nature neuroscience

感觉不确定性由编码在神经群体活动的概率分布表征；
使用fMRI结合生成模型分析我们发现，反应感觉不确定性的概率分布可以可靠的从人类视觉皮层估计出来。此外，被试在进行知觉决策时也常使用这种不确定性。

Sensory uncertainty decoded from visual cortex predicts behavior

Ruben S van Bergen[1], Wei Ji Ma[2], Michael S Pratte[3] & Janneke F M Jehee[1]

Bayesian theories of neural coding propose that sensory uncertainty is represented by a probability distribution encoded in neural population activity, but direct neural evidence supporting this hypothesis is currently lacking. Using fMRI in combination with a generative model-based analysis, we found that probability distributions reflecting sensory uncertainty could reliably be estimated from human visual cortex and, moreover, that observers appeared to use knowledge of this uncertainty in their perceptual decisions.

The information that the brain receives from the senses is typically consistent with a range of possible stimulus values; consequently, all of our perceptual decisions have to be made under uncertainty. It is well known that this sensory uncertainty can affect behavior[1,2], but how the fidelity of sensory knowledge is represented in cortex remains unclear. Bayesian theories of neural coding postulate that a probability distribution over sensory stimuli is encoded in the activity of a whole population of neurons, with the width of this distribution reflecting the degree of uncertainty about the stimulus[3–6]. Although indirect neurophysiological evidence is consistent with this notion[7,8], this hypothesis has yet to be tested directly in sensory cortex. We used functional magnetic resonance imaging (fMRI) in combination with a model-based analysis to address two fundamental questions. First, can a probability distribution that reflects sensory uncertainty be extracted from population activity in human visual cortex? Second, do observers use knowledge of this uncertainty in their perceptual decisions? Notably, unlike previous behavioral studies on probabilistically optimal inference, no external noise was added to the visual stimuli to manipulate uncertainty, as changes in physical stimulus properties could then act as external cues to reliability[9]. We hypothesized that varying degrees of internal neural noise might nonetheless affect sensory processing on a trial-by-trial basis; if uncertainty is part of the neural sensory code, then these fluctuations in uncertainty should alter human perceptual decision-making.

Participants viewed annular gratings of random orientations while we measured activity in visual cortex using fMRI. Shortly after the presentation of each grating, observers reported its orientation by rotating a bar presented at fixation. Observers generally performed well on this task. The mean angular difference between reported and actual orientations was 4.5°, ranging between 0.003° and 37.68° on individual trials. These fluctuations in behavioral accuracy are often thought to arise, in part, from internal neural noise affecting the fidelity of cortical orientation representations. We asked whether this trial-by-trial variability in the fidelity of internal knowledge was reflected in fMRI activation patterns. We addressed this question using a model-based decoding approach to analyzing fMRI data (Online Methods). Specifically, using a generative model incorporating the orientation preferences of voxels as well as their (correlated) noise, we approximated on each trial the posterior probability distribution over orientation, given the pattern of blood oxygen level–dependent (BOLD) activity. This approach differs from conventional fMRI decoding studies[10–12] in that it explicitly recovers a full probability distribution over stimulus values, rather than a single stimulus estimate. The (circular) mean of the posterior distribution serves as an estimate of the presented orientation, and its width (circular s.d.) as a measure of the degree of stimulus uncertainty in the cortical representation. Regions of interest were those portions of visual areas V1–V3 that corresponded to the retinotopic representation of the stimuli.

我们问，内部知识保真度的这种反复试验的变化是否反映在功能磁共振成像的激活模式中。
我们使用基于模型的解码方法来分析fMRI数据来解决这个问题

To benchmark our approach, we first tested its ability to identify the presented orientation from activity patterns in areas V1–V3 (**Supplementary Fig. 1a**,**b**). The decoded and presented orientations were significantly correlated ($r = 0.69$, $P \approx 0$), consistent with previous findings[13,14]. We then turned to the degree of trial-by-trial uncertainty about orientation. The decoder's estimates of uncertainty varied from trial to trial as a result of noise in the fMRI measurements (**Fig. 1a**). To the extent that our decoding approach appropriately models the fMRI data, uncertainty on a single trial should be related to variability across trials. Accordingly, to verify the decoder's assumptions and test whether the decoded uncertainty followed the actual degree of noise in the fMRI data, we binned the data according to posterior width, and calculated the trial-by-trial variability in decoded orientation estimates for each of the bins (**Supplementary Fig. 1c**). We found that the decoded uncertainty was indeed significantly correlated with the across-trial variability in decoded orientations ($r = 0.91$, $P \approx 0$). A comparison between different noise models revealed that the one used here best captured the fMRI data (**Supplementary Fig. 2**). Altogether, these findings corroborate the validity of our assumptions and suggest that posterior width captures the overall degree of uncertainty in the data on a trial-by-trial basis.

Having established that the decoded uncertainty reflects the aggregate of all fMRI noise sources, we next asked whether it captures variability in cortical stimulus representations in particular. We addressed this question in two sets of analyses, focusing first on gratings of different orientations. Consistent with previous work[15,16], behavioral orientation judgments were more accurate for cardinal than for oblique orientation stimuli (correlation between behavioral variability and the angle of the presented stimulus with the nearest cardinal

神经编码的贝叶斯理论假设，感觉刺激的概率分布被编码在整个神经元群的活动中，分布的宽度反映了刺激不确定性的程度。

[1]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, the Netherlands. [2]Center for Neural Science and Department of Psychology, New York University, New York, New York, USA. [3]Department of Psychology, Mississippi State University, Mississippi State, Mississippi, USA. Correspondence should be addressed to J.F.M.J. (janneke.jehee@donders.ru.nl).

Received 21 August; accepted 1 October; published online 26 October 2015; doi:10.1038/nn.4150

怎么测量知觉决策中是否使用了知觉的不确定性？不确定性的大小影响决策？

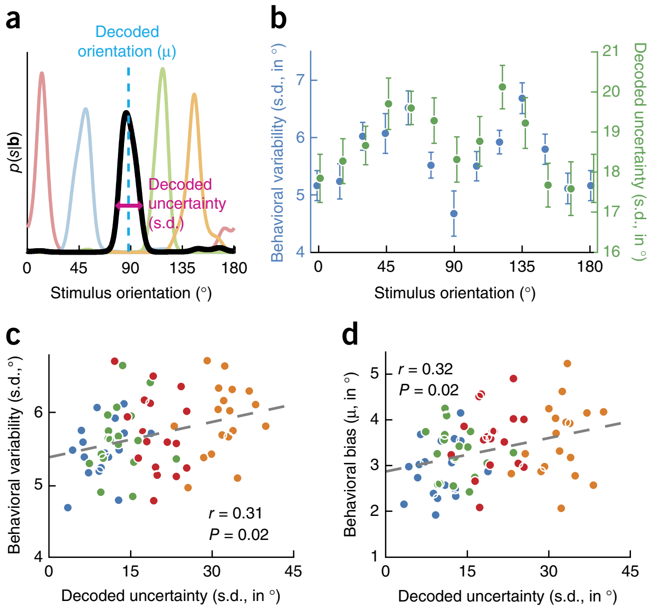

Figure 1 Stimulus uncertainty decoded from human visual cortex correlates with behavior. (**a**) Examples of probability distributions over stimulus orientation, decoded from fMRI activation patterns obtained from areas V1–V3, and five different trials (trials indicated by different colors; s, stimulus orientation; **b**, BOLD activity). (**b**) Both behavioral variability and decoded uncertainty increased for more oblique stimulus orientations in areas V1–V3 (behavioral variability, $t(53) = 4.71$, $P = 1.8 \times 10^{-5}$; decoded uncertainty, $t(53) = 2.75$, $P = 0.008$). For illustrative purposes, trials were sorted into 12 equally spaced orientation bins. (**c**) Behavioral variability was significantly correlated with posterior width in areas V1–V3 ($t(52) = 2.39$, $P = 0.021$). (**d**) Behavioral biases increased with increasing width of the posterior distribution in areas V1–V3 ($t(52) = 2.46$, $P = 0.017$). In **c** and **d**, colors indicate four within-observer bins of increasing decoded uncertainty (for each of 18 observers). In all plots, error bars represent ±1 s.e.m., dashed lines represent best linear fits and reported r values are partial correlation coefficients obtained from a linear regression analysis (Online Methods). Similar results were found for visual areas V1–V3 (**Supplementary Fig. 1**).

axis: $r = 0.54$, $P = 1.8 \times 10^{-5}$; **Fig. 1b**). Because behavioral accuracy is directly related to uncertainty, we tested whether this oblique effect in behavior was paralleled in visual cortical responses. Indeed, the width of the decoded posterior probability distribution was narrower for stimulus orientations closer to the cardinal axes ($r = 0.35$, $P = 0.008$; **Fig. 1b**). Thus, similar to behavior, horizontal and vertical orientations are represented with greater precision in visual cortex.

For the second set of analyses, we focused on repeated presentations of physically identical orientation stimuli. We reasoned that, if the posterior distribution also captures random, trial-by-trial fluctuations in cortical activity, then more certain decoder estimates should be linked to reduced variability in the observer's behavior, even for physically identical orientation stimuli[6]. To test this relationship, we divided each participant's data into bins of increasing decoded uncertainty, calculated the across-trial variability in participant behavior for each of the bins and computed the partial correlation coefficient between the two (while controlling for stimulus orientation and between-subject variability, Online Methods). This revealed that observers clearly made more accurate decisions when the information decoded from their visual activity was more precise ($r = 0.31$, $P = 0.021$; **Fig. 1c**). By contrast, neither the error in decoded orientation (**Supplementary Fig. 3**) nor the amplitude or width of the estimated neural population response (**Supplementary Fig. 4**) reliably predicted behavioral variability, demonstrating the power of the posterior distribution in capturing trial-by-trial fluctuations in cortical processing. Control analyses established that these results could not be accounted for by differences in eye movements, gross BOLD amplitude or subject head motion (**Supplementary Fig. 5**), and were specific to voxels tuned to the retinotopic location of the stimulus (that is, we found no reliable correlation between decoded uncertainty and behavioral variability when selecting V1–V3 voxels preferring other retinotopic locations, $P = 0.17$). Taken together, these results demonstrate that the fidelity of a sensory representation can reliably be extracted from fMRI activation patterns.

Armed with the ability to estimate uncertainty in sensory representations, the critical question is whether observers take this uncertainty into account when making perceptual decisions. If so, then this would provide strong empirical support for probabilistic models of perception[6]. To address this question, we relied on a well-established behavioral finding[17,18] that we replicated here: orientation judgments were biased away from the cardinal axes (**Supplementary Fig. 6**). Although the precise neural mechanisms underlying such repulsive biases have yet to be determined, all theoretical models[18–20] of these biases generate the same prediction: if the observer takes into account the uncertainty in perceptual representations, the amount of behavioral bias should depend on the degree of sensory uncertainty. Indeed, we found that behavioral biases increased when the decoded fidelity of cortical information about the visual stimulus was low ($r = 0.32$, $P = 0.017$; **Fig. 1d**), consistent with a recent theoretical prediction that the repulsive bias should increase with increasing levels of internal uncertainty[20]. Given that the physical stimulus was held constant, this suggests that human observers use knowledge of internal uncertainty in their perceptual decision-making and, moreover, that they monitor fluctuations therein on a trial-by-trial basis.

A major limitation of previous work on probabilistically optimal inference has been the use of external sources of noise, leaving open the possibility that observers simply monitor certain image properties, such as blurriness or contrast, as external cues to uncertainty. For this reason, we held physical stimulus properties constant and relied on fluctuations in internal noise to make perceptual information more or less reliable to the observer. We found that the uncertainty in perceptual representations could reliably be extracted as a probability distribution from human visual cortex with fMRI, and, moreover, that human observers appeared to rely on the uncertainty in this internal evidence when making perceptual decisions. These results suggest that neural activity encodes probability distributions over stimulus values, rather than merely point estimates, and that the brain uses this probabilistic information in its computations.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Ling for valuable comments and discussion, K. Ambroziak for help with data collection, C. Beckmann for advice on statistical analyses, and P. Gaalman for MRI support.

AUTHOR CONTRIBUTIONS

R.S.v.B., W.J.M. and J.F.M.J. conceived and designed the experiments. R.S.v.B. collected data. R.S.v.B. analyzed data, with help from W.J.M., M.S.P. and J.F.M.J. R.S.v.B., W.J.M., M.S.P. and J.F.M.J. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ernst, M.O. & Banks, M.S. *Nature* **415**, 429–433 (2002).
2. Körding, K.P. & Wolpert, D.M. *Nature* **427**, 244–247 (2004).
3. Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. *Nat. Neurosci.* **9**, 1432–1438 (2006).
4. Jazayeri, M. & Movshon, J.A. *Nat. Neurosci.* **9**, 690–696 (2006).
5. Vilares, I. & Kording, K. *Ann. N Y Acad. Sci.* **1224**, 22–39 (2011).
6. Pouget, A. *Nat. Neurosci.* **16**, 1170–1178 (2013).
7. Fetsch, C.R., Pouget, A., DeAngelis, G.C. & Angelaki, D.E. *Nat. Neurosci.* **15**, 146–154 (2012).
8. Beck, J.M. *et al. Neuron* **60**, 1142–1152 (2008).
9. Barthelmé, S. & Mamassian, P. *Proc. Natl. Acad. Sci. USA* **107**, 20834–20839 (2010).
10. Kamitani, Y. & Tong, F. *Nat. Neurosci.* **8**, 679–685 (2005).
11. Haynes, J.D. & Rees, G. *Nat. Neurosci.* **8**, 686–691 (2005).
12. Jehee, J.F.M., Brady, D.K. & Tong, F. *J. Neurosci.* **31**, 8210–8219 (2011).
13. Brouwer, G.J. & Heeger, D.J. *J. Neurophysiol.* **106**, 2108–2119 (2011).
14. Serences, J.T., Saproo, S., Scolari, M., Ho, T. & Muftuler, L.T. *Neuroimage* **44**, 223–231 (2009).
15. Appelle, S. *Psychol. Bull.* **78**, 266–278 (1972).
16. Girshick, A.R., Landy, M.S. & Simoncelli, E.P. *Nat. Neurosci.* **14**, 926–932 (2011).
17. Huttenlocher, J., Hedges, L.V. & Duncan, S. *Psychol. Rev.* **98**, 352–376 (1991).
18. Jazayeri, M. & Movshon, J.A. *Nature* **446**, 912–915 (2007).
19. Stocker, A.A. & Simoncelli, E.P. *Adv. Neural Inf. Process. Syst.* **20**, 1409–1416 (2008).
20. Wei, X.-X. & Stocker, A.A. *Nat. Neurosci.* **18**, 1509–1517 (2015).

ONLINE METHODS

Participants. 18 healthy adult volunteers (aged 22–31, seven female), with normal or corrected-to-normal vision, participated in this study. All participants provided informed written consent. The study was approved by the Radboud University Institutional Review Board. Assuming effect sizes similar to those reported here, a power analysis indicated that 18 subjects would produce a power level of 0.76–0.86. Indeed, this sample size is consistent with previous decoding studies[21,22].

Data acquisition. MRI data were acquired using a Siemens 3T Magnetom Trio scanner with an eight-channel occipital coil located at the Donders Center for Cognitive Neuroimaging. For each participant, a high-resolution T1-weighted magnetization-prepared rapid gradient echo anatomical scan (MPRAGE, FOV 256 × 256, 1-mm isotropic voxels) was collected at the start of the session. Functional imaging data were acquired using T2*-weighted gradient-echo echoplanar imaging, in 30 slices oriented perpendicular to the calcarine sulcus, covering all of the occipital and part of posterior parietal and temporal cortex (TR 2,000 ms, TE 30 ms, flip angle 90°, FOV 64 × 64, slice thickness 2.2 mm, in-plane resolution 2.2 × 2.2 mm).

Experimental design and stimuli. Stimuli were generated by a Macbook Pro computer running Matlab and the Psychophysics Toolbox[23,24], and displayed on a rear-projection screen using a luminance-calibrated EIKI projector (resolution 1,024 × 768 pixels, refresh rate 60 Hz). Observers viewed the visual display through a mirror mounted on the head coil.

Participants were required to maintain fixation on a central bull's eye target (radius: 0.25°) throughout each experimental run. Each run consisted of an initial fixation period (4 s), followed by 18 stimulus trials (12 s) and a final fixation period (4 s). Trials were separated by a 4-s inter-trial interval. Each trial started with the presentation of an orientation stimulus (1.5 s). Orientation stimuli were counterphasing sinusoidal gratings (contrast: 10%, spatial frequency: 1 cycle per °, randomized spatial phase, 2-Hz sinusoidal contrast modulation) presented in an annulus surrounding fixation (inner radius: 1.5°, outer radius: 7.5°, grating contrast decreased linearly to 0 over the outer and inner 0.5° radius of the annulus). The orientation of the stimulus was determined pseudo-randomly (from 0–179°) to ensure an approximately even sampling of orientations in each run. The grating was followed by a fixation interval (6.5 s), and then a response period (4 s) in which a black line (length: 2.8°, width: 0.1°) appeared at the center of the screen at an initially random orientation. The line disappeared gradually over the last 1 s of the response period to indicate the approaching end of this window. Subjects reported the orientation of the grating by rotating the line using separate buttons for clockwise or counterclockwise rotation on an MRI-compatible button box.

Participants completed 10–18 stimulus runs. Each scan session also included two visual localizer runs, in which flickering checkerboard patterns were presented within the same aperture as the gratings (check size: 0.5°, display rate: 10 Hz, contrast: 100%). Checkerboards were presented in 12-s blocks, interleaved with fixation blocks of equal duration.

Retinotopic maps of visual cortex were acquired in a separate scan session using conventional retinotopic mapping procedures[25–27].

fMRI data preprocessing and regions of interest. Functional images were motion corrected using FSL's MCFLIRT[28] and passed through a high-pass temporal filter with a cut-off period of 40 s to remove slow drifts in the BOLD signal. Residual motion-induced fluctuations in the BOLD signal were removed through linear regression, based on the alignment parameters generated by MCFLIRT. Functional volumes were aligned to a previously collected anatomical reference scan using FreeSurfer[29].

Regions of interest (ROIs; V1, V2 and V3) were defined on the reconstructed cortical surface using standard procedures[25–27]. In each area, we selected all voxels that responded to the localizer stimulus at a lenient threshold ($P < 0.05$ uncorrected) for subsequent analysis, in the native space for each participant. Control analyses verified that our results were not strongly affected by the number of voxels selected for analysis (**Supplementary Fig.** 7). In one of our control analyses, we tested the degree to which the decoded uncertainty was specific to the retinotopic location of the stimulus. For this analysis, we selected those voxels in areas V1–V3 combined that were not significantly activated by the localizer stimulus (at a statistical threshold of $P > 0.2$), obtaining on average 1,257 voxels per subject (by comparison, the average number of voxels in individual ROIs was 660; **Supplementary Fig. 1**).

The time series of each voxel was z-normalized using the corresponding time points of all trials in a given run. Activation patterns for each trial were defined by averaging together the first 4 s of each trial, after adding a 4-s temporal shift to account for hemodynamic delay. This relatively short time window (4 s) was chosen in order to ensure that activity from the response window was excluded from analysis. Control analyses verified that this time window was close to the peak of the hemodynamic response function (time series for this analysis were normalized to percentage signal change units, defined relative to the average activation level across each run; **Supplementary Fig. 8**). In addition, temporally expanding the time window to include an earlier time point did not greatly affect any of our results. For the control analyses of **Supplementary Figure 5**, mean BOLD intensity values were obtained by averaging across all voxels in a given ROI. Subject head motion was calculated as the Euclidian norm of the temporal derivatives of the realignment parameters generated by the motion correction algorithm – a quantity that reflects the total amount of head motion per time step. We averaged across TRs 3 and 4 of each trial, similar to our decoding analyses.

Decoding analysis. *The generative model.* Our decoding approach started with the assumption that voxels in early visual cortex are selective to orientation[10,11,13,14,30,31], and that voxel activity varies across trials due to (correlated) noise[32,33]. More specifically, we assumed that the BOLD response of voxel i to orientation s can be characterized as a linearly weighted combination of the idealized tuning functions $\boldsymbol{f}(s)$ of K neural populations[13,14] ($K = 8$), each tuned to a different orientation, combined with Gaussian noise

$$b_i = \sum_k^K W_{ik}\left(f_k(s)+\eta_k\right)+\nu_i \tag{1}$$

Here, $f_k(s)$ is the mean response (or tuning curve) of the kth population as a function of stimulus orientation s, W_{ik} is the contribution of population k to the response of voxel i, and both η_k and ν_i model sources of noise. The latent (unobserved) variable $\boldsymbol{\eta}$ is normally distributed as $\boldsymbol{\eta} \sim \mathcal{N}\left(0,\sigma^2\mathbf{I}\right)$, and specifies noise that is shared among neural populations of similar orientation preference (which is why it is weighted by $\mathbf{W}$). It models global changes in the population response due to, for example, shifts in response gain. The variable $\boldsymbol{\nu}$ was included to capture various remaining sources of noise, such as variability in the fMRI signal due to thermal, electrical or physiological noise, as well as residual sources of neural noise. The distribution of $\boldsymbol{\nu}$ is given by

$$\boldsymbol{\nu} \sim \mathcal{N}\left(0,\boldsymbol{\Sigma}\right) \tag{2}$$

$$\boldsymbol{\Sigma} = \rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + (1-\rho)\mathbf{I}\circ\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} \tag{3}$$

Thus, $\boldsymbol{\nu}$ describes noise specific to individual voxels (with variance τ_i^2 for voxel i), as well as noise shared globally among voxels irrespective of their tuning properties (scaled by ρ). Population tuning curves (or basis functions) $\boldsymbol{f}(s)$ are half wave-rectified cosine functions, raised to the fifth power[13]

$$f_k(s) = \max\left(0,\cos\left(\pi\frac{s-\varphi_k}{90}\right)\right)^5 \tag{4}$$

where φ_k is the preferred orientation of the kth population. Preferred orientations are equally spaced between 0–180°, with one basis function maximally tuned toward horizontal.

The conditional probability of a voxel activation pattern given a noisy population response ($\boldsymbol{f}(s) + \boldsymbol{\eta}$) is then defined as

$$p\left(\mathbf{b}|s,\boldsymbol{\eta};\mathbf{W},\boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}\left(\mathbf{b}-\mathbf{W}\left(\boldsymbol{f}(s)+\boldsymbol{\eta}\right)\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{b}-\mathbf{W}\left(\boldsymbol{f}(s)+\boldsymbol{\eta}\right)\right)\right) \tag{5}$$

Marginalizing over $\boldsymbol{\eta}$ results in

$$p\left(\mathbf{b}|s;\mathbf{W},\boldsymbol{\Omega}\right)=\int p\left(\mathbf{b}|\boldsymbol{\eta},s;\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\eta}\right)d\boldsymbol{\eta}=\frac{1}{\sqrt{2\pi|\boldsymbol{\Omega}|}}\exp\left(-\frac{1}{2}\left(\mathbf{b}-\mathbf{W}\boldsymbol{f}(s)\right)^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\left(\mathbf{b}-\mathbf{W}\boldsymbol{f}(s)\right)\right) \tag{6}$$

$$\boldsymbol{\Omega}=\rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}}+\left(1-\rho\right)\mathbf{I}\circ\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}}+\sigma^{2}\mathbf{W}\mathbf{W}^{\mathrm{T}} \tag{7}$$

When developing the model, we considered four different, increasingly complex models. The current model was found to best capture the data on two relevant benchmark tests (**Supplementary Fig. 2**). **Supplementary Table 1** provides an overview of all parameters.

Training and testing. Model parameters were estimated using the fMRI activation patterns for the orientation stimuli in a leave-one-run-out cross-validation procedure. Data were divided into a training data set (consisting of data from all but one fMRI run) and a testing data set (consisting of data from the remaining run). The average training data set consisted of 249 trials and 1,981 voxels. When training the model, we maximized the likelihood of the model parameters given the orientation stimuli. Model parameters were estimated in a two-step training procedure to constrain the number of free parameters and prevent overfitting of the covariance matrix. In the first step of this estimation procedure, we imposed a regularizing prior on the covariance matrix, assuming that $\sigma = 0$. This reduces the number of free parameters in the covariance matrix considerably (from $((K+1)M + 2)$ to $(M + 2)$), but results in a predictable underestimation of σ in the second step of the estimation procedure (see below), when we relaxed the prior on σ. However, simulations indicated that this did not seriously affect our ability to reconstruct the uncertainty present in the activation patterns (see **Supplementary Fig. 9**). Under the assumption that $\sigma = 0$, the maximum likelihood estimation of $\mathbf{W}$ is simplified to an ordinary least-squares (OLS) regression.

$$\hat{\mathrm{W}}_i=\mathrm{b}_i\boldsymbol{f}\left(s\right)^{\mathrm{T}}\left(\boldsymbol{f}\left(s\right)\boldsymbol{f}\left(s\right)^{\mathrm{T}}\right)^{-1} \tag{8}$$

In the second step of the estimation procedure, we relaxed the regularizing assumption on σ, and maximized the likelihood of the remaining parameters (ρ, $\boldsymbol{\tau}$ and σ), conditioned on $\hat{\mathbf{W}}$. This maximization step was performed numerically, using a Matlab implementation of the conjugate gradient method. The resulting parameter estimates are summarized in **Supplementary Figure 10**.

After fitting the model to the training data set, we tested the model on the held-out (independent) testing data set. By applying Bayes rule and using a flat stimulus prior (see below), we obtained for each test trial the posterior probability distribution over stimulus orientation given the fitted model parameters

$$p\left(s|\mathbf{b};\hat{\boldsymbol{\theta}}\right)=\frac{p\left(\mathbf{b}|s;\hat{\boldsymbol{\theta}}\right)p\left(s\right)}{\int p\left(\mathbf{b}|s;\hat{\boldsymbol{\theta}}\right)p\left(s\right)ds} \tag{9}$$

where $\hat{\boldsymbol{\theta}}=\{\hat{\mathrm{W}},\hat{\rho},\hat{\sigma},\hat{\boldsymbol{\tau}}\}$, and the normalization constant $\int p\left(\mathbf{b}|s;\hat{\boldsymbol{\theta}}\right)p\left(s\right)ds$ was computed numerically. The circular mean of the posterior served as an estimate of the presented stimulus orientation, and its circular s.d. as a measure of the degree of uncertainty in the orientation estimate. Both of these summary statistics were computed using numerical integration. The cross-validation procedure was repeated until each run served as a test run once, resulting in a single posterior probability distribution for each trial of BOLD activity.

The prior. Unlike our decoding algorithm, human observers appear to employ in their behavior a more naturalistic prior favoring cardinal orientations[16]. Would it have been more appropriate to apply such a non-uniform prior in our analyses? The answer to this question is negative. From a decoding perspective, a uniform prior correctly represents the a priori knowledge that all orientations were presented equally often in the current experiment. Had we instead used a prior favoring cardinal orientations, this would have (wrongfully) biased the decoder's estimates toward horizontal and vertical orientations, and automatically changed the width of the posterior distribution for cardinal and oblique orientations, even when there was no evidence for an oblique effect in brain activity.

Behavioral data. The observer's behavioral error on a given trial was computed as the acute-angle difference between the reported orientation and the presented orientation. Behavioral variability was calculated after correcting for an orientation-dependent shift in mean (**Supplementary Fig. 6**), by fitting a 4-degree polynomial to each observer's behavioral errors as a function of stimulus orientation. We then used the residuals from this fit, i.e. the bias-corrected behavioral errors, in our calculation of behavioral variability.

In general, participants finished adjusting the probe's orientation well before the end of the response window, taking on average 2,254 ± 47 ms (mean ± s.e.m.) to respond. On 0–6 trials (out of 180–324), the behavioral error exceeded the mean error (for that observer) by more than 3 s.d., suggesting that the participant randomly guessed the orientation of the grating. These trials were excluded from further analysis.

Eye tracking. Eye-tracking data were acquired for 12 of 18 subjects, using an SR Research Eyelink 1000 eye-tracking system. Gaze position was sampled at 1 kHz. After removing blinks, we applied a band-pass temporal filter with a low-frequency cut-off period of 100 s, and a high-frequency cut-off period of 0.1 s. Similar to all fMRI analyses, the mean and s.d. of eye position were computed for the time window corresponding to stimulus presentation. Specifically, with respect to mean eye position, we first calculated the average x- and y-coordinates of the gaze data during the first 4 s of each trial, and then took the absolute distance from this average (x, y) position to the central fixation target. Eye movement (that is, variability in eye position) was quantified by first calculating, for each sample of (x, y) gaze coordinates, the absolute distance to the mean (x, y) eye position. This distance was then averaged across the first 4 s of each trial.

Statistical procedures. Decoding accuracy was computed by taking, for each individual observer, the circular correlation coefficient between the presented and decoded orientations. This correlation coefficient was subsequently Fisher transformed and standardized to a Z-score across observers. Its (two-sided) P value was derived from the inverse cumulative normal distribution, and the average of the Fisher transformed values was converted back to the correlation scale for reporting.

Many of our analyses focused on the relationship between decoded uncertainty and across-trial error distributions. For these analyses, data were binned and summary statistics were computed for each bin. When testing for an oblique effect in BOLD activity, trials were sorted into four equally spaced bins of increasing angle between the stimulus orientation and the nearest cardinal axis (for illustrative reasons, **Fig. 1b** shows the data sorted into 12 equally spaced orientation bins). Behavioral variability was computed as the circular s.d. of all (bias-corrected) behavioral errors in each of the bins. When testing for the link between decoded uncertainty and behavioral variability or bias, trials were sorted into four bins of increasing decoded uncertainty. Behavioral variability was computed as the circular s.d. of all (bias-corrected) behavioral errors in each of the bins, while behavioral bias was calculated as the mean of all (biased) behavioral errors in the bins. We used the same number of bins for each participant (four), with a constant number of samples across bins within each participant. To test whether decoded uncertainty was correlated with the variable of interest, we calculated partial correlation coefficients via a multiple linear regression analysis. Independent variables were distance to cardinal axis (**Fig. 1b–d** and **Supplementary Fig. 1d–f**) and decoded uncertainty (**Fig. 1c**,**d** and **Supplementary Fig. 1c**,**e–f**). All regression analyses furthermore included subject-specific intercepts to remove between-subject variability. Dependent variables were decoded uncertainty (**Fig. 1b** and **Supplementary Fig. 1d**), behavioral variability (**Fig. 1b**,**c** and **Supplementary Fig. 1e**), and behavioral bias (**Fig. 1d** and **Supplementary Fig. 1f**). The significance of individual regression coefficients was assessed with a (two-sided) t test. We verified that the residuals of all regression analyses were independent (Durbin-Watson test) and normally distributed (Anderson-Darling test), and that our results were not affected by violations of homoscedasticity. For ease of exposition, regression coefficients were standardized to partial correlation coefficients, and in **Figure 1c**,**d**, data are displayed in the form of partial residual plots, showing the partial relationships of interest and residuals obtained from the regressions (but omitting any effects of the variables-of-no-interest, as well as between-subject differences in intercept).

Code availability. Custom code written in Matlab is available on request.

A **Supplementary Methods Checklist** is available.

21. Kok, P., Jehee, J.F.M. & De Lange, F.P. *Neuron* **75**, 265–270 (2012).

22. Ester, E.F., Anderson, D.E., Serences, J.T. & Awh, E. *J. Cogn. Neurosci.* **25**, 754–761 (2013).
23. Brainard, D.H. *Spat. Vis.* **10**, 433–436 (1997).
24. Pelli, D.G. *Spat. Vis.* **10**, 437–442 (1997).
25. Sereno, M.I. *et al. Science* **268**, 889–893 (1995).
26. DeYoe, E.A. *et al. Proc. Natl. Acad. Sci. USA* **93**, 2382–2386 (1996).
27. Engel, S.A., Glover, G.H. & Wandell, B.A. *Cereb. Cortex* **7**, 181–192 (1997).
28. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. *Neuroimage* **17**, 825–841 (2002).
29. Fischl, B., Sereno, M.I. & Dale, A.M. *Neuroimage* **9**, 195–207 (1999).
30. Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. *Nature* **452**, 352–355 (2008).
31. Jehee, J.F.M., Ling, S., Swisher, J.D., van Bergen, R.S. & Tong, F. *J. Neurosci.* **32**, 16747–16753 (2012).
32. Smith, M.A. & Kohn, A. *J. Neurosci.* **28**, 12591–12603 (2008).
33. Goris, R.L.T., Movshon, J.A. & Simoncelli, E.P. *Nat. Neurosci.* **17**, 858–865 (2014).