameters for the decision-making model M and $\hat{C}_i$ is the subject's decision on the i^{th} trial in the training set. The term $p(\hat{C}|M, \Theta)$ is given by the Eq. 32 with either the unmodified $\mathcal{L}(\theta)$ in the Full-Likelihood Model or a Gaussian approximation to $\mathcal{L}(\theta)$ in the Fixed-Uncertainty Model.

The optimizations were performed using three algorithms: `fmincon` and `ga` from MATLAB's optimization toolbox and Bayesian Adaptive Direct Search (BADS)[52]. When applicable, the optimization was repeated with 50 or more random parameter initializations. For each cross-validation fold, we retained the parameter combination $\hat{\Theta}$ that yielded the best training set score (i.e. lowest negative log likelihood) among all optimization runs across different algorithms and parameter initializations. We subsequently tested the model M with the best training set parameter $\hat{\Theta}$ and reported the score on the test set. For each contrast-session, all analyses on the trained model presented in the main text were performed on the aggregated test sets scores.

Likelihood shuffling analysis To assess the contribution of the trial-by-trial fluctuations in the decoded likelihood functions in predicting the animal's decisions under the Full-Likelihood Model, for each contrast-session we shuffled the likelihood functions among trials in the same stimulus orientation bin, while maintaining the trial to trial relationship between the point estimate of the stimulus orientation (i.e. mean of the normalized likelihood) and the perceptual decision. Specifically, we binned trials to the nearest orientation degree such that each bin was centered at an integer degree (i.e. bin center $\in \mathbb{Z}$) with the bin width of $1°$. We then shuffled the likelihood functions among trials in the same orientation bin. This effectively removed the stimulus orientation conditioned correlation between the likelihood function and the subject's classification $\hat{C}$, while preserving the expected likelihood function for each stimulus orientation.

However, we were specifically interested in decoupling the uncertainty information contained in the shape of the likelihood function from the decision while minimally disrupting the trial-by-trial correlation between the point estimate of the stimulus orientation and the subject's classification decision. To achieve this, for each trial, the newly assigned likelihood function was shifted such that the mean of the normalized likelihood function, μ_L (Eq. 22), remained the same for each trial before and after the likelihood shuffling (Fig. 5c). This allowed us to specifically assess the effect of distorting the shape of the likelihood function conditioned on both the (binned) stimulus orientation and the point estimate of the stimulus orientation (i.e. μ_L) (Fig. 5c). We then trained both the Full-Likelihood Model and the Fixed-Uncertainty Model on the shuffled data, following the exact procedure used when training on the original (unshuffled) data.

Classification simulation We computed the expected effect size of the model fit difference between the Full-Likelihood Model and the Fixed-Uncertainty Model by generating simulated data using the trained Full-Likelihood Model as the ground truth. Specifically, for each trial for each contrast-session, we computed the probability of responding $\hat{C} = 1$ from Eq. 32, utilizing the full decoded likelihood function $\mathcal{L}(\theta)$ for the given trial, and sampled a classification decision from that probability. This procedure yielded simulated data where all monkeys' decisions were replaced by decisions made by the trained Full-Likelihood Models. We repeated this procedure 5 times, thereby producing 5 sets of simulated data. For each set of simulated data, we trained the two decision-making models (Full-Likelihood Model and Fixed-Uncertainty Model) on each contrast-session with 10-fold cross-validation, and reported the aggregated test set scores as was done for the original data.

Data availability All figures except for Figure 1 were generated from raw data or processed data. The data generated and/or analyzed during the current study are available from the corresponding author upon reasonable request. No publicly available data was used in this study.

Code availability Code used for modeling and training the deep neural networks as well as for figure generation can be viewed and downloaded from `https://github.com/eywalker/v1_likelihood`. All other code used for analysis including data selection and decision model fitting can be found at `https://github.com/eywalker/v1_project`. Finally, code used for elecrophysiology data processing can be found in the Tolias lab GitHub organization `https://github.com/atlab`.

Statistics Throughout the paper, the level of significance is indicated as * for p <0.05, ** for p <0.01 and *** for p <0.005. Exact p values less than 0.001 were reported as p <0.001. All statistical tests used were two-tailed paired two-sample t-test, unless specified otherwise. Wherever reported, data are means and error bars indicate standard error of the means computed as $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation and n is the size of the sample within the bin, unless specified otherwise.

References

1. Laplace, P.-S. *Theorie Analytique des Probabilites* (Ve Courcier, Paris, 1812).
2. von Helmholtz, H. Versuch einer erweiterten Anwendung des Fechnerschen Gesetzes im farbensystem. *Z. Psychol. Physiol. Sinnesorg* **2**, 1–30 (1891).
3. Knill, D. C. & Richards, W. (eds.) *Perception As Bayesian Inference* (Cambridge University Press, New York, NY, USA, 1996).
4. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annual review of psychology* **55**, 271–304 (2004).
5. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27**, 712–719 (2004).
6. Ma, W. J. & Jazayeri, M. Neural Coding of Uncertainty and Probability. *Annual review of neuroscience* **37**, 205–220 (2014).
7. Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* **14**, 257–262 (2004).
8. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002). NIHMS150003.
9. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**, 1432–1438 (2006). NIHMS150003.

10. Beck, J. M. *et al.* Probabilistic Population Codes for Bayesian Decision Making. *Neuron* **60**, 1142–1152 (2008). 1507.01561.

11. Pouget, A., Dayan, P. & Zemel, R. Information processing with population codes. *Nature reviews. Neuroscience* **1**, 125–32 (2000).

12. Pouget, A., Dayan, P. & Zemel, R. S. Inference and Computation with Population Codes. *Annu. Rev. Neurosci* **26**, 381–410 (2003).

13. Ma, W. J., Beck, J. M. & Pouget, A. Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology* **18**, 217–222 (2008).

14. Graf, A. B. A., Kohn, A., Jazayeri, M. & Movshon, J. A. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Publishing Group* **14**, 239–245 (2011).

15. Yang, T. & Shadlen, M. N. Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080 (2007). NIHMS150003.

16. Fetsch, C. R., Pouget, A., Deangelis, G. C. & Angelaki, D. E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience* **15**, 146–154 (2012). NIHMS150003.

17. Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences* **110**, 20332–20337 (2013).

18. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

19. Goodfellow, Ian, Bengio, Yoshua, Courville, A. Deep Learning. *MIT Press* (2016). `arXiv:1312.6184v5`.

20. Seung, H. S. & Sompolinsky, H. Simple models for reading neuronal population codes. *Proc.Natl.Acad.Sci.* **90**, 10749–10753 (1993). `NIHMS150003`.

21. Sanger, T. D. Probability density estimation for the interpretation of neural population codes. *Journal of neurophysiology* **76**, 2790–3 (1996).

22. Zemel, R. S., Dayan, P. & Pouget, A. Probabalistic interpretation of population codes. *Neural Comp.* **10**, 403–430 (1998).

23. Jazayeri, M. & Movshon, J. A. Optimal representation of sensory information by neural populations. *Nature Neuroscience* **9**, 690–696 (2006). `10.1021/nl3012853|NanoLett.2012,12,36023608`.

24. Averbeck, B. B. & Lee, D. Effects of Noise Correlations on Information Encoding and Decoding. *J Neurophysiol* **95**, 3633–3644 (2006).

25. Ecker, A. S. *et al.* Decorrelated neuronal firing in coritcal micorcircuits. *Science* **327**, 584–587 (2010).

26. Ecker, A. S., Berens, P., Tolias, A. S. & Bethge, M. The Effect of Noise Correlations in Populations of Diversely Tuned Neurons. *Journal of Neuroscience* **31**, 14272–14283 (2011). `NIHMS150003`.

27. Ecker, A. S. *et al.* State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248 (2014).

28. van Bergen, R. S. & Jehee, J. F. Modeling correlated noise is necessary to decode uncertainty. *NeuroImage* (2017). 1708.04860.

29. Denfield, G. H., Ecker, A. S., Shinn, T. J., Bethge, M. & Tolias, A. S. Attentional fluctuations induce shared variability in macaque primary visual cortex. *Nature Communications* **9**, 2654 (2018).

30. Ma, W. J. Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research* **50**, 2308–2319 (2010).

31. Van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience* **18**, 1728–1730 (2015). 15334406.

32. Tolhurst, D. J., Movshon, J. A. & Dean, A. F. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* **23**, 775–785 (1983).

33. Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **18**, 3870–96 (1998).

34. Angelaki, D. E., Humphreys, G. & DeAngelis, G. C. Multisensory Integration. *Journal Of The Theoretical Humanities* **19**, 452–458 (2009).

35. Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R. & Pouget, A. Behavior and neural basis of near-optimal visual search. *Nature Neuroscience* **14**, 783–790 (2011). NIHMS150003.

36. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in Neural Circuits with Divisive Normalization. *Journal of Neuroscience* **31**, 15310–15319 (2011). NIHMS150003.

37. Ma, W. J. & Rahmati, M. Towards a Neural Implementation of Causal Inference in Cue Combination. *Multisensory Research* **26**, 159–176 (2013).

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

39. Brainard, D. H. The Psychophysics Toolbox. *Spatial vision* **10**, 433–6 (1997).

40. Tolias, A. S. *et al.* Recording Chronically From the Same Neurons in Awake, Behaving Primates. *Journal of Neurophysiology* **98**, 3780–3790 (2007).

41. Subramaniyan, M., Ecker, A. S., Berens, P. & Tolias, A. S. Macaque Monkeys Perceive the Flash Lag Illusion. *PLoS ONE* **8**, e58788 (2013).

42. Quiroga, R. Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering. *Neural Computation* **16**, 1661–1687 (2004).

43. Kohn, A. & Movshon, J. A. Adaptation changes the direction tuning of macaque MT neurons. *Nature Neuroscience* **7**, 764–772 (2004).

44. Richard, M. D. & Lippmann, R. P. Neural Network Classifiers Estimate Bayesian a posteriori Probabilities. *Neural Computation* **3**, 461–483 (1991).

45. Kline, D. M. & Berardi, V. L. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing and Applications* **14**, 310–318 (2005).

46. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86 (1951). 1511.00860.

47. MacKay, D. J. C. *Information Theory , Inference , and Learning Algorithms*, vol. 22 (Cambridge University Press, Cambridge, UK, 2003). arXiv:1011.1669v3.

48. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).

49. Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, 55–69 (Springer-Verlag, London, UK, UK, 1998).

50. Mamassian, P. & Landy, M. S. Observer biases in the 3D interpretation of line drawings. *Vision Research* **38**, 2817–2832 (1998).

51. Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the Origins of Suboptimality in Human Probabilistic Inference. *PLoS Computational Biology* **10**, e1003661 (2014).

52. Acerbi, L. & Ma, W. J. Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In *Advances in Neural Information Processing Systems 30*, 1836–1846 (2017). 1705.04405.

Acknowledgements The research was supported by National Science Foundation Grant IIS-1132009 (to W.J.M. and A.S.T.), DP1 EY023176 Pioneer Grant (to A.S.T.), F30 EY025510 (to. E.Y.W.) and R01 EY026927 (to A.S.T and W.J.M.). We thank Fabian Sinz for helpful discussion and suggestions on the deep neural network fitting to likelihood functions. We also thank Tori Shin for assistance in monkeys behavioral training and experimental data collection.

Contributions All authors designed the experiments and developed the theoretical framework. R.J.C. trained the first monkey, and R.J.C. and E.Y.W. recorded data from this monkey. E.Y.W. trained and recorded from the second monkey. E.Y.W. performed all data analyses. E.Y.W. wrote the manuscript, with contributions from all authors.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to
A.S.T. (astolias@bcm.edu).