Behavioral/Cognitive

Neural Evidence for Boundary Updating as the Source of the Repulsive Bias in Classification

Heeseung Lee, Hyang-Jung Lee, Kyoung Whan Choe, and Sang-Hun Lee
Department of Brain and Cognitive Sciences, Seoul National University, Seoul 08826, Republic of Korea

Binary classification, an act of sorting items into two classes by setting a boundary, is biased by recent history. One common form of such bias is repulsive bias, a tendency to sort an item into the class opposite to its preceding items. Sensory-adaptation and boundary-updating are considered as two contending sources of the repulsive bias, yet no neural support has been provided for either source. Here, we explored human brains of both men and women, using functional magnetic resonance imaging (fMRI), to find such support by relating the brain signals of sensory-adaptation and boundary-updating to human classification behavior. We found that the stimulus-encoding signal in the early visual cortex adapted to previous stimuli, yet its adaptation-related changes were dissociated from current choices. Contrastingly, the boundary-representing signals in the inferior-parietal and superior-temporal cortices shifted to previous stimuli and covaried with current choices. Our exploration points to boundary-updating, rather than sensory-adaptation, as the origin of the repulsive bias in binary classification.

Key words: **Bayesian inference; history effect; perceptual decision; repulsive bias; sensory adaptation**

Significance Statement

Many animal and human studies on perceptual decision-making have reported an intriguing history effect called "repulsive bias," a tendency to classify an item as the opposite class of its previous item. Regarding the origin of repulsive bias, two contending ideas have been proposed: "bias in stimulus representation because of sensory adaptation" versus "bias in class-boundary setting because of belief updating." By conducting model-based neuroimaging experiments, we verified their predictions about which brain signal should contribute to the trial-to-trial variability in choice behavior. We found that the brain signal of class boundary, but not stimulus representation, contributed to the choice variability associated with repulsive bias. Our study provides the first neural evidence supporting the boundary-based hypothesis of repulsive bias.

Introduction

We commit to a proposition about a specific world state when making a perceptual decision. One basic form of such commitment is binary classification. It is to decide whether an item's magnitude lies on the smaller or larger side of the magnitude distribution across items of interest (Fig. 1*A*). For example, when uttering "this tree is tall" while walking in a wood, we are implicitly judging the height of that tree to be taller than the typical height of the trees in the wood (Klein, 1980; Bierwisch, 1989), where "typical height" works as the boundary dividing the "short" and "tall" classes. Like this, binary classification is exercised in our daily language use, whenever modifying a subject with relative adjectives (Rips and Turnbull, 1980; Tribushinina, 2011; Solt, 2015; Lassiter and Goodman, 2017), and has been adopted as an essential paradigm for studying perceptual decision-making (Lages and Treisman, 1998; Grinband et al., 2006; Kepecs et al., 2008; Nahum et al., 2010; Lak et al., 2014; Bosch et al., 2020; Hachen et al., 2021).

Humans and nonhuman animals show various forms of history bias in binary classification. One frequent form of such history biases is a tendency to classify an item as the class opposite to its preceding items, dubbed repulsive bias (Lages and Treisman, 1998, 2010; Bosch et al., 2020; Hachen et al., 2021). For instance, we tend to classify a tree of intermediate height as "tall" after seeing a short tree. Currently, it remains unclear why and how repulsive bias occurs.

As one most straightforward scenario for repulsive bias, the previous stimuli may repel away our perception of the current stimulus from themselves because the sensory system adapts to earlier stimuli (Gibson and Radner, 1937; Stocker

Received Jan. 27, 2023; revised May 29, 2023; accepted June 1, 2023.

Author contributions: H.L., H.-J.L., K.W.C., and S.-H.L. designed research; H.L. and K.W.C. performed research; H.L., H.-J.L., and K.W.C. contributed unpublished reagents/analytic tools; H.L. and K.W.C. analyzed data; H.L. wrote the first draft of the paper; H.L. and S.-H.L. edited the paper; H.L. and S.-H.L. wrote the paper.

This research was supported by the Seoul National University Research Grant 339-20220013 and by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and Information and Communications Technology Grant No. NRF-2021R1F1A1052020.

The authors declare no competing financial interests.

Correspondence should be addressed to Sang-Hun Lee at visionsl@snu.ac.kr.

https://doi.org/10.1523/JNEUROSCI.0166-23.2023

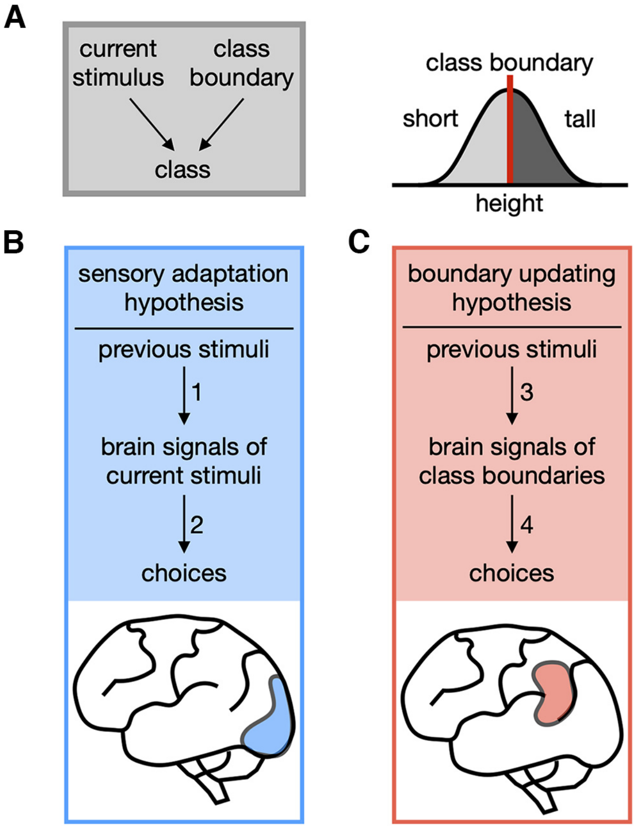
A
current stimulus
class boundary
class
class boundary
short
tall
height
B
sensory adaptation hypothesis
previous stimuli
1
brain signals of current stimuli
2
choices
C
boundary updating hypothesis
previous stimuli
3
brain signals of class boundaries
4
choices

Figure 1. Two contending hypotheses on the origin of the repulsive bias in binary classification. ***A***, Task structure (left) and statistical knowledge (right) for binary classification. For any given item, its class is determined by its position relative to the class boundary in the distribution of feature magnitudes relevant to a given task (e.g., a tree is classified as "tall" if its height is in the side greater than the typical height of the trees in the wood of interest). This relativity of binary classification makes the "biased sensory encoding" and the "biased knowledge about boundary position" because of previous stimuli, in principle, have equal footings in inducing the repulsive bias. ***B***, Sensory-adaptation hypothesis. It points to the adaptation of a low-level stimulus-encoding signal to past stimuli (arrow 1) as the origin of the repulsive bias (arrow 2). In the case of visual classification tasks, the task-relevant sensory signals in the early visual cortex (blue patch), which are subject to adaptation, have been hypothesized to mediate the repulsive bias. ***C***, Boundary-updating hypothesis. It points to the attractive shift of a classifier's internal class boundary toward previous stimuli (arrow 3) as the origin of the repulsive bias (arrow 4). Such boundary-representing signals are expected to reside not in the early sensory cortex but in the high-tier associative cortices (red patch).

and Simoncelli, 2006; Clifford et al., 2007; Knapen et al., 2010; Pavan et al., 2012; Morgan, 2014; Nakashima and Sugita, 2017; Fig. 1*B*). According to this "sensory-adaptation" hypothesis, the current tree is biasedly classified as "tall" since the sensory system's adaptation to the previous short tree makes the current tree appear taller than its physical height. However, there is an alternative scenario, which considers the possibility that the internal class boundary adaptively shifts toward recent samples of property magnitude (Treisman and Williams, 1984; Lages and Treisman, 1998, 2010; Dyjas et al., 2012; Raviv et al., 2014; Norton et al., 2017; Hachen et al., 2021; Fig. 1*C*). According to this "boundary-updating" hypothesis, the current tree is biasedly classified as "tall" since the shift of the class boundary toward the previous short tree makes the current tree be positioned in the taller side of the boundary.

As discussed previously (Hachen et al., 2021), it is hard to assess which hypothesis is more viable based on behavioral data. This difficulty arises because binary classification is a matter of the relativity between the perceived stimulus and the class boundary: the identical bias in classification can be caused either by sensory-adaptation or boundary-updating. However, the two hypotheses involve distinct neural routes through which repulsive bias transpires. The sensory-adaptation hypothesis predicts that the sensory brain signals subject to adaptation, such as those in the early sensory cortex with substantive adaptation to earlier stimuli, contribute to the choice variability. By contrast, the boundary-updating hypothesis predicts that the brain signals of the shifting boundary, such as those in the high-tier cortices involved in the working memory of previous stimuli, contribute to the choice variability.

Here, we tested these two predictions by analyzing functional magnetic resonance imaging (fMRI) data. We found that the stimulus-encoding signal in primary visual cortex (V1) exhibited adaptation, but its bias induced by adaptation was dissociated from current choices. By contrast, the boundary-representing signals in the posterior-superior-temporal gyrus and the inferior-parietal lobe not only shift to previous stimuli but also covaried with current choices. Our findings contribute to the resolution of the competing ideas regarding the source of repulsive bias by providing the first neural evidence supporting the boundary-updating scenario.

Materials and Methods

The data of experiment 1 (Exp1) and experiment 2 (Exp2) were acquired from 19 (nine females, aged 20–30 years) and 18 (nine females, aged 20–30 years) participants, respectively. Among the participants, 17 of them participated in both experiments. The Research Ethics Committee of Seoul National University approved the experimental procedures. All participants gave informed consent and were naive to the purpose of the experiments. High-spatial-resolution images were acquired only from the early visual cortex in Exp1 while the images in Exp2 were acquired from the entire brain with a conventional spatial resolution. The 17 people who provided the data for both experiments participated in three to six behavior-only sessions for training and stimulus calibration, one fMRI session for retinotopy, and two experimental fMRI sessions (one for each experiment). The remaining people also completed the behavioral and retinotopy fMRI sessions with the same protocols but participated in only one of the two experiments.

The data from Exp1 had been used for our previous work (Choe et al., 2014). The data of Exp2 has never been used in any previous publication. In the current paper, we describe some basic procedures of Exp1. For more details on Exp1, please refer to the original work (Choe et al., 2014).

Experimental setup

MRI data were collected using a 3 Tesla Siemens Tim Trio scanner equipped with a 12-channel Head Matrix coil at the Seoul National University Brain Imaging Center. Stimuli were generated using MATLAB (MathWorks) in conjunction with MGL (http://justingardner.net/mgl) on a Macintosh computer. Observers looked through an angled mirror attached to the head coil to view the stimuli displayed via an LCD projector (Canon XEED SX60) onto a back-projection screen at the end of the magnet bore at a viewing distance of 87 cm, yielding a field of view of 22 × 17°.

Behavioral data acquisition

Figure 2 illustrates the experimental procedures. On each trial, the observer initially viewed a small fixation dot (diameter in visual angle, 0.12°; luminance, 321 cd/m^2) appearing at the center of a dark (luminance, 38 cd/m^2) screen. A slight increase in the size of the fixation dot (from 0.12° to 0.18° in diameter), which was readily detected with foveal vision, forewarned the observer of an upcoming presentation of a test stimulus. The test stimulus was a brief (0.3 s) presentation of a thin (full-width at half-maximum of a Gaussian envelope, 0.17°), white (321 cd/m^2), dashed (radial frequency, 32 cycles/360°) ring that counter-phase-

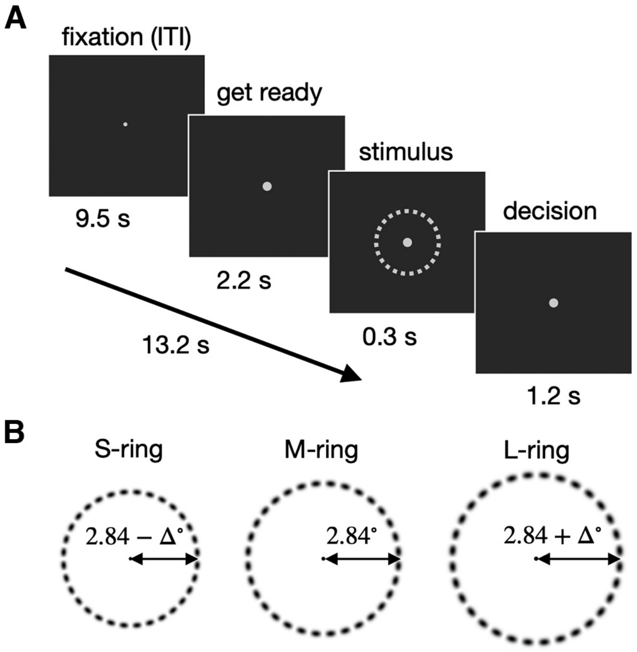

Figure 2. Binary classification task on ring size. ***A***, Within-trial procedure. With the eyes fixed, human participants were prewarned (2.2 s), with the increase of the fixation dot, to get ready for the upcoming trial after a long intertrial interval (9.5 s), briefly viewed the ring stimulus (0.3 s), and judged its size as large or small in respect to the medium size ring within a limited window of time (1.5 s). ***B***, Ring stimuli with threshold-level differences in size. On each trial, a participant viewed one of the three rings, small (S), medium (M), large (L), the size contrast (Δ) of which was optimized to ensure threshold-level classification performance on a participant-to-participant basis in a separate calibration run inside the MR scanner, right before the main session of fMRI scan runs. The order of ring sizes over trials was constrained with an m-sequence to preclude the temporal correlation among stimuli. Here, the luminance of the rings is inverted here for an illustrative purpose.

flickered at 10 Hz. After each presentation, participants classified the ring size into small or large using a left-hand or right-hand key, respectively, within 1.5 s from stimulus onset. They were instructed to maintain strict fixation on the fixation dot throughout experimental runs. This behavioral task was performed in three different environments: (1) the training sessions, (2) the practice runs of trials inside the MR scanner, and (3) the main scan runs inside the MR scanner, in the following order.

In the training sessions, participants practiced the task intensively over several (three to six) sessions (~1000 trials per session) in a dim room outside the scanner until they reached an asymptotic level of accuracy. Note that we opted to train observers with the stimuli that were much larger than those for the main experiments (mean radius of 9°) to avoid any unwanted perceptual learning effects at low sensory levels and to train participants to learn the task structure of classification.

In the practice runs of trials inside the MR scanner, participants performed 54 practice trials and then 180 threshold-calibration trials while lying in the magnet bore. On each of the threshold-estimation trials in which consecutive trials were apart from one another by 2.7 s., one of 20 different-sized rings was presented according to a multiple random staircase procedure (four randomly interleaved one-up-two-down staircases, two starting from the easiest stimulus and the other two starting from the hardest one) with trial-to-trial feedback based on the class boundary with the radius of 2.84°. A Weibull function was fit to the psychometric curves obtained from the threshold-calibration trials using a maximum-likelihood procedure. From the fitted Weibull function, the threshold difference in size (Δ in Fig. 2*B*) associated with a 70.7% correct proportion of responses was estimated. By finding this threshold for each participant, three threshold-level ring sizes were individually tailored as 2.84−Δ° (S-ring), 2.84° (M-ring), 2.84+Δ° (L-ring).

In the main scan runs, one of these rings with threshold-level differences was presented in the order defined by an m-sequence (base = 3, power = 3; nine S and L-rings and eight M-rings were presented; all scan runs started with two M-rings; Buracas and Boynton, 2002) to null the autocorrelation between stimuli. Participants were not informed of the existence of medium-ring. Importantly, participants did not receive trial-to-trial feedback. Instead, only their run-averaged percent correct based on the trials of S-ring and L-ring was shown during a break after each run, to prevent trial-to-trial feedback from evoking any unwanted brain responses associated with rewards (Marco-Pallarés et al., 2007; Carlson et al., 2011) or errors (Carter et al., 1998; Holroyd et al., 2004; Cavanagh and Frank, 2014). Consecutive trials were apart from one another by 13.2 s. In the main scan runs of Exp1 and Exp2, observers performed 156 (six runs × 26 trials) and 208 (eight runs × 26 trials) trials in total, respectively.

MRI equipment and acquisition

We acquired three types of MRI images. (1) 3D, T1-weighted, whole-brain images were acquired at the beginning of each functional session: MPRAGE; resolution, 1 × 1 × 1 mm; field of view (FOV), 256 mm; repetition time (TR), 1.9 s; time for inversion, 700 ms; time to echo (TE), 2.36 ms; and flip angle (FA), 9°. (2) 2D, T1-weighted, in-plane images were acquired at the beginning of each functional session. The parameters for the retinotopy-mapping, the V1 mapping, and the whole brain mapping differed slightly as follows (retinotopy, followed by the V1 mapping, and then by the whole brain mapping): MPRAGE; resolution, 1.078 × 1.078 × 2.0 mm, 1.083 × 1.083 × 2.3 mm 1.08 × 1.08 × 3.3 mm; TR, 1.5 s; T1, 700 ms; TE, 2.79 ms; and FA, 9°). (3) 2D, T2*-weighted, functional images were acquired during each functional session: gradient EPI; TR, 2.7, 2.2, 2.2 s; TE, 40 ms; FA, 77°, 73°, 73°; FOV, 208 mm, 207 mm, 208 mm; image matrix, 104 × 104, 90 × 90, 90 × 90; slice thickness, 1.8 mm with 11% gap, 2 mm with 15% slice gap, 3 mm with 10% space gap; slice, 30, 22, 32 oblique transfers slices; bandwidth, 858 Hz/px, 750 Hz/px, 790 Hz/px; and effective voxel size, 2.0 × 2.0 × 1.998 mm, 2.3 × 2.3 × 2.3 mm, 3.25 × 3.25 × 3.3 mm).

Retinotopy-mapping protocol

Standard traveling wave methods (Engel et al., 1994; Sereno et al., 1995) were used to define V1, to estimate each participant's hemodynamic impulse response function (HIRF) of V1, and to estimate V1 voxels' receptive field center and width. High-contrast and flickering (1.33 Hz) dartboard patterns were presented either as 0.89°-thick expanding or contracting rings in two scan runs, as 40°-width clockwise or counterclockwise rotating wedges in four runs or in one run as four stationary, 15°-wide wedges forming two bowties centered on the vertical and horizontal meridians. Each scanning run consisted of nine repetitions of 27-s period of stimulation. The fixation behavior during the scans was assured by monitoring participants' performance on a fixation task, in which they had to detect any reversal in direction of a small dot rotating around the fixation.

Data preprocessing of V1 images in the retinotopy-mapping session and the main session of Exp1

All functional EPI images were motion-corrected using SPM8 (http://www.fil.ion.ucl.ac.uk/spm; Friston et al., 1996; Jenkinson et al., 2002) and then co-registered to the high-resolution reference anatomic volume of the same participant's brain via the high-resolution in-plane image (Nestares and Heeger, 2000). After co-registration, the images of the retinotopy-mapping scan were resliced, but not spatially smoothed, to the spatial dimensions of the main experimental scans. The area V1 was manually defined on the flattened gray matter cortical surface mainly based on the meridian representations, resulting in 825.4±140.7 (mean±SD across observers) voxels. The individual voxels' time series were divided by their means to convert them from arbitrary intensity units to percentage modulations and were linearly detrended and high-pass filtered (Smith et al., 1999) using custom scripts in MATLAB (MathWorks). The cutoff frequency was 0.0185 Hz for the retinotopy-mapping session and 0.0076 Hz for the main session. The first 10 (of 90; a length of a cycle) and 6 (of 156; a length of a trial) frames of each run of the retinotopy-mapping session and main session, respectively, were discarded to minimize the effect of transient magnetic saturation and allow the hemodynamic response to reach a steady state. The "blood-vessel-clamping" voxels, which show unusually high variances of fMRI responses, were discarded (Olman et al., 2007; Shmuel et al., 2007); a voxel was classified as "blood-vessel-clamping" if its variance exceeds 10 times of the median variance value of the entire voxels. As the final step of data preprocessing, we removed a stimulus-nonspecific (untuned) component from the detrended BOLD time series by subtracting the across-eccentricity-bin average from the individual bins' time series at each time frame t, which resulted in the tuned responses (TR_i):

$$TR_i(t) = RR_i(t) - \sum_{i=1}^{n_e} RR_i(t)/n_e,$$

where RR_i is the i-th bin's BOLD time series, and n_e is the number of eccentricity bins (21). This subtraction procedure is exactly the same as we did in our previous work (Choe et al., 2014). We used $TR_i(t)$ to extract the size-encoding signal in V1.

Data preprocessing of whole-brain images in the main session of Exp2

The whole-brain images of the participants in Exp2 were normalized to the MNI template in the following steps: motion correction, coregistration to whole-brain anatomic images via the in-plane images (Nestares and Heeger, 2000), spike elimination, slice timing correction, resampling to 3 × 3 × 3-mm voxel size with the SPM DARTEL Toolbox (Ashburner, 2007). Spatial smoothing was not applied to avoid the blurring of the patterns of activity. All the procedures were implemented using SPM8 and SPM12 (https://www.fil.ion.ucl.ac.uk/spm-statistical-parametric-mapping/; Friston et al., 1996; Jenkinson et al., 2002), except for spike elimination, for which we used the AFNI toolbox (Cox, 1996). The first 6 frames of each functional scan, which correspond to the first trial of each run, were discarded to allow the hemodynamic responses to reach a steady state. Then, the normalized BOLD time series at each voxel, each run, and each brain underwent linear detrending, high-pass filtering (0.0076-Hz cutoff frequency with a Butterworth filter), conversion into percent-change signals, and correction for non-neural nuisance signals, which was done by regressing out the mean BOLD activity of CSF.

The anatomic masks of CSF, white matter, and gray matter were defined by generating the probability tissue maps for individual participants from T1-weighted images, by smoothing those maps to the normalized MNI space using SPM12, and then by averaging them across participants. Finally, the masks were defined as respective groups of voxels whose probabilities exceed 0.5.

Unfortunately, in a few of the sessions, functional images did not cover the entire brain. Especially, the lost part was much larger in one participant's session than the others including the orbitofrontal cortex and posterior cerebellum. Thus, not to lose too many of voxels for analysis because of this single session, we relaxed the criterion of voxel selection a bit by including the voxels that were shared by >16 brains in the normalized MNI space. As a result, some voxels in the temporal pole, ventral orbitofrontal, and posterior cerebellum were excluded from data analysis.

Estimation of the eccentricities in retinotopic space for V1 voxels

For each V1 voxel in Exp1, its eccentricity (e, as shown in Fig. 3*E*,*H*) was defined by fitting a one-dimensional Gaussian function simultaneously to the time-series of fMRI responses to the expanding and contracting ring stimuli in the retinotopy session, which were also used for the definition of V1. The essence of this procedure is as follows (additional details can be found in the original paper; Choe et al., 2014).

First, the time series of fMRI were extracted only from a relevant group of voxels with SNR > 3 in both of the ring scan runs. Second, an eccentricity-tuning curve (gain over eccentricity, in other words) of a single voxel, $g(\varepsilon)$, was modeled by a Gaussian as a function of the eccentricity in a visuotopic space, ε, and it was parameterized by a peak eccentricity, e, and a tuning width, σ:

$$g_e(\varepsilon) = exp^{-\left(\frac{(\varepsilon - e)^2}{2\sigma^2}\right)}.$$

Third, the collective responses of visual neurons within that voxel with a particular $g(\varepsilon)$ at a given time frame t, $n(t)$, were predicted by multiplying $g(\varepsilon)$ by spatial layout of stimulus input at that time frame, $s(\varepsilon, t)$:

$$n(t) = \sum_{\varepsilon} s(\varepsilon, t) g(\varepsilon).$$

Fourth, the predicted time-series of fMRI responses of that voxel, $fMRI_p(t)$, were generated by convoluting $n(t)$ with a scaled (by β) copy of the HIRF acquired from the meridian scans, $h(t)\beta$, and plus a baseline response, b:

$$fMRI_p(t) = n(t) * h(t)\beta + b.$$

Fifth, the eccentricity e and the other model parameters (σ, β, b) were found by fitting $fMRI_p(t)$ to the predicted time-series of fMRI responses to the actual stimulation, $fMRI_o(t)$, by minimizing the residual sum of squared errors between $fMRI_p(t)$and $fMRI_o(t)$ over all time frames, RSS:

$$RSS = \sum_t \left(fMRI_o(t) - fMRI_p(t)\right)^2.$$

Extraction of the size-encoding signal from V1 voxels

The three different weighting profiles, each representing the contributions of the individual eccentricity bins assessed by the three different schemes (the uniform, the discriminability, and the log-likelihood ratio schemes), were defined as follows. The uniform scheme (Fig. 4*B*, blue) assigned three discrete values to the eccentricity bins depending on which flanking side of the M-ring (r_M) their preferred eccentricities (e) belonged to:

$$w(e) = \begin{cases} -1, \, for \, e < r_M \\ 0, \, for \, e = r_M \\ 1, \, for \, e > r_M \end{cases}.$$

The discriminability scheme (Fig. 4*B*, red) defined the weights in proportion to the differential responses of given eccentricity bins to the L (r_L) and the S-rings (r_S), which were derived from the eccentricity-tuning curves defined from the retinotopy-mapping session:

$$w(e) = g_e(r_L) - g_e(r_S) - \delta,$$

where g_e is the eccentricity-tuning curve of the eccentricity bin with preferred eccentricity, e, and the baseline offset, δ, is as follows:

$$\sum_e \left[g_e(r_L) - g_e(r_S)\right]/n_e.$$

The log-likelihood ratio scheme (Fig. 4*B*, yellow) defined the weights by taking the differences between the log-likelihoods of obtaining a given response if the stimulus were the L-ring, $\log L_L$, and if the stimulus were the S-ring, $\log L_S$. Because the eccentricity-tuning curves were assumed to be described by a Gaussian function, the log-likelihood ratio weights at preferred eccentricity, e, can be simplified to the following formula:

$$w(e) = \log L_L - \log L_S = -\frac{1}{2\sigma_L^2}(e - r_L)^2 + \frac{1}{2\sigma_S^2}(e - r_S)^2 - \delta,$$

where σ_L and σ_S are the tuning widths with r_L and r_S, and the baseline offset, δ, is as follows:

$$\sum_{e}\left[-\frac{1}{2\sigma_L^2}(e-r_L)^2+\frac{1}{2\sigma_S^2}(e-r_S)^2\right]/n_e.$$

A Bayesian model of boundary-updating (BMBU)

The generative model

The generative model is the observers' causal account for noisy sensory measurements, where the true ring size, S, causes a noisy sensory measurement on a current trial, $m_{(t)}$, which becomes noisier as i trials elapse, thus turning into a noisy retrieved measurement of the value of S on trial $t-i$, $r_{(t-i)}$ (Fig. 5*D*). Hence, the generative model can be specified with the following three probabilistic terms: a prior of S, $p(S)$, a likelihood of S given $m_{(t)}$, $p(m_{(t)}|S)$, and a likelihood of S given $r_{(t-i)}$, $p(r_{(t-i)}|S)$. These three terms were all modeled as normal distribution functions, the shape of which is specified with mean and standard deviation parameters, μ and σ: μ_0 and σ_0 for the prior, $\mu_{m_{(t)}}$ and $\sigma_{m_{(t)}}$ for the likelihood for $m_{(t)}$, and $\mu_{r_{(t-i)}}$ and $\sigma_{r_{(t-i)}}$ for the likelihood for $r_{(t-i)}$. The mean parameters of the two likelihoods, $\mu_{m_{(t)}}$ and $\mu_{r_{(t-i)}}$, are identical to $m_{(t)}$ and $r_{(t-i)}$; therefore, the parameters that must be learned are reduced to μ_0, σ_0, $\sigma_{m_{(t)}}$, and $\sigma_{r_{(t-i)}}$.

$\sigma_{m_{(t)}}$ is assumed to be invariant across different values of $m_{(t)}$, as well as across trials. Therefore, $\sigma_{m_{(t)}}$ is reduced to a constant σ_m. Finally, because $\sigma_{r_{(t-i)}}$ is assumed to originate from σ_m and to increase as trials elapse (Gorgoraptis et al., 2011; Zokaei et al., 2015), $\sigma_{r_{(t-i)}}$ is also reduced to the following parametric function: $\sigma_{r_{(t-i)}}=\sigma_m(1+\kappa)^i$, where $\kappa>0$. As a result, the generative model is completely specified by the four parameters, $\Theta=\{\mu_0,\sigma_0,\sigma_m,\kappa\}$.

The primary purpose of BMBU is to build a generative Bayesian model which allows us to estimate the trial-to-trial latent states of the class boundary variable that are likely to be used by human observers whose class boundary is continually attracted to previous stimuli as posited by the boundary-updating hypothesis on "repulsive bias." In doing so, we intended to build a parsimonious model with minimal free parameters as long as the model implements the strategy essential to the boundary-updating hypothesis. For this reason, we had to introduce several arbitrary assumptions in building BMBU. For example, although we assumed that memory precision decays exponentially, other forms of decay function are also possible, such as hyperbolic, power, and logarithmic ones. We also assumed that the noisy sensory measurement on a current trial, $m_{(t)}$, becomes the noisy retrieved measurement of the value of S as trials elapse. However, it is equally possible that the memory measurements of S in the elapsed trials can be retrieved independently from the sensory measurement used for decision-making. Whether or not these assumptions are valid might be an interesting research question but is beyond the scope of the current work, especially in that the alternative assumptions about such detailed modeling aspects are unlikely to affect the way BMBU shifts the class boundary toward previous stimuli.

Stimulus inference (s)

A Bayesian estimate of the value of S on a current trial, $s_{(t)}$, was distributed as a posterior function of a given sensory measurement $m_{(t)}$:

$$\begin{aligned}p(s_{(t)}) &= p(S|m_{(t)})\\ &\propto p(m_{(t)}|S)p(S)\end{aligned}.$$

The posterior $p(S|m_{(t)})$ is a conjugate normal distribution of the prior and likelihood of S given the evidence $m_{(t)}$ whose mean $\mu_{s_{(t)}}$ and standard deviation $\sigma_{s_{(t)}}$ were calculated as follows (Fig. 5*D*):

$$\mu_{s_{(t)}}=\frac{\sigma_0^2 m_{(t)}+\sigma_m^2\mu_0}{\sigma_0^2+\sigma_m^2};\ \sigma_{s_{(t)}}=\frac{\sigma_0\sigma_m}{\sqrt{\sigma_0^2+\sigma_m^2}}.$$

Class boundary inference (b)

The Bayesian observer infers the value of class boundary on a current trial, $b_{(t)}$, by inferring the posterior function of a given set of retrieved sensory measurements $\vec{r}_{(t)}=\{r_{(t-1),}r_{(t-2),}\ldots r_{(t-n)}\}$:

$$b_{(t)}=\widetilde{S}=\arg\max_{S} p(S|\vec{r}_{(t)}),$$

where the maximum number of measurements that can be retrieved, n, was set to 7. We set 7 because it is much longer than the effective trial lags of the previous stimulus effect (Fig. 5*C*). Here, $p(S|\vec{r}_{(t)})$ is a conjugate normal distribution of the prior and likelihoods of S given the evidence $\vec{r}_{(t)}$:

$$\begin{aligned}p(S|\vec{r}_{(t)}) &\propto p(\vec{r}_{(t)}|S)p(S)\\ &= p(r_{(t-1)}|S)p(r_{(t-2)}|S)\ldots p(r_{(t-7)}|S)p(S)\end{aligned},$$

whose mean and standard deviation were calculated (Bromiley, 2003) based on the knowledge of how the retrieved stimulus becomes noisier as trials elapse:

$$\mu_{b_{(t)}}=\beta_0\mu_0+\sum_{i=1}^{7}\beta_i r_{(t-i)};\ \sigma_{b_{(t)}}=\sqrt{\beta_0^2\sigma_0^2+\sum_{i=1}^{7}\beta_i^2\sigma_{r_{(t-i)}}^2},$$

where $\beta_0=\frac{\sigma_0^{-2}}{\sigma_0^{-2}+\sum_{i=1}^{7}\sigma_{r_{(t-i)}}^{-2}}$ and $\beta_i=\frac{\sigma_{r_{(t-i)}}^{-2}}{\sigma_0^{-2}+\sum_{i=1}^{7}\sigma_{r_{(t-i)}}^{-2}}$. We postulated that the uncertainty of $b_{(t)}$ is equivalent to $\sigma_{b_{(t)}}$ (Fig. 5*G*).

Deduction of decision variable (v), *decision* (d), *and decision uncertainty* (u)

On each trial, the Bayesian observer makes a binary decision $d_{(t)}$ by calculating the probability of $s_{(t)}$ is larger than $b_{(t)}$, which is called the decision variable, $v_{(t)}$, defined as

$$v_{(t)}=p(s_{(t)}>b_{(t)})=\Phi\left[\frac{s_{(t)}-b_{(t)}}{\sqrt{\sigma_{s_{(t)}}^2+\sigma_{b_{(t)}}^2}}\right].$$

Then, if $v_{(t)}$ is larger than 0.5, $d_{(t)}$ is *large*. Otherwise, $d_{(t)}$ is *small*. Also, we defined the decision uncertainty, $u_{(t)}$, which represents the odds that the current decision will be incorrect (Sanders et al., 2016), as follows:

$$u_{(t)}=\Phi\left[\frac{-|s_{(t)}-b_{(t)}|}{\sqrt{\sigma_{s_{(t)}}^2+\sigma_{b_{(t)}}^2}}\right].$$

Fitting the parameters of BMBU

For each human participant, the parameters of the generative model, $\Theta=\{\mu_0,\sigma_0,\sigma_m,\kappa\}$, were estimated as those maximizing the sum of log-likelihoods for T individual choices made by the observer, $\vec{D}_{(T)}=[D_{(1)},D_{(2)},\ldots,D_{(T)}]$:

$$\hat{\Theta}=\arg\max_{\Theta}\sum_{t=1}^{T}\log p(D_{(t)}|\Theta).$$

For each participant, estimation was conducted in the following steps. First, we found local minima of parameters using a MATLAB function, *fminsearchbnd.m*, with the iterative evaluation number set to 50. We repeated this step by choosing 1000 different initial parameter sets, that were randomly sampled within uniform prior bounds, and acquired 1000 candidate sets of parameter estimates. Second, from these candidate sets of parameters, we selected the top 20 in terms of goodness-of-fit (sum of log-likelihoods) and searched the minima using each of those 20 sets as initial parameters by increasing the iterative evaluation number to 100,000 and setting tolerances of function and parameters to 10^{-7} for reliable estimation. Finally, using the parameters fitted via the second step, we repeated the second step one more time. Then, we selected the parameter set that showed the largest sum of likelihoods as the final parameter estimates. We discarded (1) the first trial of each run and (2) the trials in which RTs were too short (less than 0.3 s) for parameter estimation for any further analyses because (1) the first trial of each run does not have its previous trial, which is necessary for investigating the repulsive bias, and (2) the response made during the stimulus is shown (0–0.3 s) can be considered too hasty to reflect a normal cognitive decision-making process.

A constant-boundary model

The constant-boundary model has two parameters, bias of class boundary μ_0 and measurement noise σ_m. Stimulus estimates, $s_{(t)}$, were assumed to be sampled from a normal distribution, $\mathcal{N}(S_{(t)}, \sigma_m)$. Each stimulus sample has uncertainty $\sigma_{s_{(t)}} = \sigma_m$. Class boundary $b_{(t)}$ was assumed to be a constant, μ_0; so $\sigma_{p(b_{(t)})} = \sigma_{b_{(t)}} = 0$.

Estimation of the latent states of the variables of BMBU

Fitting the model parameters separately for each human participant ($\hat{\Theta} = \{\hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_m, \hat{\kappa}\}$) allowed us to create the same number of Bayesian observers, each tailored to each human individual. We repeated the experiment on these Bayesian observers using the stimulus sequences identical to those presented to their human partners for the following two purposes. First, we wanted to examine whether BMBU's choice ($d_{(t)}$) can reproduce the human partners' repulsive bias. Second, we need to estimate the trial-to-trial latent states of the model variables ($s_{(t)}$, $b_{(t)}$, $v_{(t)}$, $u_{(t)}$) that were used by the human partners, thus represented in their brains engaged in the binary classification task. We acquired a sufficient number (10^6 repetitions) of simulated choices, $d_{(t)}$, and decision uncertainty values, $u_{(t)}$, which were determined by the corresponding number of the stimulus estimates, $s_{(t)}$, and the boundary estimates, $b_{(t)}$, for each Bayesian observer. Then, the averages across those 10^6 simulations were taken as the final outcomes. When estimating $s_{(t)}$, $b_{(t)}$, $v_{(t)}$, and $u_{(\mathrm{t})}$ for the observed choice $D_{(t)}$, we only included the simulation outcomes in which the simulated choice $d_{(t)}$ matched the observed choice $D_{(t)}$.

Recovery of the true states of the model variables

To ascertain the validity of our procedure of estimating the latent variables of BMBU described above, we checked how accurately it recovers the true states of the variables. This recovery test was conducted in the following procedure.

First, we created 256 different sets of parameter values by taking the possible combinations of the four different values of each of the four model parameters, where the four different values corresponded to the 20th, 40th, 60th, and 80th percentiles of the parameter values fitted to the observers' choices. Second, we acquired the synthetic choices and the *true* model variables *b*, *s*, *v*, and *u* by plugging one parameter set into BMBU and simulating it on the actual stimulus sequence presented to the observers. Third, we fitted the parameters of BMBU to the synthetic choices in the same procedure conducted for fitting BMBU to the observed choices. Fourth, we simulated a set of the *recovered* states of the model variables using the fitted model parameters. Fifth, we calculated the R^2 between the true and the recovered variables to assess how reliably our model fitting procedure can recover the true states of the model variables. Finally, we repeated the above procedure for all the remained parameter sets and used the R^2 averaged across the 256 parameter sets as the performance measure of the recovery test.

The multiple logistic regression model for capturing the repulsive bias

To capture the repulsive bias in human classification, we logistically regressed the current choice onto stimuli and choices using the following regression model to obtain regression coefficients $\vec{p} = \{p_{(1)}, \cdots, p_{(11)}\}$ for each observer:

$$D_{(t)} = \frac{e^{K_{(t)}}}{1 + e^{K_{(t)}}},$$

where $K_{(t)} = p_{(0)} + p_{(1)}S_{(t)} + \sum_{i=1}^{5}(p_{(1+i)}S_{(t-i)} + p_{(6+i)}D_{(t-i)})$, the independent variables were each standardized to *z* scores for each participant. $S_{(t)}$ and $D_{(t)}$ are the stimulus and the observed choice values at trial *t*. $S_{(t-i)}$ and $D_{(t-i)}$ are the stimulus and the observed choice at the *i*th trial lags from trial *t*.

To capture the repulsive bias of the Bayesian observers, the Bayesian observers' choices were also regressed with the logistic regression model by substituting $d_{(t)}$ and $d_{(t-i)}$, the simulated choices, for $D_{(t)}$ and $D_{(t-i)}$, the observed choices. The regression was repeatedly conducted for each simulation, and the regression coefficients that were averaged across simulations were taken as final outcomes. The simulation was repeated 10^5 times. We confirmed that the simulation number was sufficiently large to produce stable simulation outcomes.

The average marginal effect analysis

Average marginal effect (AME) was calculated by using the R-package "margins" (Leeper et al., 2018). AME quantifies the average marginal effect between an ordinal dependent variable (i.e., binary choice) and an independent variable of a multiple logistic (or probit) regression model (Williams and Jorgensen, 2023). To calculate the AMEs of any given variable on the current choice ($D_{(t)}$) without controlling the previous ($S_{(t-1)}$) and current stimuli ($S_{(t)}$; i.e., the baseline AME), we implemented a logistic regression model with two regressors –the variable of interest *X* (i.e., V1, *b*, *s*, or *v*) and the previous choice ($D_{(t-1)}$):

$$D_{(t)} \sim logit(\beta_0 + \beta_X X + \beta_{\mathrm{D}_1} D_{(t-1)}).$$

We always included $D_{(t-1)}$ as a regressor because the effect of $D_{(t-1)}$ would confound the effect of $S_{(t-1)}$, if $D_{(t-1)}$ is not included in the regression model. Specifically, because $S_{(t-1)}$ and $D_{(t-1)}$ are highly correlated, it would be unclear whether the AME difference before and after controlling $S_{(t-1)}$ is ascribed to the effect of $S_{(t-1)}$ or that of $D_{(t-1)}$, if $D_{(t-1)}$ is not controlled. The effect of $D_{(t-1)}$ was controlled in all regression models.

To test whether the AME of *X* decreased after controlling $S_{(t-1)}$ (or $\mathrm{S}_{(t)}$), we calculated the AME of *X* from the logistic regression model including $S_{(t-1)}$ (or $S_{(t)}$) as an additional regressor, as follows:

$$D_{(t)} \sim logit\left(\beta_0 + \beta_X X + \beta_{D_{(t-1)}} D_{(t-1)} + \beta_{S_{(t-1)}(or S_{(t)})} S_{(t-1)} (or\, S_{(t)})\right),$$

and subtracted the new AME from the baseline AME to see whether the baseline AME significantly changed after controlling previous or current stimuli.

Searching for the multivoxel patterns of activity representing the latent variables of BMBU

We assumed that (1) activity patterns of neural population for representing the latent variables are different between participants, but (2) locations and (3) timings of the activity patterns overlap across participants. Therefore, to identify the brain signals of the latent variables of BMBU in fMRI responses, the support vector regression (SVR) decoding was conducted for each human participant within specific spatial and temporal windows.

As for the spatial window, we implemented a searchlight technique (Kahnt et al., 2011b; Haynes, 2015). A searchlight has a radius of 9 mm (= 3 voxels; Soon et al., 2008) and thus can contain 123 voxels at most. Of the 123 voxels, we excluded the voxels located in CSF or white matter because they reflect non-neural signals. Thus, the effective number of voxels in a searchlight used for the analysis can vary searchlight by searchlight.

As for the temporal windows, we implemented the time-resolved decoding technique in which a target variable is decoded from the BOLD responses at each of the within-trial time points (Fig. 6*B*). We used the first four time points (out of six in total) because the BOLD responses associated with the action of button press, the last process of the sensory-to-motor decision-making stream, is maximized at the fourth time point (the result is not shown here). In sum, SVR is trained for each participant, each time point, and each searchlight.

Before training SVR, the BOLD responses in a searchlight and a target latent variable were z-scored across trials. Then, the z-scored variable

Table 1. The sets of regressions that BMBU requires the brain signals of its latent variables to satisfy

Test index	$b_{(t)}$			$s_{(t)}$			$v_{(t)}$		
Regressor	Regressor	Threshold p-value	Tail	Regressor	Threshold p-value	Tail	Regressor	Threshold p-value	Tail
1	b	0.001 ↓	Right	s	0.001 ↓	Right	v	0.001 ↓	Right
2	b	0.05 ↓ (fdr)	Right	s	0.05 ↓ (fdr)	Right	v	0.05 ↓ (fdr)	Right
3	$b_{\perp v}$	0.05 ↓	Right	$s_{\perp v}$	0.05 ↓	Right	$v_{\perp b}$	0.05 ↓	Right
4	$b_{\perp d}$	0.05 ↓	Right	$s_{\perp d}$	0.05 ↓	Right	$v_{\perp s}$	0.05 ↓	Right
5	s	0.05 ↑	Both	b	0.05 ↑	Both	$v_{\perp d}$	0.05 ↓	Right
6	v	0.05 ↓	Left	v	0.05 ↓	Right	b	0.05 ↓	Left
7	$v_{\perp b}$	0.05 ↑	Left	$v_{\perp s}$	0.05 ↑	Right	$b_{\perp v}$	0.05 ↑	Left
8	d	0.05 ↓	Left	d	0.05 ↓	Right	s	0.05 ↓	Right
9	u	0.05 ↑	Both	u	0.05 ↑	Both	$s_{\perp v}$	0.05 ↑	Right
10	$S_{(t)}$	0.05 ↑	Both	$S_{(t)}$	0.05 ↓	Right	d	0.05 ↓	Right
11	$S_{(t-1)}$	0.05 ↓	Right	$S_{(t-1)}$	0.05 ↑	Both	u	0.05 ↑	Both
12	$S_{(t-2)}$	0.05 ↑	Left	$S_{(t-2)}$	0.05 ↑	Both	$S_{(t)}$	0.05 ↓	Right
13	$D_{(t-1)}$	0.05 ↑	Both	$D_{(t-1)}$	0.05 ↑	Both	$S_{(t-1)}$	0.05 ↓	Left
14	$D_{(t-2)}$	0.05 ↑	Both	$D_{(t-2)}$	0.05 ↑	Both	$S_{(t-2)}$	0.05 ↑	Right
15							$D_{(t)}$	0.05 ↓	Right
16							$D_{(t-1)}$	0.05 ↑	Both
17							$D_{(t-2)}$	0.05 ↑	Both

The regressions required for the brain signal of the inferred class boundary ($b_{(t)}$; left sector), the inferred stimulus ($s_{(t)}$; middle sector), and the decision variable ($v_{(t)}$; right sector). The top sector (#1∼#9 for $b_{(t)}$; #1∼#9 for $s_{(t)}$; #1∼#11 for $v_{(t)}$) specifies the individual, simple regression models in which the brain signal of interest is regressed on a single regressor (second column). Any regressor subscripted with another variable with the perpendicular symbol (e.g., $b_{\perp v}$) means that the residuals of the left-side variable (e.g., b) from the regression of the right-side variable with the perpendicular symbol (e.g., v) were used as the regressor. This regression with the residual regressor was created to check whether the brain variable of interest has a unique covariation with the original regressor by withholding the influence of the perpendiculared variable (e.g., pSTG_{b5} must be positively correlated with b even when the part of b's variability associated with v is withheld). The bottom sector of each table (#10∼#14 for $b_{(t)}$; #10∼#14 for $s_{(t)}$; #12∼#17 for $v_{(t)}$) specifies the multiple-regression model in which the brain signal of interest is regressed concurrently on the current and previous stimuli and the past or current choices. The third and fourth column of each table specify the statistical criteria used for significance test, where fdr indicates a multiple comparison test controlling the false discovery rate.

was decoded for each searchlight using the cross-validation method of leave-one-run-out (eightfold cross-validation). As a result, for each searchlight and at each time point, we acquired a set of decoded latent variables in all trials. In other words, on each time point, we acquired the 4-dimensional map of the decoded variable (i.e., three spatial dimensions and 1 trial dimension). The 3D spatial dimensions of the decoded variables were smoothed with a 5 mm FWHM Gaussian kernel on each trial.

After this subject-wise decoding analysis, we conducted the across-subject analysis to test whether the decoded variables are significantly informative. To do so, for each searchlight locus and each time point, we regressed the smoothed decoded variable onto the regression conditions of the target variable by using a generalized linear mixed effect regression model (GLMM) with a random effect of subjects. The number of regression conditions was 14, 14, and 17 for $b_{(t)}$, $s_{(t)}$, and $v_{(t)}$, respectively (Table 1). Those regression models were deduced from the causal structure between the variables of BMBU (see the next section). We accepted a given cluster as the brain signals of $b_{(t)}$, $s_{(t)}$, or $v_{(t)}$ only when they satisfied those regression models over >12 contiguous searchlights. For the ROI analysis, the decoded variables were averaged over all searchlights within each ROI.

SVR was conducted using LIBSVM (https://www.csie.ntu.edu.tw/∼cjlin/libsvm/) with a linear kernel and constant regularization parameter of 1 (Soon et al., 2008; Kahnt et al., 2011b). The brain imaging results were visualized using Connectome Workbench (Marcus et al., 2011) and xjview.

The regression-model test for verifying the brain signals of $b_{(t)}$, $s_{(t)}$, and $v_{(t)}$

To identify the brain signals of $b_{(t)}$, $s_{(t)}$, and $v_{(t)}$, we defined three respective lists of regressions that must be satisfied by the brain signals. We stress that each of these lists consists of the necessary conditions to be satisfied because the conditions are deduced from the causal structure of the variables in BMBU (Fig. 5G). Below, we specify the specific regression tests for $s_{(t)}$ and $v_{(t)}$ that constitute these lists. For the tests for $b_{(t)}$, see Results.

The 14 regressions for the brain signal of $s_{(t)}$ (Table 1): (#1–4), y_s, s decoded from brain signals, must be regressed positively onto s, the variable it represents, even when the false discovery rate is controlled (Benjamini and Hochberg, 1995), and s orthogonalized to v or d because it should reflect the variance irreducible to the offspring variables of s; (#5), y_s must not be regressed onto b because s and b are independent of each other ($b \nleftrightarrow s$; Fig. 5G); (#6, 7), y_s must be positively regressed onto v ($s \rightarrow v$; Fig. 5G) but not when v is orthogonalized to s because the influence of s on v is removed; (#8, 9) y_s must be positively regressed onto d ($s \rightarrow v \rightarrow d$; Fig. 5G) but not onto u because u cannot be linearly correlated with s ($s \rightarrow v \rightarrow u$ is blocked by the interaction between u and v; Fig. 5G); (#10–12), y_s must be positively regressed onto the current stimuli and not the past stimuli because s is inferred solely from the current stimulus measurement; (#13, 14), y_s must not be regressed onto previous decisions because s is inferred solely from the current stimulus measurement. #10–14 were investigated by a multiple regression with regressors $\left[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t-1)}, D_{(t-2)}\right]$. We did not include $D_{(t)}$ as a regressor because $D_{(t)}$ may induce a spurious correlation between b and s by controlling the collider v (Elwert and Winship, 2014; $b \rightarrow v \leftarrow s$ and $v \rightarrow d$; Fig. 5G).

The 17 regressions for the brain signal of $v_{(t)}$ (Table 1). (#1–5), y_v, v decoded from brain signals, must be positively regressed onto v, the variable it represents, even when the false discovery rate is controlled (Benjamini and Hochberg, 1995), and v orthogonalized to b, s, or d, because it should reflect the variance irreducible to the offspring variables of v; (#6, 7), y_v must be negatively regressed onto one of its parents b ($b \rightarrow v$; Fig. 5G), but not when b is orthogonalized to v, because the influence of b on v is removed; (#8, 9), y_v must be positively regressed onto one of another parent s ($s \rightarrow v$; Fig. 5G), but not when s is orthogonalized to v, because the influence of s on v is removed; (#10, 11), y_v must be regressed onto d but not onto u because u's correlation with its parent v cannot be revealed without holding the variability of d (the interaction between u and v); (#12–14), y_v must be positively regressed onto the current stimulus because the influence of the current stimulus on v is propagated via s ($S_{(t)} \rightarrow s \rightarrow v$), and negatively regressed onto the past stimuli because the influence of the past stimuli on v is propagated via b ($S_{(t-1)} \rightarrow b \rightarrow v$) —strongly onto the 1-back stimulus and more weakly onto the two-back stimulus (thus, nonsignificant regression with one-tailed regression in the opposite sign is modeled moderately); (#15–17), y_v must be regressed onto the current decision and not the past decisions because the current decision is a dichotomous translation of v ($v \rightarrow d$; Fig. 5G), whereas past decisions have nothing to do with the current state of v. #12–17 were investigated by a multiple regression with regressors $\left[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t)}, D_{(t-1)}, D_{(t-2)}\right]$. $D_{(t)}$ was included as a regressor because v does not suffer from a spurious correlation that arises by controlling a collider variable which is absent in this case.

Bayesian network analysis

To investigate whether the relationship between decoded *b*, *s*, and *v* is consistent with the causal structure postulated by BMBU, we calculated the BIC values for all the three-node networks consisting of the time series of three brain signals $\{y_b, y_s, y_v\}$ (Scutari, 2010) and determined the causal graph whose likelihood is maximal. The three-node network has 162 possible structures, as follows. A total of 27 edge structures can be created out of three nodes since three types of edges are possible for any given pair of nodes (i.e., $x \rightarrow y$, $x \leftarrow y$ or $x \nleftrightarrow y$) and there are three pairs (i.e., $\{b, v\}$, $\{v, s\}$, $\{s, b\}$; 3^3). Also, a total of 6 combinations of three nodes exist for $\{y_c, y_s, y_v\}$ since we have three (IPL$_{b1}$, pSTG$_{b3}$, pSTG$_{b5}$), two (DLPFC$_{s3}$, Cereb$_{s5}$), and single (aSTG$_{v5}$) brain signals of *b*, *s*, and *v*, respectively ($3 \times 2 \times 1$). Thus, because each of the 6 possible node combinations can have 27 edge structures, there are 162 possible three-node causal networks.

We opted to apply this Bayesian network analysis to the three-node networks instead of the six-node network consisting of all the six brain signals identified by the searchlight analysis because the number of possible six-node networks ($N = 3^{6C2} = 3^{15} = 14,348,907$) was unrealistically large so that the statistical results are likely to suffer from type I errors. In addition, guided by BMBU, we were interested in identifying the causal structure of the three brain signals, each corresponding to one of the three model variables (*b*, *s*, and *v*). In other words, we were not interested in the causal relationship between the brain signals representing the same model variable (e.g., between pSTG$_{b3}$, pSTG$_{b5}$).

Statistics

We used the searchlight technique to look for brain signals related to the latent variables of the BMBU. To make the searchlight analysis statistically powerful by reducing the noise effect in the BOLD signals, we applied a generalized linear mixed effect model (GLMM) with the random effect of observers to calculate the association between the true and the decoded model variables. We applied the mixed effect model only to the searchlight analysis (Fig. 6; Table 1). For the other regression analyses, we conducted the analysis for each individual, respectively, because the mixed effect model was too computationally demanding to be applied to all other analyses. For instance, applying GLMM to the model simulation depicted in Figure 5*C* requires 10^5 repetitions of regression analysis. The significance tests were two-tailed except for the searchlight analysis as specified in Table 1. Also, for the time-resolved searchlight analysis, we implemented the multiple-comparison test (the false discovery rate (fdr) correction; Benjamini and Hochberg, 1995) for each of the fMRI time frames. In the figures summarizing statistical results, all confidence intervals are the 95% confidence intervals of the mean across individual observers.

Results

Experimental paradigm

Over consecutive trials, participants sorted ring sizes into two classes, *small* and *large*, under moderate time pressure (Fig. 2*A*). To ensure decision-makings with uncertainty, we presented three rings (small, medium, and large) differing by a threshold size (Δ), which was tailored for individuals (Fig. 2*B*; see Materials and Methods). The ring sizes were presented in m-sequence to rule out any correlation between consecutive stimulus sizes (Buracas and Boynton, 2002). We provided participants with feedback after each scan run by summarizing their performance with the proportion of correct trials.

To verify the sensory-adaptation hypothesis, we conducted experiment 1, where 19 participants performed the classification task while BOLD measurements with a high spatial resolution were acquired only from their early visual cortices. To verify the boundary-updating hypothesis, we conducted experiment 2, where 18 participants performed the same task while their whole brains were imaged. The data of experiment 1 had been used in our published work (Choe et al., 2014).

Repulsive bias in experiment 1

The participants in experiment 1 displayed a substantive amount of repulsive bias. As anticipated, the proportion of large choices (PL) increased as the ring size on the current trial ($S_{(t)}$) increased. Importantly, when the psychometric curves were conditioned on the previous stimulus ($S_{(t-1)}$), they shifted upward as the ring size in the previous trial decreased (the contrasts between the solid, dotted, and dashed lines in Fig. 3*A*), which indicates the presence of repulsive bias. By contrast, the psychometric curves were not affected much by the previous choice (the contrasts between the gray and black lines in Fig. 3*A*). To quantify the impact of the previous stimulus on the current choice, we subtracted the PLs acquired when the previous ring size was S from those when L separately for each of the six combinatorial conditions of the current stimulus (three sizes) and previous choice (two alternatives) and then averaged those six PL differences. The averaged PL difference (-0.20) was significantly smaller than zero ($t_{(18)} = -8.9, p = 5.1 \times 10^{-8}$; Fig. 3*B*, left). We also quantified the impact of the previous choice on the current choice similarly: the PL differences of previous *large* from *small* choices were calculated separately for the nine combinatorial conditions of the current and previous stimulus and then averaged. The averaged PL difference (-0.018) did not significantly differ from zero ($t_{(18)} = -0.68, p = 0.50$; Fig. 3*B*, right).

To ensure this repulsive effect of the previous stimulus on the current choice, we logistically regressed each participant's current choice ($D_{(t)}$) simultaneously onto the previous stimulus and choice. The regression coefficients for the previous stimuli were significant up to two trial lags across participants ($S_{(t-1)}$, $\beta = -0.39$, $t_{(18)} = -9.6$, $p = 1.6 \times 10^{-8}$; $S_{(t-2)}$, $\beta = -0.11$, $t_{(18)} = -2.4$, $p = 0.026$; $D_{(t-1)}$, $\beta = -0.17$, $t_{(18)} = -2.8, p = 0.012$), which confirms the robust presence of repulsive bias in experiment 1 (Fig. 3*C*).

Sensory adaptation in V1

As a first step toward the verification of the sensory-adaptation hypothesis, we defined the size-encoding signal in V1. As our group showed previously (Choe et al., 2014); the eccentricity-tuned BOLD responses in V1 (Fig. 3*D*) readily resolved the threshold-level differences in ring size, as anticipated by the retinotopic organization of the V1 architecture (Fig. 3*E*). Thus, the subtraction of the BOLD responses at the voxels preferring S-ring to L-ring from those at the voxels preferring L-ring to S-ring (Fig. 3*F*) was significantly greater when $S_{(t)}$ was large than when small (the third and the fourth time points, $\beta = 0.11$, $t_{(18)} = 4.8$, $p = 1.5 \times 10^{-4}$and $\beta = 0.13$, $t_{(18)} = 6.6, p = 3.7 \times 10^{-6}$; Fig. 3*G*).

Next, having defined the size-encoding signal in V1, which will be referred to as "$V1$," we sought evidence of sensory-adaptation in that signal. According to the previous work on sensory-adaptation (Clifford et al., 2007; Kohn, 2007; Solomon and Kohn, 2014; Weber et al., 2019), we expected $V1$ to decrease following the large size and to increase following the small size because of the selective gain reduction at the sensory neurons tuned to previous stimuli. In line with this expectation, $V1$ indeed significantly decreased when preceded by L-ring than when preceded by S-ring (the fourth time point, $\beta = -0.45$, $t_{(18)} = -2.2$, $p = 0.040$; Fig. 3*H*,*I*). Although we rendered ineffective the autocorrelation between consecutive stimuli using an m-sequence (see Materials and Methods), we additionally checked the possibility that the observed adaptation of $V1$ might have spuriously occurred because of any imbalance in the ring size of the current

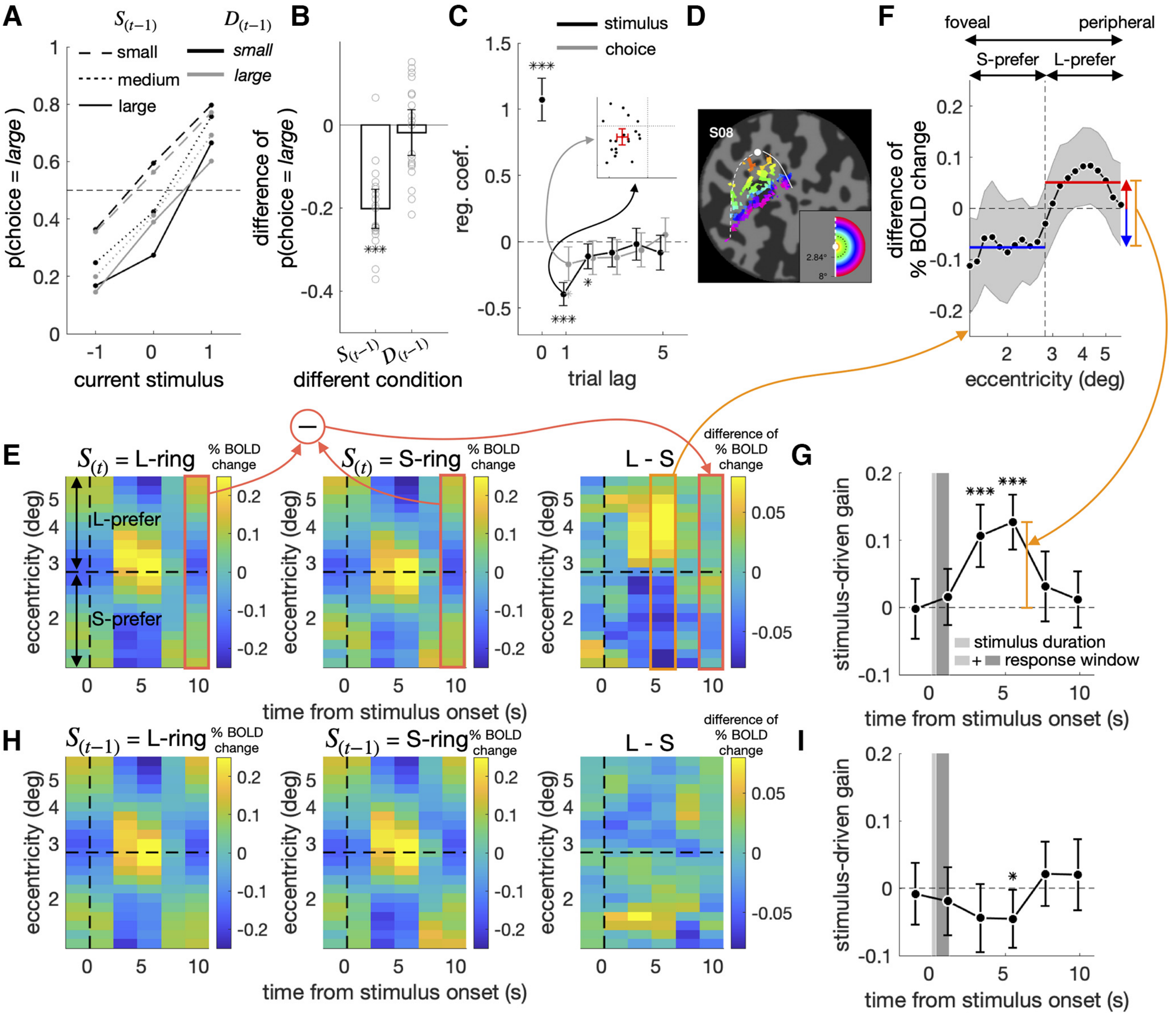

Figure 3. Influences of previous and current stimuli on classification behavior and V1 activity in experiment 1. ***A–C***, Repulsive bias in psychometric curves (***A***, ***B***) and regression analysis (***C***). The psychometric curves, where the fractions of *large* choices are plotted against the current stimulus, are shown separately for the six possible combinations defined by the previous stimulus and choice (***A***). As the summary of the effects of the previous stimulus on the current choice, the differences in the fractions of *large* choices between the previous stimuli were L-ring and S-ring ($p(D_{(t)} = large|S_{(t-1)} = 1) - p(D_{(t)} = large|S_{(t-1)} = -1)$) are computed separately for the six combinations of the current stimulus and previous choice and then averaged (***B***, left). As the summary of the effects of the previous choice on the current choice, the differences in the fractions of *large* choices between the previous choices were *large* and *small* ($p(D_{(t)} = large|D_{(t-1)} = large) - p(D_{(t)} = large|D_{(t-1)} = small)$) are computed separately for the nine combinations of the current and previous stimuli and then averaged (***B***, right). The small gray circles represent the individual observers. The multiple logistic regression coefficients of the current choice are plotted against trial lags (***C***). In the inset, the regression coefficients for the previous-stimulus ($S_{(t-1)}$) regressor are plotted against those for the previous-choice ($D_{(t-1)}$) regressor for individual observers, where the red error bars demarcate the 95% CIs of the means. ***D***, Eccentricity map of V1 on the flattened left occipital cortex of a representative brain, S08. The dot, curves, and colors correspond to those in the inset depicting the visual field. The image is borrowed from our previous work (Choe et al., 2014). ***E***, ***H***, Spatiotemporal BOLD V1 responses to L-ring (left) and S-ring (middle), and their differentials (right), presented on the current (***E***) and previous (***H***) trials. The color bars indicate BOLD changes in the unit of % signal, averaged across all participants. The vertical dashed line marks the time point for stimulus onset. The horizontal dashed line corresponds to the eccentricity of M-ring, splitting the voxels into "L-prefer" and "S-prefer" groups based on their preferred ring size. ***F***, The differential of BOLD responses at peak between the small and large ring on the current trial. The vertical dashed line marks the eccentricity of M-ring. The horizontal red and blue lines mark the average BOLD signals of the L-prefer and S-prefer voxels, respectively. The vertical orange line quantifies the stimulus-driven gain of V1 responses. ***G***, ***I***, Time courses of the stimulus-driven gain of V1 responses to the current (***G***) and previous (***I***) stimuli. The stimulus duration and response window are demarcated by the light and dark gray bars demarcate (***G***, ***I***). The 95% CIs of the mean across observers are indicated by the shaded areas (***F***) or by the vertical error bars (***B***, ***C***, ***G***, ***I***). Asterisks indicate the statistical significance (*$P<0.05$, **$P<10^{-3}$, ***$P<10^{-4}$; ***B***, ***C***, ***G***, ***I***). The orange boxes and arrows are drawn to help the relationships between the panels (***E–G***).

stimuli. To do so, we first calculated the differences in *V*1 between the previous S-rings and L-rings separately for the three current stimuli and then averaged those three differences. We confirmed that the averaged *V*1 differences were smaller when preceded by L-ring than when preceded by S-ring (the fourth time point, $\beta = -0.44$, $t_{(18)} = -2.1$, $p = 0.049$).

In sum, the V1 population activity reliably encoded the ring size and exhibited sensory adaptation.

The variability of *V*1 associated with previous stimuli fails to contribute to the choice variability

Next, we verified the critical prediction of the sensory-adaptation hypothesis on repulsive bias. Below, we will define what this

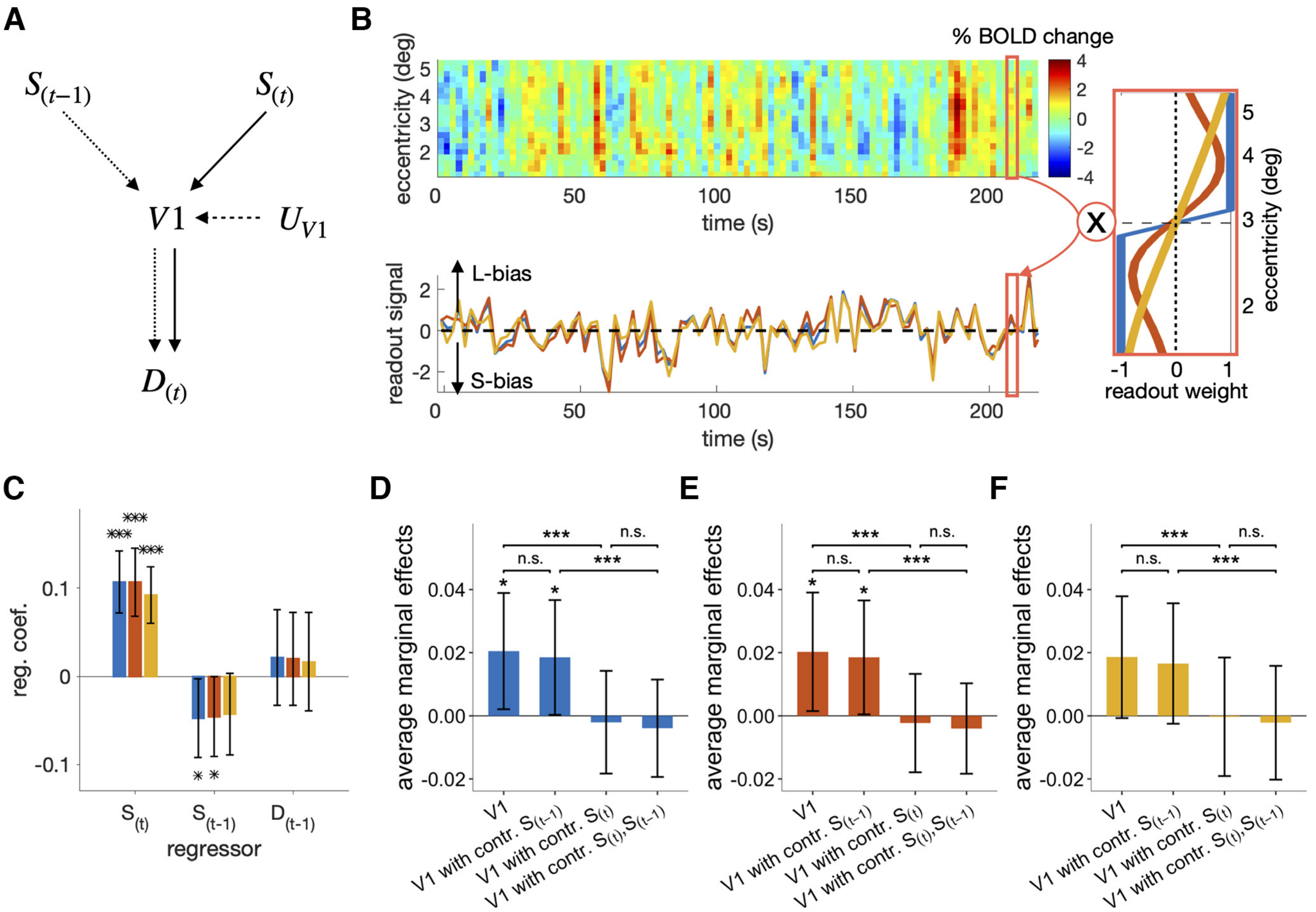

Figure 4. Origin of the covariation between the stimulus-encoding signal of V1 and the current choice. ***A***, The causal structure of the variables implied by the sensory-adaptation hypothesis. The stimulus-encoding signal of V1 ($V1$) is influenced by the current stimulus ($S_{(t)}$), the previous stimulus ($S_{(t-1)}$), and the unknown sources (U_{V1}). In turn, $V1$ influences the current choice ($D_{(t)}$). If the sensory-adaptation hypothesis is true, part of the causal influence of $V1$ on $D_{(t)}$ must originate from $S_{(t-1)}$, as indicated by the connected chain of the dotted arrows. ***B***, Extraction of the stimulus-encoding signal of V1. For any given run from any participant, the matrix of spatiotemporal BOLD responses in V1 (top left) was multiplied by one of the three weighting vectors (right; blue, red, and yellow lines represent the uniform, discriminability, and log-likelihood ratio readout schemes, respectively) to result in the vector of stimulus-encoding signal ($V1$) in the same trial length (bottom left). The positive and negative values of $V1$ indicate the larger and smaller sizes of the ring, respectively. ***C***, Multiple linear regression of the stimulus-encoding signal of V1 on $S_{(t)}$, $S_{(t-1)}$, and $D_{(t-1)}$. The colors correspond to the three different readout schemes in ***B***. ***D–F*** The average marginal effects (AMEs) of $V1$ on $D_{(t)}$, with $V1$ extracted by the uniform (***D***), discriminability (***E***), and log-likelihood ratio (***F***) readout schemes. In each panel, the influence of $V1$ on $D_{(t)}$ that can be ascribed to $S_{(t-1)}$ and $S_{(t)}$ were assessed by checking i) whether the AME of $V1$ on $D_{(t)}$ (left) significantly decreased or not after controlling the influence of $S_{(t-1)}$ (second from the left) and $S_{(t)}$ (second from the right), respectively, or ii) whether the AME of $V1$ on $D_{(t)}$ controlling the influence of both $S_{(t-1)}$ and $S_{(t)}$ (right) significantly increased or not after only controlling the influence of $S_{(t)}$ (second from the right) and $S_{(t-1)}$ (second from the left), respectively. Asterisks indicate the statistical significance (*$P<0.05$, **$P<0.01$, ***$P<0.001$), and "n.s." stands for the nonsignificance of the test (***C–F***). The 95% CIs of the mean across participants are indicated by the vertical error bars (***C–F***).

crucial prediction is and how we empirically examine that prediction.

Above, we confirmed that the ring size, not only on the current trial ($S_{(t)}$) but also on the previous trial ($S_{(t-1)}$), affects $V1$ on the current trial ($S_{(t-1)} \rightarrow V1 \leftarrow S_{(t)}$ in Fig. 4A). What we do not know yet is whether the variabilities of $V1$ that originate from $S_{(t)}$ and $S_{(t-1)}$, respectively, flow all the way into the observer's current choice ($S_{(t)} \rightarrow V1 \rightarrow D_{(t)}$ and $S_{(t-1)} \rightarrow V1 \rightarrow D_{(t)}$ in Fig. 4A). Critically, if the sensory-adaptation hypothesis is true, the variability of $V1$ associated with $S_{(t-1)}$ must contribute to the current choice ($D_{(t)}$; $S_{(t-1)} \rightarrow V1 \rightarrow D_{(t)}$), just as that associated with $S_{(t)}$ must do so ($S_{(t)} \rightarrow V1 \rightarrow D_{(t)}$). Here, it is important to realize that the mere association between $S_{(t)}$ and $V1$ ($S_{(t)} \rightarrow V1$) does not warrant their contribution to $D_{(t)}$ ($S_{(t)} \rightarrow V1 \rightarrow D_{(t)}$). Likewise, the association between $S_{(t-1)}$ and $V1$ ($S_{(t-1)} \rightarrow V1$) does not warrant their contribution to $D_{(t)}$ ($S_{(t-1)} \rightarrow V1 \rightarrow D_{(t)}$).

We can test the critical implication of the sensory-adaptation hypothesis by comparing the average marginal effect (AME; Williams and Jorgensen, 2023) of $V1$ on $D_{(t)}$ ($V1 \rightarrow D_{(t)}$) to that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ controlled ($S_{(t-1)} \not\rightarrow V1 \rightarrow D_{(t)}$). The rationale behind this comparison is that the contribution of $V1$ to $D_{(t)}$ must be substantially smaller when $S_{(t-1)}$ was controlled than when not if the contribution of $S_{(t-1)}$ to $D_{(t)}$ via $V1$ (i.e., $S_{(t-1)} \rightarrow V1 \rightarrow D_{(t)}$) is substantial. In addition, the critical implication can also be tested by comparing the AME of $V1$ on $D_{(t)}$ with $S_{(t)}$ only controlled ($S_{(t)} \not\rightarrow V1 \rightarrow D_{(t)}$) to that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($S_{(t-1)} S_{(t)} \not\rightarrow V1 \rightarrow D_{(t)}$). In this case, the contribution of $V1$ to $D_{(t)}$ must be greater when only $S_{(t)}$ is controlled than when both $S_{(t-1)}$ and $S_{(t)}$ are controlled if the contribution of $S_{(t-1)}$ to $D_{(t)}$ via $V1$ is substantial. AME was adopted instead of comparing regression coefficients because it does not suffer from the scale problem, unlike logistic and probit regression coefficients (Mize et al., 2019).

In doing so, the trial-to-trial measures of $V1$ were acquired by taking the sum of BOLDs across the eccentricity bins with the same readout weights used in the previous section (Fig. 4B). At first, we confirmed that $V1$ contains both current stimuli and adaptation signals by regressing $V1$ on to $S_{(t)}$, $S_{(t-1)}$, and $D_{(t-1)}$ concurrently for each participant (Fig. 4C). This multiple regression analysis indicates that the previously observed adaptation to $S_{(t-1)}$ (Fig. 3H,I) was still significant across participants ($\beta = -0.047$, $t_{(18)} = -2.2$, $p = 0.039$), even when we controlled the variability of $D_{(t-1)}$ ($\beta = 0.021$, $t_{(18)} = 0.82$, $p = 0.42$), a potential confounding variable.

The AME of $V1$ on $D_{(t)}$ was significant across participants ($\beta = 0.020, t_{(18)} = 2.3, p = 0.031$; Fig. 4D, the first bar). Importantly, it did not significantly decrease across participants when the influence of $S_{(t-1)}$ was controlled ($t_{(18)} = -1.6, p = 0.13$; Fig. 4D, the change of the first to second bars). Given the significant repulsive bias associated with $S_{(t-1)}$ presented on the two-back trial, we also controlled $S_{(t-2)}$ in addition to $S_{(t-1)}$. Despite this additional control, the AME of $V1$ on $D_{(t)}$ did not significantly decrease ($t_{(18)} = -1.5, p = 0.15$). By contrast, the AME of $V1$ on $D_{(t)}$ substantially decreased across participants, almost to none, when the influence of $S_{(t)}$ was controlled ($t_{(18)} = -6.0, p = 1.1 \times 10^{-5}$; Fig. 4D, the change of the first to third bars). Likewise, the AME of $V1$ on $D_{(t)}$ with $S_{(t)}$ only controlled did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 1.4, p = 0.17$; Fig. 4D, the change of the fourth to third bars), whereas the AME of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ controlled was greater than that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 6.02, p = 1.1 \times 10^{-5}$; Fig. 4D, the change of the fourth to second bars). These results coherently indicate that the contribution of the previous stimuli to $D_{(t)}$ via $V1$ is absent or negligible, which is at odds with the sensory-adaptation hypothesis.

The analyses above were conducted for $V1$ acquired at the fourth time point, where sensory adaptation was significant. However, an insignificant but substantial amount of sensory adaption occurred also at the preceding (third) time point (Fig. 3I). To check the possibility that the contribution of $S_{(t-1)}$ to $D_{(t)}$ via $V1$ might be present if $V1$ is alternatively defined, we redefined $V1$ by averaging those acquired at the third and fourth points and repeated the same AME analyses as above. However, the contribution of the previous stimuli to $D_{(t)}$ via $V1$ is still absent or negligible: the AME of $V1$ on $D_{(t)}$ did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ controlled ($t_{(18)} = -1.4$, $p = 0.19$); the AME of $V1$ on $D_{(t)}$ with $S_{(t)}$ only controlled did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 1.03, p = 0.32$).

Furthermore, the same pattern of AMEs was observed when we used two alternative readout schemes for extracting $V1$. The AME of $V1$ on $D_{(t)}$ decreased after $S_{(t)}$ was controlled (the discriminability scheme: $t_{(18)} = -5.4, p = 4.3 \times 10^{-5}$; the log likelihood scheme: $t_{(18)} = -6.0$, $p = 1.1 \times 10^{-5}$; Fig. 4E,F, the change of the first to third bars) but not after $S_{(t-1)}$ was controlled (the discriminability scheme: $t_{(18)} = -1.4, p = 0.19$; the log likelihood scheme: $t_{(18)} = -1.5$, $p = 0.14$; Fig. 4E,F, the change of the first to second bars). Likewise, the AME of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ only controlled was larger than that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled (the discriminability scheme: $t_{(18)} = 5.4$, $p = 4.0 \times 10^{-5}$; the log likelihood scheme: $t_{(18)} = 6.0, p = 1.2 \times 10^{-5}$; Fig. 4E,F, the change of the fourth to second bars), while that with $S_{(t)}$ only controlled did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled (the discriminability scheme: $t_{(18)} = 1.3$, $p = 0.22$; the log likelihood scheme: $t_{(18)} = 1.4$, $p = 0.18$; Fig. 4E,F, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of $V1$ to the current choice is ascribed mostly to the current stimulus but hardly to the previous stimuli, which is inconsistent with the sensory-adaptation hypothesis.

Repulsive bias in experiment 2

Having failed to find the evidence supporting the sensory-adaptation hypothesis in experiment 1, we conducted experiment 2 to search the whole brain for the signal representing the class boundary and to test whether that signal relates to the previous stimuli and the current choice in a manner consistent with the boundary-updating hypothesis. As mentioned earlier (see above, Experimental paradigm), the experimental procedure in experiment 2 was the same as in experiment 1, except for the fMRI protocol.

The behavioral performance in experiment 2 closely matched that in experiment 1 (Fig. 3A–C) in many aspects. The PL difference induced by the previous stimulus (−0.25) substantially differed from zero ($t_{(17)} = -7.3$, $p = 1.3 \times 10^{-6}$) indicating the existence of repulsive bias, whereas that by the previous choice (0.027) did not significantly differ from zero ($t_{(17)} = 1.3$, $p = 0.19$; Fig. 5A,B). The logistic regression analysis confirmed the significant presence of repulsive bias across participants ($S_{(t-1)}$, $\beta = -0.54$, $t_{(17)} = -7.9$, $p = 4.6 \times 10^{-7}$; $S_{(t-2)}$, $\beta = -0.24$, $t_{(17)} = -4.7, p = 2.3 \times 10^{-4}$; $D_{(t-1)}$, $\beta = 0.0055$, $t_{(17)} = 0.13, p = 0.90$; Fig. 5C).

Bayesian model of boundary-updating (BMBU)

As we identified $V1$ in experiment 1, we first need to identify the brain signal that reliably represents the class boundary. However, it is challenging to identify such signals in two aspects. First, unlike in experiment 1, where V1 was the obvious cortical region to bear the size-encoding signal susceptible to adaptation given a large volume of previous work (Kohn, 2007; Patterson et al., 2013; Morgan, 2014; Solomon and Kohn, 2014; Weber et al., 2019; Fritsche et al., 2022) and our own work (Choe et al., 2014), we have no such a priori region where the boundary-representing signal resides. This aspect requires us to explore the whole brain. Second, unlike in experiment 1, where the size variable was physically prescribed by the experimental design, we need to infer the trial-to-trial states (i.e., sizes) of the class boundary, which is an unobservable, thus latent, variable. This aspect requires us to build a model. To address these challenges, we inferred the latent state of the class boundary using a Bayesian model of boundary-updating (BMBU) and searched the whole brain for the boundary-representing signal using a searchlight multivariate pattern analysis technique.

We developed BMBU by formalizing the binary classification task in terms of Bayesian decision theory (Knill and Richards, 1996), a powerful framework for modeling human decision-making behavior under uncertainty. Binary classification is to judge whether the "ring size on the current trial t ($S_{(t)}$)" is larger or smaller than the "the typical size of rings appearing across the entire trials ($\widetilde{S}$)." Therefore, a classifier must infer them based on the measurements of stimulus size in the sensory and memory systems.

The generative model

On trial t, $S_{(t)}$ is randomly sampled from a probability distribution $p(S)$ and engenders a measurement in the sensory system $m_{(t)}$, which is a random sample from a probability distribution $p(m_{(t)}|S_{(t)})$ (Fig. 5D, bottom, black dotted curve). Critically, as i trials elapse, $m_{(t)}$ is re-encoded into a mnemonic measurement in the working-memory system $r_{(t-i)}$, which is a random sample from a probability distribution $p(r_{(t-i)}|S_{(t)})$ (Fig. 5D, bottom,

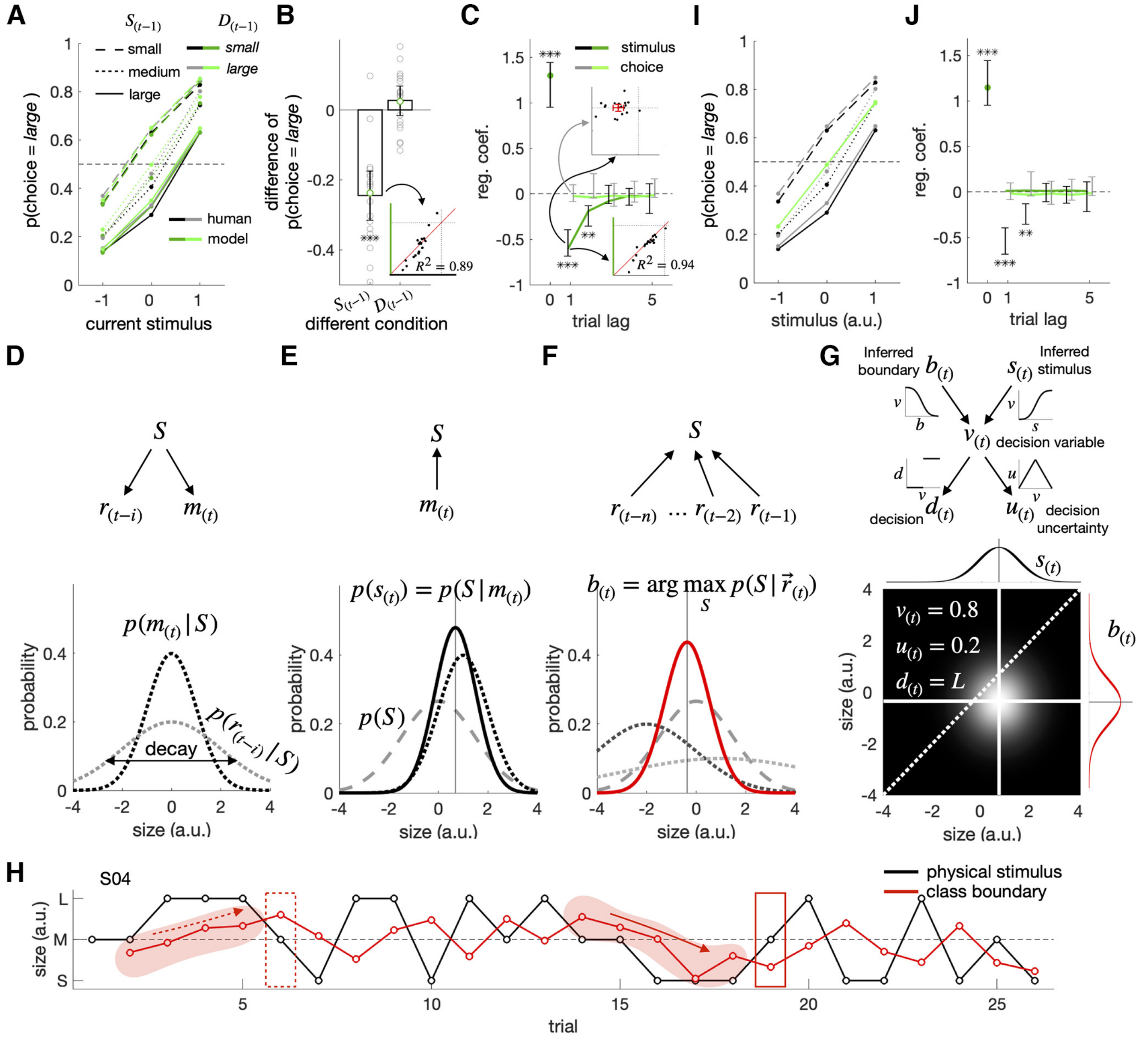

Figure 5. Repulsive bias in experiment 2 and a Bayesian model of boundary updating (BMBU). ***A–C***, Repulsive bias in psychometric curves (***A***, ***B***) and regression analysis (***C***). The formats were identical to those in the corresponding figure panels for experiment 1 (Fig. 3*A–C*), except that the *ex post* model simulation results (green lines and symbols) are added. In the bottom insets of ***B***, the observed (*x*-axis) and simulated (*y*-axis) average differences in the fractions of *large* choices between the trials in which the previous stimulus was L-ring and those in which it was S-ring are plotted against one another, where the red diagonal demarcates the identity line. In the bottom insets of ***C***, the observed (*x*-axis) and simulated (*y*-axis) regression coefficients for the previous stimulus ($S_{(t-1)}$) regressor are plotted against one another for individual observers, where the red diagonal demarcates the identity line. ***D–G***, The measurement generation (***C***), stimulus inference (***D***), class-boundary inference (***E***), and decision-variable deduction (***F***) processes of BMBU. BMBU posits that the Bayesian decision-maker has an internal causal model of how a physical stimulus size (S) engenders a current sensory measurement ($m_{(t)}$) and a retrieved memory measurement from ith preceding trial ($r_{(t-i)}$; ***D***, top), which specifies the probability distribution of $m_{(t)}$ and $r_{(t-i)}$ conditioned on S, respectively (***D***, bottom). In turn, $p\left(m_{(t)}|S\right)$ allows the Bayesian decision-maker to infer S on observing $m_{(t)}$ by combining it with the prior knowledge about S, $p(S)$, to compute the posterior probability of S given $m_{(t)}$, $p\left(S|m_{(t)}\right)$ (***E***). Similarly, $p\left(r_{(t-i)}|S\right)$ allows for inferring the class boundary ($b_{(t)}$) on retrieving the memory of previous sensory measurements ($\vec{r}_{(t)} = [r_{(t-1)}, r_{(t-2)}, \ldots]$) by combining it with $p(S)$ to compute the posterior probability of S given $\vec{r}_{(t)}$, $p\left(S|\vec{r}_{(t)}\right)$ (***F***). In ***D–F***, black dotted curves, $p\left(m_{(t)}|S\right)$; gray dotted curves, $p\left(r_{(t-i)}|S\right)$, the darker the dotted curve is, the more recent the memory is; gray dashed curves, $p(S)$; black solid curve, $p\left(S|m_{(t)}\right)$; red solid curve, $p\left(S|\vec{r}_{(t)}\right)$. Finally, the inferred stimulus, $s_{(t)}$, and the inferred class boundary, $b_{(t)}$, allow for deducing the decision variable, $v_{(t)}$, the choice variable, $d_{(t)}$, and the uncertainty variable, $u_{(t)}$ (***G***, top), as illustrated in an example bivariate distribution of $s_{(t)}$ and $b_{(t)}$, from which $v_{(t)}$, $d_{(t)}$ and $u_{(t)}$ are derived (***G***, bottom). ***H***, An example temporal trajectory of the class boundary inferred by BMBU in a single scan run of a representative subject 04. The black and red lines indicate the sizes of physical stimulus and the boundary inferred by BMBU, respectively. ***I***, ***J***, *Ex post* simulation results of the constant-boundary model. The formats are identical to those of ***A*** and ***B***.

light-gray dotted curve). Here, we assumed that the width of $p(r_{(t-i)}|S_{(t)})$ increases as i increases reflecting the working memory decay (Gorgoraptis et al., 2011; Zokaei et al., 2015).

Inferring the current stimulus size

On trial t, the Bayesian classifier infers $S_{(t)}$ by inversely propagating $m_{(t)}$ in the generative model (Fig. 5*E*, top). As a result, the inferred size ($s_{(t)}$) is defined as the value of S given $m_{(t)}$, as captured by the following equation:

$$p(s_{(t)}) = p\left(S|m_{(t)}\right), \tag{1}$$

where the width of $p\left(S|m_{(t)}\right)$ reflects the precision of $s_{(t)}$ (Fig. 5*E*, bottom).

Inferring the class boundary

On trial t, the Bayesian classifier infers the class boundary ($b_{(t)}$), i.e., the inferred value of $\widetilde{S}$, by inversely propagating a set of retrieved measurements in the working memory system $\vec{r}_{(t)} = \{r_{(t-1),}r_{(t-2),}r_{(t-3),}\cdots, r_{(t-n)}\}$ (Fig. 5*F*, top). $b_{(t)}$ is defined as the most probable value of S given $\vec{r}_{(t)}$, as captured by the following equation:

$$b_{(t)} = \arg\max_{S} p\left(S|\vec{r}_{(t)}\right) \tag{2}$$

where the width of $p\left(S|\vec{r}_{(t)}\right)$ reflects the precision of $b_{(t)}$. Notably, Equation 2 implies that $b_{(t)}$ must be attracted more to recent stimuli than to old ones because (1) the precision of working memory evidence decreases as trials elapse (Fig. 5*F*, bottom, dotted curves) and (2) the more uncertain the evidence is, the less weighed the evidence is for class-boundary inference.

Making a decision with the inferred current stimulus size and the inferred class boundary

Having estimated $s_{(t)}$ and $b_{(t)}$, the Bayesian classifier deduces a decision variable ($v_{(t)}$) from $s_{(t)}$ and $b_{(t)}$ and translating it into a binary decision ($d_{(t)}$) with a degree of uncertainty ($u_{(t)}$; Fig. 5*G*). Here, $v_{(t)}$ is the probability that $s_{(t)}$ will be greater than $b_{(t)}$ ($v_{(t)} = p\left(s_{(t)} > b_{(t)}\right)$); $d_{(t)}$ is *large* or *small* if $v_{(t)}$ is greater or smaller than 0.5, respectively; $u_{(t)}$ is the probability that $d_{(t)}$ will be incorrect ($u_{(t)} = p\left(s_{(t)} < b_{(t)} | d_{(t)} = large\right)$ or $p\left(s_{(t)} > b_{(t)} | d_{(t)} = small\right)$; Sanders et al., 2016).

In sum, BMBU models a human decision-maker as the Bayesian classifier who, over consecutive trials, continuously infers the class boundary (b) and the current stimulus size (s), deduces the decision variable (v) from s and b, and makes a decision (d) with a varying degree of uncertainty (u). As shown below, BMBU well predicts human participants' choices and reproduces their repulsive bias.

The prediction and simulation of human choices and repulsive bias by BMBU

We assessed BMBU's accountability for human behavior in the binary classification task in two aspects, comparing its (1) predictability of the choices and (2) reproducibility of repulsive bias to those of the control model which does not update the class boundary ("constant-boundary model"; see Materials and Methods).

We assessed the predictability of BMBU and the constant-boundary model by fitting them to human choices using the maximum likelihood rule (see Materials and Methods). BMBU excels over the constant-boundary model in goodness-of-fit. The average AIC difference across participants is −10.48 and was significantly less than the conventional threshold (−4; Anderson and Burnham, 2004; $t_{(17)} = -2.6$, $p = 0.020$). The variance explained by BMBU, measured by the Nagelkerke R^2, is equal to 132% of that by the constant-boundary model.

After equipping the models with their best-fit parameters, we assessed their reproducibility by making them simulate the decisions over the same sequence of ring sizes presented to the human participants (see Materials and Methods). From this simulation, we can also vividly appreciate how BMBU updates its class boundary ($b_{(t)}$) depending on the ring sizes encountered over a sequence of classification trials (Fig. 5*H*). As implied by Equation 2, BMBU continuously shifts $b_{(t)}$ toward the ring sizes shown in previous trials. Such attractive shifts are pronounced especially when streaks of S-ring (Fig. 5*H*, solid arrow) or L-ring (Fig. 5*H*, dashed arrow) appeared over trials. Importantly, we confirmed that such boundary-updating of BMBU reproduces the repulsive bias displayed by the human participants with a remarkable level of resemblance across participants, both for the psychometric curves (the R^2 of the effect of previous stimulus on PL between humans and BMBU was 0.89; Fig. 5*A*,*B*) and for the coefficients of the stimulus and choice regressors (the R^2 of coefficients of the immediately preceding stimulus between humans and BMBU was 0.94; Fig. 5*C*). None of the simulated PLs and coefficients, a total of 17 points, fell outside the 95% confidence intervals of the corresponding human PLs and coefficients. Not surprisingly, the constant-boundary model failed to show any slightest hint of repulsive bias (Fig. 5*I*,*J*). Although we used m-sequences to prevent any auto-correlation among ring sizes, the failure of the constant-boundary model in reproducing repulsive bias reassures that the actual stimulus sequences used in the experiment do not contain any unwanted statistics that might induce spurious kinds of repulsive bias.

In sum, BMBU's inferences of the class boundary based on past stimuli accounted for a substantive fraction of the choice variability of human classifiers and successfully captured their repulsive bias.

Brain signals of the class boundary and the other latent variables

In the previous section, we demonstrated that BMBU accounted well for the variability of human choices and successfully reproduced the observed repulsive bias. However, such correspondences between the humans' and the models' choices do not necessarily warrant the validity of our procedure of estimating the latent states of the model variables (b, s, and v), which is crucial in testing the boundary-updating hypothesis. To validate our estimation procedure, we tested whether it could accurately recover the true states of the model variables based on the synthetic datasets simulated with 256 ground-truth model parameter sets (see the Materials and Methods). The recovered states of the model variables well matched the corresponding true states (R^2 = 0.98±0.0044, 0.96±0.0073, and 0.96±0.0040 for b, s, and v, respectively; mean±95% confidence interval), which ascertains the validity of our procedure of estimating the latent states of the model variables.

Then, with the trial-to-trial states of the simulated latent variables, we identified the brain signals of those variables with the following rationale and procedure. On any given trial t, a classifier makes a decision in the manner constrained by the causal structure of BMBU (Fig. 5*G*). This causal structure implies two important points to be considered when identifying the neural representations of b, s and v. First, for any cortical activity, its significant correlation with the variable of interest does not necessarily imply that it represents that variable per se but is open to the possibility that it may represent the other variables that are associated with the variable of interest. Second, if any given cortical activity represents the variable of interest, that activity must not violate any of its relationships with the other variables that are implied by the causal structure (Table 1; see Materials and Methods).

We incorporated these two points in our search of the brain signals of b, s and v, as follows. Initially, we identified the candidate brain signals of b, s, and v by localizing the patterns of activities that closely reflect the trial-to-trial states of b, s, and v. For localization, we used the support vector regressor decoding with the searchlight technique (Kahnt et al., 2011a; Hebart et al., 2016), which is highly effective in detecting the local patterns of population fMRI responses associated with the latent variables of computational models (Kriegeskorte et al., 2006). Next, we put those candidate brain signals to a strong test of whether their trial-to-trial states satisfy the causal relationships with the other variables. Specifically, we converted those causal relationships into the empirically testable sets of regression models (Table 1), respectively for b (14 regressions), s (14 regressions), and v (17 regressions) and checked whether all the regressors' coefficients derived from the brain signals were consistent with the regression models (see Materials and Methods). In what follows, we will describe how the regression tests for the brain signal of b (y_b) were derived from the causal structure of the variables defined by BMBU (see Materials and Methods for those for the two remaining variables s and v).

According to the causal relationship of b with the latent variables, y_b must satisfy the following single linear regression models: y_b must be positively regressed onto b (#1) and be so even when the false discovery rate (Benjamini and Hochberg, 1995) is applied (#2); y_b must be positively regressed onto b even when b is orthogonalized to v (#3) or d (#4) because y_b should reflect the variance irreducible to the offspring variables of b; y_b must not be regressed onto s because b and s are independent of one another ($b \nleftrightarrow s$; Fig. 5*G*, #5); y_b must be negatively regressed onto v ($b \rightarrow v$; Fig. 5*G*, #6) but not when v is orthogonalized to b because such orthogonalization removes the influence of b on v (#7); y_b must be negatively regressed onto d ($b \rightarrow v \rightarrow d$; Fig. 5*G*, #8) but not onto u because u is not linearly correlated with b ($b \rightarrow v \rightarrow u$ is blocked by the nonlinear relationship between u and v, Fig. 5*G*, #9). In addition, according to the causal relationship of the latent variables with the stimuli and choices (Fig. 5*D–G*), y_b must satisfy the following multiple linear regression model defined by the observable variables $\left[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t-1)}, D_{(t-2)}\right]$: y_b must not be regressed onto the current stimulus (#10) because b is independent of $S_{(t)}$; y_b must be positively regressed onto the 1-back stimulus for sure (#11) because b firmly shifts toward $S_{(t-1)}$; the regression of y_b onto the two-back stimulus must be weaker than that onto the 1-back stimulus (#12) because of memory decay (Fig. 5*D*; accordingly, the sign of the regression coefficient of $S_{(t-2)}$ was defined as the complementary part of that of $S_{(t-1)}$); y_b must not be regressed onto previous decisions because previous decisions do not have any influence on b (#13, 14). We did not include $D_{(t)}$ as a regressor in the multiple regression because $D_{(t)}$ may induce a spurious correlation between b and s by controlling the collider (common offspring) variable v (Elwert and Winship, 2014; $b \rightarrow v \leftarrow s$; Fig. 5*G*) via its relationship with v ($v \rightarrow d$; Fig. 5*G*).

As a result, the brain signals that survived the exhaustive regression tests clustered in six separate regions (Fig. 6; Table 2). The signal of b appeared in three separate regions at different time points relative to stimulus onset, a region in the left inferior parietal lobe at 1.1s (IPL$_{\text{b1}}$) and two regions in the left posterior superior temporal gyrus at 3.3 and 5.5 s (pSTG$_{\text{b3}}$, pSTG$_{\text{b5}}$). The signal of s appeared in the left dorsolateral prefrontal cortex at 3.3 s (DLPFC$_{\text{s3}}$) and in the right cerebellum at 5.5 s (Cereb$_{\text{s5}}$). The signal of v appeared in the left anterior superior temporal gyrus at 5.5 s (aSTG$_{\text{v5}}$). To ascertain the robustness of the neural representations of the latent variables in these six areas, we repeated the searchlight decoding analysis using a different searchlight size (87 voxels, which is smaller than the original one, 123 voxels). Despite the change in searchlight size, we could detect the clusters that survived all regression tests around the six regions (Table 2).

Lastly, we investigated whether the probable causal structures between the brain signals of b, s, and v are consistent with BMBU in the following two critical aspects. First, the brain signal of v should be concurrently affected by the brain signals of b and s: $b \rightarrow v \leftarrow s$. Second, there should be no causal connection between b and s because BMBU is built on the assumption that b and s are independent of one another (i.e., b and s are biased by previous and current stimuli, respectively): $b \nleftrightarrow s$ (Fig. 5*G*). To examine these aspects, we investigated all of the three-node networks ($N = 162$) composed of the brain signals of b, s, and v, and calculated their Bayesian Information Criterion (BIC; see Materials and Methods).

The outcomes of BIC evaluation were consistent with BMBU. First, out of the 162 possible causal graphs, the smallest (best) BIC value was found for "pSTG$_{\text{b5}}$→aSTG$_{\text{v5}}$ ←Cereb$_{\text{s5}}$" (Fig. 7). Second, We found that any graph with the causal arrows between $b_{(t)}$ and $s_{(t)}$ is significantly less likely than the best causal graph (BIC $>$ 2; shown at the bottom of Fig. 7; Kass and Raftery, 1995). The results indicate that the relationship between the identified brain signals faithfully reflects the causal relationship of the latent variables implied by BMBU.

The variability of the class-boundary brain signals associated with previous stimuli contributes to the variability of choice

Finally, with the brain signals that represent the class boundary (IPL$_{\text{b1}}$, pSTG$_{\text{b3}}$, and pSTG$_{\text{b5}}$) in our hands, we verified the boundary-updating hypothesis with the rationale and analysis identical to those for the verification of the sensory-adaptation hypothesis.

We stress that the respective associations of the brain signal of b with the previous stimulus ($S_{(t-1)}$; Table 1, eleventh row) and with the variable d (Table 1, eighth row) do not necessarily imply that the variability of the brain signal of b that is associated with $S_{(t-1)}$ contributes to the choice variability (as implied by the causal information flows through b depicted in Fig. 8*A*), for the same reasons mentioned when verifying the sensory-adaptation hypothesis. To verify such contribution, we need to compare the AME of the brain signals of b on the current choice ($D_{(t)}$; pSTG$_{\text{b5}}$→ $D_{(t)}$) to the AME of the brain signals of b on $D_{(t)}$ with $S_{(t-1)}$ controlled ($S_{(t-1)} \nrightarrow$pSTG$_{\text{b5}}$→ $D_{(t)}$).

As anticipated, the AME of pSTG$_{\text{b5}}$ on $D_{(t)}$ was negatively significant across participants ($t_{(17)} = -4.8, p = 1.7 \times 10^{-4}$; Fig. 8*B*, the first bar). Importantly, unlike the size-encoding signal in V1, the negative AME significantly weakened across participants when the contribution of $S_{(t-1)}$ was controlled ($t_{(17)} = 2.8, p = 0.012$; Fig. 8*B*, the change of the first to second bars). On the other hand, controlling $S_{(t)}$ did not affect the AME of pSTG$_{\text{b5}}$ on $D_{(t)}$ at all ($t_{(17)} = 0.29, p = 0.77$; Fig. 8*B*, the change of the first to third bars), which is consistent with the absence of the contribution of $S_{(t)}$ on b in the causal relationship defined by BMBU (Fig. 5*G*). Likewise, the null effect of $S_{(t)}$ on the AMEs of pSTG$_{\text{b5}}$ on $D_{(t)}$ was confirmed by the insignificant difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($t_{(17)} = -0.31, p = 0.77$; Fig. 8*B*, the change of the fourth to second bars). Also, the effect of $S_{(t-1)}$ on the AMEs of pSTG$_{\text{b5}}$ on $D_{(t)}$ was confirmed by the significant difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($t_{(17)} = -2.7, p = 0.014$; Fig. 8*B*, the change of the fourth to third bars).

The same patterns were also observed for IPL$_{\text{b1}}$ and pSTG$_{\text{b3}}$ (Fig. 8*C*,*D*). Especially, the AMEs of pSTG$_{\text{b3}}$ and IPL$_{\text{b1}}$ on $D_{(t)}$ both weakened after controlling $S_{(t-1)}$ (pSTG$_{\text{b3}}$: $t_{(17)} = 2.2, p = 0.046$;

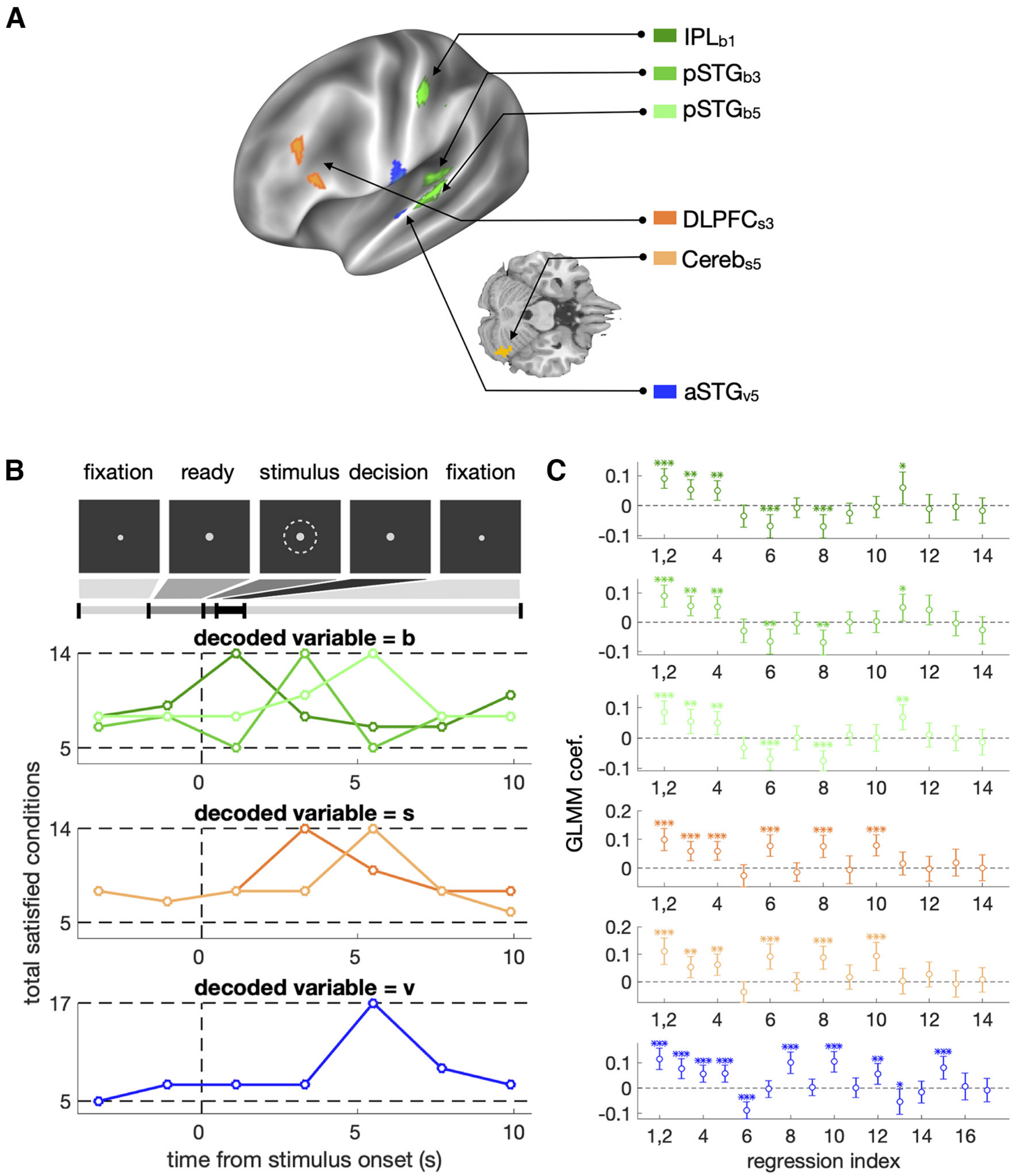

Figure 6. Brain signals of the latent variables of BMBU. ***A***, Loci of the brain signals. The brain regions where BOLD activity patterns satisfied all the regressions implied by the causal structure of the variables in BMBU are overlaid on the inflated cortex and the axial view of the cerebellum of the template brain. ***B***, Within-trial time courses of the satisfied regressions in number. The within-trial task phases are displayed (top panel) to help appreciate when the brain signals become pronounced, with the hemodynamic delay (4–5 s) in BOLD (bottom three panels). ***C***, The coefficients and the 95% CIs of the generalized linear mixed effect model (GLMM) of the decoded variable averaged across the searchlights of each ROI on the time points on which each ROI was detected. The regression index indicates the index specified in Table 1. ***B***, ***C***, The colors of the symbols and lines correspond to those of the brain regions shown in ***A***. Asterisks indicate the statistical significance (*P,0:05, **P,0:01, ***P,0:001). The 95% CIs of the mean across participants are indicated by the vertical error bars.

Fig. 8*D*, the change of the first to second bars; IPL$_{b1}$: $t_{(17)} = 2.1, p = 0.0503$; Fig. 8*C*, the change of the first to second bars), but not after controlling $S_{(t)}$ (pSTG$_{b3}$: $t_{(17)} = -0.57, p = 0.58$; Fig. 8*D*, the change of the first to third bars; IPL$_{b1}$: $t_{(17)} = 0.22, p = 0.83$; Fig. 8*C*, the change of the first to third bars). The null effect of $S_{(t)}$ was confirmed by the insignificance difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled (pSTG$_{b3}$: $t_{(17)} = 0.43, p = 0.67$, Fig. 8*D*; IPL$_{b1}$: $t_{(17)} = -0.41$, $p = 0.69$, Fig. 8*C*, the change of the fourth to second bars). Also, the effect of $S_{(t-1)}$ was confirmed by the significant or marginally significant differences between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled (pSTG$_{b3}$: $t_{(17)} = -1.9, p = 0.081$, Fig. 8*D*, the change of the fourth to third bars; IPL$_{b1}$: $t_{(17)} = -2.2, p = 0.045$, Fig. 8*C*, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of the class boundary to the current choice is significantly ascribed to the previous stimuli supporting the boundary-updating hypothesis on repulsive bias.

Having found the evidence supporting the boundary-updating hypothesis in the brain signals of *b*, we also conducted the

Table 2. Specification of the brain signals of the latent variables of BMBU

Name	Cortical area	Decoded variable	Detected time from stimulus onset (s)	Contiguous searchlights number	Peak searchlight MNI coordinate	GLMM p-value (right-tailed)
IPL$_{b1}$	Left inferior parietal lobe	$b_{(t)}$	1.1	15 (10)	$[-54, -27, 48]$ $([-54, -27, 48])$	8.8×10^{-8} (7.4×10^{-7})
pSTG$_{b3}$	Left posterior superior temporal gyrus	$b_{(t)}$	3.3	13 (7)	$[-45, -30, 9]$ $([-54, -27, 12])$	5.8×10^{-7} (8.6×10^{-8})
pSTG$_{b5}$	Left posterior superior temporal gyrus	$b_{(t)}$	5.5	18 (14)	$[-66, -21, 9]$ $([-66, -21, 9])$	2.3×10^{-7} (2.6×10^{-7})
DLPFC$_{s3}$	Left dorsolateral prefrontal cortex	$s_{(t)}$	3.3	33 (37)	$[-51, 27, 24]$ $([-51, 27, 24])$	1.9×10^{-7} (7.7×10^{-6})
Cereb$_{s5}$	Right cerebellum	$s_{(t)}$	5.5	36 (19)	$[36, -63, -21]$ $([36, -69, -18])$	1.4×10^{-6} (5.0×10^{-7})
aSTG$_{v5}$	Left anterior superior temporal gyrus	$v_{(t)}$	5.5	15 (4)	$[-60, -9, 15]$ $([-60, -9, 15])$	4.9×10^{-8} (3.7×10^{-7})

The results outside of the parentheses indicate the main result obtained by using the searchlight composed of 123 voxels. The values inside of the parentheses are the results calculated by using different size of searchlight (87 voxels).

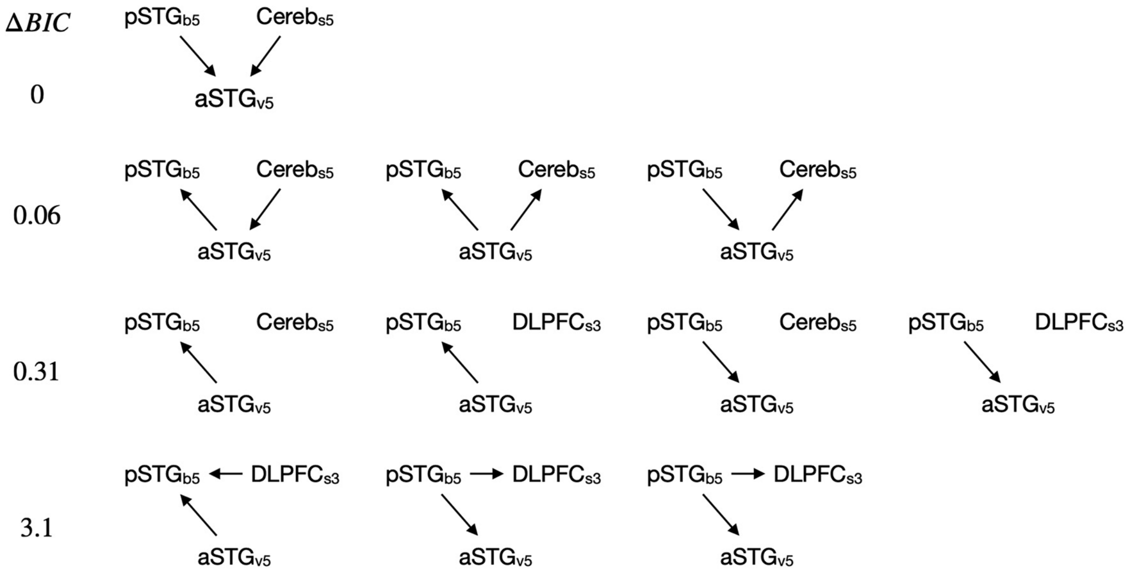

Figure 7. The probable causal structures between the brain signals of the latent variables in BMBU. For each row, the value in the left indicates the relative BIC scores of the causal structures in reference to the most probable one at the top.

same AME analysis on the signals of s and v below. Given the causal structure of b, s, and v, the validity of the boundary-updating hypothesis will be reinforced if the brain signals of s and v also turn out acting as fulfilling their causal roles defined by BMBU. According to BMBU, the contribution of s to $D_{(t)}$ must originate not from $S_{(t-1)}$ but from the $S_{(t)}$ (the causal route indicated by the solid arrows in Fig. 8*A*). In line with this implication, the AMEs of DLPFC$_{s3}$ and Cereb$_{s5}$ on $D_{(t)}$ were both significant across participants ($t_{(17)} = 3.8$, $p = 0.0014$ for DLPFC$_{s3}$; $t_{(17)} = 3.3$, $p = 0.0041$ for Cereb$_{s5}$; Fig. 8*E*,*F*, first bars) and significantly decreased after controlling $S_{(t)}$ ($t_{(17)} = -4.4$, $p = 4.1 \times 10^{-4}$ for DLPFC$_{s3}$; $t_{(17)} = -3.7, p = 0.0019$ for Cereb$_{s5}$; Fig. 8*E*,*F*, the change of the first to third bars) but not after controlling $S_{(t-1)}$ ($t_{(17)} = 1.2$, $p = 0.26$ for DLPFC$_{s3}$; $t_{(17)} = 0.69, p = 0.50$ for Cereb$_{s5}$; Fig. 8*E*,*F*, the change of the first to second bars). Likewise, the AMEs of DLPFC$_{s3}$ and Cereb$_{s5}$ on $D_{(t)}$ with $S_{(t-1)}$ controlled were both larger than those with both $S_{(t)}$ and $S_{(t-1)}$ controlled ($t_{(17)} = 4.3, p = 0.0050$ for DLPFC$_{s3}$; $t_{(17)} = 3.8$, $p = 0.0016$for Cereb$_{s5}$; Fig. 8*E*,*F*, the change of the fourth to second bars), whereas the AMEs of DLPFC$_{s3}$ and Cereb$_{s5}$ on $D_{(t)}$ with $S_{(t)}$ controlled did not differ from those with both $S_{(t)}$ and $S_{(t-1)}$ controlled ($t_{(17)} = -0.92, p = 0.37$ for DLPFC$_{s3}$; $t_{(17)} = -0.057$, $p = 0.96$ for Cereb$_{s5}$; Fig. 8*E*,*F*, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of the inferred stimulus to the current choice is significantly ascribed to the current but not to the previous stimuli supporting the boundary-updating hypothesis.

On the contrary, the contribution of v to $D_{(t)}$ must originate not only from $S_{(t-1)}$ but also from $S_{(t)}$ (Fig. 8*A*). In line with this implication, the AME of aSTG$_{v5}$ on $D_{(t)}$ was significant ($t_{(17)} = 5.1, p = 9.7 \times 10^{-5}$; Fig. 8*G*, the first bar) and significantly decreased both after controlling $S_{(t-1)}$ ($t_{(17)} = -2.8, p = 0.012$; Fig. 8*G*, the change of the first to second bars) and after controlling $S_{(t)}$ ($t_{(17)} = -4.1, p = 7.5 \times 10^{-4}$; Fig. 8*G*, the change of the first to third bars). Likewise, the AME of aSTG$_{v5}$ on $D_{(t)}$ with controlled both $S_{(t)}$ and $S_{(t-1)}$ significantly increased both after controlling $S_{(t-1)}$ ($t_{(17)} = 4.1, p = 6.7 \times 10^{-4}$; Fig. 8*G*, the change of the fourth to second bars) and after controlling $S_{(t)}$ ($t_{(17)} = 2.8, p = 0.012$; Fig. 8*G*, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of the decision variable to the current choice is significantly ascribed to both current and the previous stimuli supporting the boundary-updating hypothesis.

On a separate note, the six loci of the brain signals of b,s, and d were defined by applying the conservative criterion that any given cluster satisfying all the regression tests (Table 1) should be the same or larger than 12. We note that there was a focal region in the right-hemisphere medial visual cortex that survived the regression tests for $s_{(t)}$ on the 3 s after stimulus onset (VC$_{s3}$) but failed to reach the threshold size (N voxels = 6).

To examine the neural loci of the inferred stimulus further, we checked the possibility that VC$_{s3}$ might carry the signal via

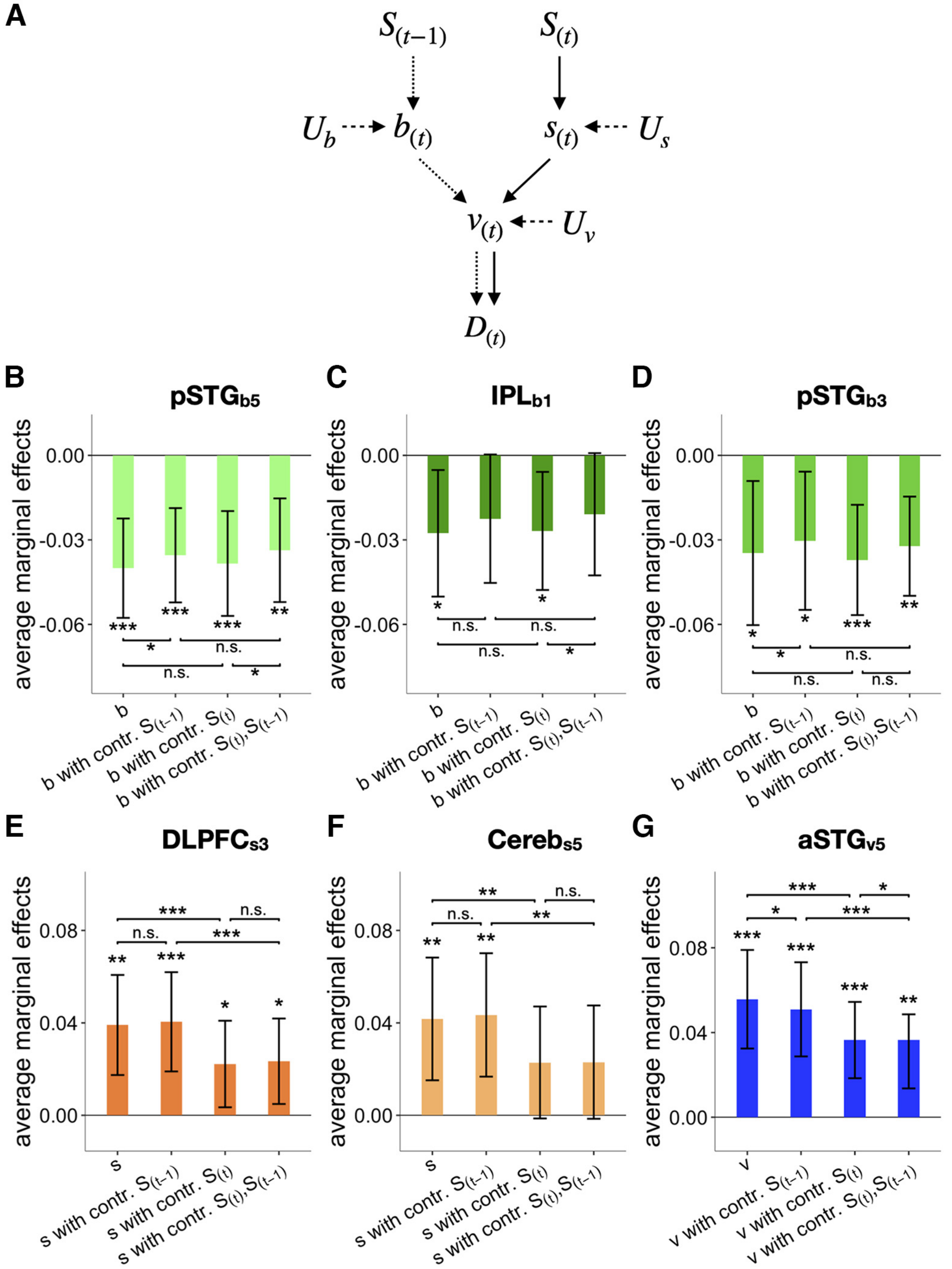

Figure 8. Origin of the covariation between the current choice and the brain signals of the latent variables in BMBU. ***A***, The causal structure of the variables implied by the boundary-updating hypothesis. The brain signal of the decision variable ($v_{(t)}$) is influenced by the brain signal of the inferred class criterion ($b_{(t)}$), brain signal of the inferred stimulus ($s_{(t)}$), and the unknown sources (U_v). In turn, $b_{(t)}$ is influenced by the previous stimulus ($S_{(t-1)}$) and the unknown sources (U_b) whereas $s_{(t)}$ is influenced by the current stimulus ($S_{(t)}$) and the unknown sources (U_s). Lastly, $v_{(t)}$ influences the current choice ($D_{(t)}$). If the boundary-updating hypothesis is true, part of the causal influence of $b_{(t)}$ on $D_{(t)}$ must originate from $S_{(t-1)}$, as indicated by the connected chain of the dotted arrows. ***B–G***, The average marginal effects (AMEs) of the brain signals on $D_{(t)}$, with the brain signals of $b_{(t)}$ from pSTG$_{b5}$ (***B***), IPL$_{b1}$ (***C***), and pSTG$_{b3}$ (***D***), $s_{(t)}$ from DLPFC$_{s3}$ (***E***), and Cereb$_{s5}$ (***F***), and $v_{(t)}$ from aSTG$_{v5}$ (***G***). In each panel, the influences of the given brain signal on $D_{(t)}$ that can be ascribed to $S_{(t-1)}$ and $S_{(t)}$ were assessed by checking (1) whether the AME of the given brain signal on $D_{(t)}$ (left) is significantly reduced or not after controlling the influence of $S_{(t-1)}$ (second from the left) and $S_{(t)}$ (second from the right), respectively, or (2) whether the AME of *V*1 on $D_{(t)}$ controlling the influence of both $S_{(t-1)}$ and $S_{(t)}$ (right) significantly increased or not after only controlling the influence of $S_{(t)}$ (second from the right) and $S_{(t-1)}$ (second from the left), respectively. The colors of the bars correspond to those of the brain regions shown in Figure 6*A*. Asterisks indicate the statistical significance (*$P<0.05$, **$P<0.01$, ***$P<0.001$), and "n.s." stands for the nonsignificance of the test. The 95% CIs of the mean across participants are indicated by the vertical error bars.

which the current stimulus ($S_{(t)}$) contributes to the current choice ($D_{(t)}$). The AME of VC$_{s3}$ on $D_{(t)}$ was significant ($t_{(17)} = 3.0, p = 0.0074$), but no longer when $S_{(t)}$ was controlled ($t_{(17)} = 1.6$, $p = 0.14$), which indicates that the noise variability of VC$_{s3}$ is not tightly linked to the variability of the current choice. However, the AME of DLPFC$_{s3}$ on $D_{(t)}$ with $S_{(t)}$ controlled was significant ($t_{(17)} = 2.5, p = 0.023$; Fig. 8*E*, the third bar) and that of Cereb$_{s5}$ was marginally significant ($t_{(17)} = 2.0, p = 0.063$; Fig. 8*F*, the third bar). The results indicate that DLPFC$_{s3}$ and Cereb$_{s5}$ carry the signal via which the current stimulus ($S_{(t)}$) contributes to the current choice ($D_{(t)}$), whereas such contribution is not evident for VC$_{s3}$.

Furthermore, to test the sensory-adaptation hypothesis, we examined whether VC_{s3} carries the stimulus signal via which the previous stimulus ($S_{(t-1)}$) contributes to the current choice ($D_{(t)}$). However, the AME of VC_{s3} on $D_{(t)}$ did not decrease when the contribution of $S_{(t-1)}$ was controlled ($t_{(17)} = 0.28, p = 0.78$). Likewise, the AME of VC_{s3} on $D_{(t)}$ with $S_{(t)}$ controlled did not differ from that with both $S_{(t-1)}$ and $S_{(t)}$ controlled ($t_{(17)} = 0.70$, $p = 0.49$). These results corroborate the AME analyses on $V1$ in experiment 1 (Fig. 4*D*–*F*), confirming that the previous stimulus is unlikely to contribute to the current choice via the stimulus-related signals in the early visual cortex.

In sum, the results suggest that neural signals of *b* and *s* transferred previous and current stimuli to current decisions, respectively, and the neural signal of *v* transferred both previous and current stimuli to current decisions as BMBU implies, which is consistent with the boundary-updating hypothesis.

Discussion

Here, we explored the two possible origins of repulsive bias, sensory-adaptation versus boundary-updating, in binary classification tasks. Although $V1$ adapted to the previous stimulus, its variability associated with the previous stimulus failed to contribute to the choice variability. By contrast, the variability associated with the previous stimulus in the boundary-representing signals in IPL and pSTG contributed to the choice variability. These results suggest that the repulsive bias in binary classification is likely to arise as the internal class boundary continuously shifts toward the previous stimulus.

Dissociation between sensory-adaptation in V1 and repulsive bias

What makes sensory-adaptation a viable origin of repulsive bias is not its mere presence but its contribution to repulsive bias. The presence of sensory-adaptation in V1 has been firmly established (Clifford et al., 2007; Kohn, 2007; Solomon and Kohn, 2014; Weber et al., 2019) and is the necessary premise for the sensory-adaptation hypothesis to work. What matters is whether the trial-to-trial variability of V1 because of such adaptation exerts its influence on the current choice. Such an influence was not observed in our data.

From a general perspective, our findings demonstrate a dissociation between the impact of previous decision-making episodes on the sensory-cortical activity and the contribution of that sensory-cortical activity to decision-making behavior. In this regard, V1 in the current work acts like the binocular-disparity-encoding signal of V2 neurons in a recent single-cell study on monkeys (Lueckmann et al., 2018), where, despite the impact of the history on V2 activity, the variability of V2 activity associated with the history failed to contribute to the history effects on decision-making behavior. Similarly, our findings also echo the failure of the sensory-adaptation of V1 in influencing the visual orientation estimation in an fMRI study on human participants (Sheehan and Serences, 2022). There, while sensory-adaptation was evident along the hierarchy of visual areas including V1, V2, V3, V4, and interaparietal sulcus (IPS), the history effect of the previous stimulus on the current estimation behavior was opposite to that expected from sensory-adaptation, which suggests that a downstream mechanism compensates for sensory-adaptation. Such a mechanism was also called for when the single-cell-recording work on monkeys tried to explain their intriguing adaptation effects found along the visual processing hierarchy (McLelland et al., 2009). For instance, static visual stimuli engendered prolonged, on the order of tens of seconds, adaptation in the lateral geniculate nucleus but the adaptation in V1 was paradoxically short-lived, on the order of 100 ms.

The representations of the class boundary in IPL and pSTG

To account for the repulsive bias in binary classification, previous studies proposed descriptive models based on the common idea that the internal boundary continuously shifts toward the previous stimuli (Treisman and Williams, 1984; Lages and Treisman, 1998, 2010; Dyjas et al., 2012; Raviv et al., 2014; Norton et al., 2017; Hachen et al., 2021). However, the neural concomitant of class-boundary updating has rarely been demonstrated.

To our best knowledge, this issue has so far been addressed by one fMRI work (White et al., 2012); which reported the class-boundary signal in the left inferior temporal pole. However, several aspects of this work make it hard to consider the reported brain signal to represent the class boundary inducing repulsive bias. First, they experimentally manipulated the class boundary in a block-by-block manner. Thus, it is unclear whether the reportedly boundary-representing signal was updated by previous stimuli trial-to-trial, which is required to induce repulsive bias. Second, the class boundary size correlated with the average stimulus size block-by-block in their experiments. Because of this confounding factor, one cannot rule out the possibility that the reported brain signal reflects the sensory signal associated with the average stimulus size induced by the current stimulus. By contrast, the brain signal of the class boundary in our work is free from these methodological limitations, because it is updated on a trial-to-trial basis and survived the rigorous set of tests, including those addressing possible confounding variables (Table 1). In this sense, the current work can be considered the first demonstration of the brain signals representing the class boundary that is dynamically updated in such a way that it can account for repulsive bias.

We emphasize that we developed BMBU to infer the trial-to-trial latent states of the class boundary used by human observers for the purpose of verifying the boundary-updating hypothesis on repulsive bias. In this sense, BMBU should not be taken as a unified account of the history effects reported by previous studies. For example, BMBU does not account for the influence of previous decisions on subsequent decision-making, another significant contributor to the history effects (Akaishi et al., 2014; Urai and Donner, 2022). To be sure, we are open to the possibility that there might be a unified mechanism relating the previous, and current, as well, stimuli and previous decisions to the current decision in an integrative manner. To incorporate the previous decisions into such a unified mechanism, it is important to distinguish the influence of the previous choice from that of the previous motor response, which we could not do in the current work because choices and motor responses covaried. In this regard, the weak but significant negative regression coefficient of the previous decision in experiment 1 (Fig. 3*C*) could have been reflective of the influence of the previous motor response, as previously suggested (Zhang and Alais, 2020).

The representations of *inferred* stimuli in DLPFC and cerebellum

The brain signals of the inferred ring size ($s_{(t)}$) in dorsolateral prefrontal cortex (DLPFC) and cerebellum share many features with $V1$ in that their covariation with the current choice did not decrease after controlling the previous stimulus but decreased after controlling the current stimulus (Figs. 4*D*–*F*, 8*E*,*F*). This commonality suggests that DLPFC, cerebellum, and $V1$ alike route the flow of information originating from the current stimulus. Then, what made $V1$ ineligible for the brain signal of $s_{(t)}$?

It is notable that BMBU treats $s_{(t)}$ as the random variable that has the noise variability in addition to being influenced by the physical stimulus (Fig. 8*A*). This means that the brain signal of $s_{(t)}$ is supposed to be associated with the choice even when the current stimulus was controlled because the noise variability can also influence the current choice, as captured by the concept of "choice probability" (Macke and Nienborg, 2019). However, unlike DLPFC and cerebellum, the AME of $V1$ on the current choice disappeared after controlling the current stimuli, which disqualifies $V1$ as the brain signal of $s_{(t)}$. In line with this, the AME of VC_{s3} on $D_{(t)}$ also disappeared after $S_{(t)}$ was controlled in experiment 2, which again disqualifies VC_{s3} as the valid brain signal of $s_{(t)}$.

The residence of the inferred, i.e., subjective or perceived, stimulus representation in DLPFC and cerebellum, instead of the visual cortex, seems consistent with previous reports. DLPFC and cerebellum have been well known for their critical involvement in visual awareness (Gao et al., 1996; Rees et al., 2002; Dehaene and Changeux, 2011; Lau and Rosenthal, 2011; Baumann et al., 2015). By contrast, the visual cortex is likely to be involved more in a faithful representation of physical input than its subjective representation (Renart and Machens, 2014), consistent with the previous findings of our group (Lee et al., 2007; Choe et al., 2014).

The representation of the decision variable in aSTG

Whereas previous single-cell studies have reported that the decision variable is represented in the prefrontal cortex (Kim and Shadlen, 1999; Hanks et al., 2015; Hebart et al., 2016), we identified the brain signal of v only in aSTG but not in PFC. This inconsistency may reflect the poor spatial and temporal resolution of fMRI measurements. For example, if any given signal of interest is encoded in the sequential or dynamical activity patterns across a neural population, as recently demonstrated theoretically (Orhan and Ma, 2019) or empirically (Wutz et al., 2018), such signals cannot be decoded from fMRI responses. Alternatively, the inconsistency may have been a result of the previous studies not taking into account the history effect in defining the decision variable, in contrast to our study which did, given the prevalence of diverse history effects in various decision-making tasks (Fründ et al., 2014; Lak et al., 2020). In this scenario, the brain signal of the inferred stimulus in DLPFC in our study hints at the possibility that the previously reported decision variable signal in PFC could have reflected the inferred stimulus, which is closely associated with the decision variable when the decision boundary is assumed to be fixed (Gold and Shadlen, 2007). Understanding the functional role of DLPFC in perceptual decision-making seems to require further future studies, especially those in which the history effects are considered in decision variable definition while neural responses are probed at a sufficiently high spatiotemporal resolution.

References

Akaishi R, Umeda K, Nagase A, Sakai K (2014) Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron 81:195–206.

Anderson D, Burnham K (2004) Model selection and multi-model inference. New York: Springer.

Ashburner J (2007) A fast diffeomorphic image registration algorithm. Neuroimage 38:95–113.

Baumann O, Borra RJ, Bower JM, Cullen KE, Habas C, Ivry RB, Leggio M, Mattingley JB, Molinari M, Moulton EA, Paulin MG, Pavlova MA, Schmahmann JD, Sokolov AA (2015) Consensus paper: the role of the cerebellum in perceptual processes. Cerebellum 14:197–220.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.

Bierwisch M (1989) The semantics of gradation. In Manfred Bierwisch, Ewald Lang (eds.), Dimensional adjectives. Grammatical structure and conceptual interpretation, 71–261. Berlin, etc: Springer.

Bosch E, Fritsche M, Ehinger BV, de Lange FP (2020) Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias. J Vis 20:9.

Bromiley P (2003) Products and convolutions of Gaussian probability density functions. Tina-Vision Memo 3:1.

Buracas GT, Boynton GM (2002) Efficient design of event-related fMRI experiments using M-sequences. Neuroimage 16:801–813.

Carlson JM, Foti D, Mujica-Parodi LR, Harmon-Jones E, Hajcak G (2011) Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: a combined ERP and fMRI study. Neuroimage 57:1608–1616.

Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280:747–749.

Cavanagh JF, Frank MJ (2014) Frontal theta as a mechanism for cognitive control. Trends Cogn Sci 18:414–421.

Choe KW, Blake R, Lee S-H (2014) Dissociation between neural signatures of stimulus and choice in population activity of human V1 during perceptual decision-making. J Neurosci 34:2725–2743.

Clifford CW, Webster MA, Stanley GB, Stocker AA, Kohn A, Sharpee TO, Schwartz O (2007) Visual adaptation: neural, psychological and computational aspects. Vision Res 47:3125–3131.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Dehaene S, Changeux J-P (2011) Experimental and theoretical approaches to conscious processing. Neuron 70:200–227.

Dyjas O, Bausenhart KM, Ulrich R (2012) Trial-by-trial updating of an internal reference in discrimination tasks: evidence from effects of stimulus order and trial sequence. Atten Percept Psychophys 74:1819–1841.

Elwert F, Winship C (2014) Endogenous selection bias: the problem of conditioning on a collider variable. Annu Rev Sociol 40:31–53.

Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky E-J, Shadlen MN (1994) fMRI of human visual cortex. Nature 369:525–525.

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI time-series. Magn Reson Med 35:346–355.

Fritsche M, Solomon SG, de Lange FP (2022) Brief stimuli cast a persistent long-term trace in visual cortex. J Neurosci 42:1999–2010.

Fründ I, Wichmann FA, Macke JH (2014) Quantifying the effect of intertrial dependence on perceptual decisions. J Vis 14:9.

Gao JH, Parsons LM, Bower JM, Xiong J, Li J, Fox PT (1996) Cerebellum implicated in sensory acquisition and discrimination rather than motor control. Science 272:545–547.

Gibson JJ, Radner M (1937) Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. J of experimental psychology 20:453–467.

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574.

Gorgoraptis N, Catalao RF, Bays PM, Husain M (2011) Dynamic updating of working memory resources for visual objects. J Neurosci 31:8502–8511.

Grinband J, Hirsch J, Ferrera VP (2006) A neural representation of categorization uncertainty in the human brain. Neuron 49:757–763.

Hachen I, Reinartz S, Brasselet R, Stroligo A, Diamond M (2021) Dynamics of history-dependent perceptual judgment. Nat Commun 12:15.

Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD (2015) Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature 520:220–223.

Haynes JD (2015) A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron 87:257–270.

Hebart MN, Schriever Y, Donner TH, Haynes JD (2016) The relationship between perceptual decision variables and confidence in the human brain. Cereb Cortex 26:118–130.

Holroyd CB, Nieuwenhuis S, Yeung N, Nystrom L, Mars RB, Coles MG, Cohen JD (2004) Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. Nat Neurosci 7:497–498.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Kahnt T, Grueschow M, Speck O, Haynes JD (2011a) Perceptual learning and decision-making in human medial frontal cortex. Neuron 70:549–559.

Kahnt T, Heinzle J, Park SQ, Haynes JD (2011b) Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. Neuroimage 56:709–715.

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. Nature 455:227–231.

Kim JN, Shadlen MN (1999) Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. Nat Neurosci 2:176–185.

Klein E (1980) A semantics for positive and comparative adjectives. Linguist Philos 4:1–45.

Knapen T, Rolfs M, Wexler M, Cavanagh P (2010) The reference frame of the tilt aftereffect. J Vis 10:8–13.

Knill DC, Richards W (1996) Perception as Bayesian inference. Cambridge: Cambridge University Press.

Kohn A (2007) Visual adaptation: physiology, mechanisms, and functional benefits. J Neurophysiol 97:3155–3164.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.

Lages M, Treisman M (1998) Spatial frequency discrimination: visual long-term memory or criterion setting? Vision Res 38:557–572.

Lages M, Treisman M (2010) A criterion setting theory of discrimination learning that accounts for anisotropies and context effects. Seeing Perceiving 23:401–434.

Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A (2014) Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron 84:190–201.

Lak A, Hueske E, Hirokawa J, Masset P, Ott T, Urai AE, Donner TH, Carandini M, Tonegawa S, Uchida N, Kepecs A (2020) Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. Elife 9:e49834.

Lassiter D, Goodman ND (2017) Adjectival vagueness in a Bayesian model of interpretation. Synthese 194:3801–3836.

Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness. Trends Cogn Sci 15:365–373.

Lee SH, Blake R, Heeger DJ (2007) Hierarchy of cortical responses underlying binocular rivalry. Nat Neurosci 10:1048–1054.

Leeper TJ, Arnold J, Arel-Bundock V (2018) Margins: marginal effects for model objects. R package version 03 23. https://cran.r-project.org/web/packages/margins/margins.pdf.

Lueckmann J-M, Macke JH, Nienborg H (2018) Can serial dependencies in choices and neural activity explain choice probabilities? J Neurosci 38:3495–3506.

Macke JH, Nienborg H (2019) Choice (-history) correlations in sensory cortex: cause or consequence? Curr Opin Neurobiol 58:148–154.

Marco-Pallarés J, Müller SV, Münte TF (2007) Learning by doing: an fMRI study of feedback-related brain activations. Neuroreport 18:1423–1426.

Marcus D, Harwell J, Olsen T, Hodge M, Glasser M, Prior F, Jenkinson M, Laumann T, Curtiss S, Van Essen D (2011) Informatics and data mining tools and strategies for the human connectome project. Front Neuroinform 5:4.

McLelland D, Ahmed B, Bair W (2009) Responses to static visual images in macaque lateral geniculate nucleus: implications for adaptation, negative afterimages, and visual fading. J Neurosci 29:8996–9001.

Mize TD, Doan L, Long JS (2019) A general framework for comparing predictions and marginal effects across models. Sociol Methodol 49:152–189.

Morgan M (2014) A bias-free measure of retinotopic tilt adaptation. J Vis 14:7.

Nahum M, Daikhin L, Lubin Y, Cohen Y, Ahissar M (2010) From comparison to classification: a cortical tool for boosting perception. J Neurosci 30:1128–1136.

Nakashima Y, Sugita Y (2017) The reference frame of the tilt aftereffect measured by differential Pavlovian conditioning. Sci Rep 7:40525.

Nestares O, Heeger DJ (2000) Robust multiresolution alignment of MRI brain volumes. Magn Reson Med 43:705–715.

Norton EH, Fleming SM, Daw ND, Landy MS (2017) Suboptimal criterion learning in static and dynamic environments. PLoS Comput Biol 13:e1005304.

Olman CA, Inati S, Heeger DJ (2007) The effect of large veins on spatial localization with GE BOLD at 3 T: displacement, not blurring. Neuroimage 34:1126–1135.

Orhan AE, Ma WJ (2019) A diverse range of factors affect the nature of neural representations underlying short-term memory. Nat Neurosci 22:275–283.

Patterson CA, Wissig SC, Kohn A (2013) Distinct effects of brief and prolonged adaptation on orientation tuning in primary visual cortex. J Neurosci 33:532–543.

Pavan A, Marotti RB, Campana G (2012) The temporal course of recovery from brief (sub-second) adaptations to spatial contrast. Vision Res 62:116–124.

Raviv O, Lieder I, Loewenstein Y, Ahissar M (2014) Contradictory behavioral biases result from the influence of past stimuli on perception. PLoS Comput Biol 10:e1003948.

Rees G, Kreiman G, Koch C (2002) Neural correlates of consciousness in humans. Nat Rev Neurosci 3:261–270.

Renart A, Machens CK (2014) Variability in neural activity and behavior. Curr Opin Neurobiol 25:211–220.

Rips LJ, Turnbull W (1980) How big is big? Relative and absolute properties in memory. Cognition 8:145–174.

Sanders JI, Hangya B, Kepecs A (2016) Signatures of a statistical computation in the human sense of confidence. Neuron 90:499–506.

Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software 35:1–22.

Sereno MI, Dale A, Reppas J, Kwong K, Belliveau J, Brady T, Rosen B, Tootell R (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889–893.

Sheehan TC, Serences JT (2022) Attractive serial dependence overcomes repulsive neuronal adaptation. PLoS Biol 20:e3001711.

Shmuel A, Yacoub E, Chaimow D, Logothetis NK, Ugurbil K (2007) Spatio-temporal point-spread function of fMRI signal in human gray matter at 7 Tesla. Neuroimage 35:539–552.

Smith AM, Lewis BK, Ruttimann UE, Frank QY, Sinnwell TM, Yang Y, Duyn JH, Frank JA (1999) Investigation of low frequency drift in fMRI signal. Neuroimage 9:526–533.

Solomon SG, Kohn A (2014) Moving sensory adaptation beyond suppressive effects in single neurons. Curr Biol 24:R1012–R1022.

Solt S (2015) Vagueness and imprecision: empirical foundations. Annu Rev Linguist 1:107–127.

Soon CS, Brass M, Heinze H-J, Haynes J-D (2008) Unconscious determinants of free decisions in the human brain. Nat Neurosci 11:543–545.

Stocker AA, Simoncelli EP (2006) Sensory adaptation within a Bayesian framework for perception. In: Advances in neural information processing systems, pp 1289–1296. MIT Press, Vancouver.

Treisman M, Williams TC (1984) A theory of criterion setting with an application to sequential dependencies. Psychol Rev 91:68–111.

Tribushinina E (2011) Once again on norms and comparison classes. Linguistics 49:525–553.

Urai AE, Donner TH (2022) Persistent activity in human parietal cortex mediates perceptual choice repetition bias. Nat Commun 13:6015.

Weber AI, Krishnamurthy K, Fairhall AL (2019) Coding principles in adaptation. Annu Rev Vis Sci 5:427–449.

White CN, Mumford JA, Poldrack RA (2012) Perceptual criteria in the human brain. J Neurosci 32:16716–16724.

Williams R, Jorgensen A (2023) Comparing logit and probit coefficients between nested models. Soc Sci Res 109:102802.

Wutz A, Loonis R, Roy JE, Donoghue JA, Miller EK (2018) Different levels of category abstraction by different dynamics in different prefrontal areas. Neuron 97:716–726.e8.

Zhang H, Alais D (2020) Individual difference in serial dependence results from opposite influences of perceptual choices and motor responses. J Vis 20:2.

Zokaei N, Burnett Heyes S, Gorgoraptis N, Budhdeo S, Husain M (2015) Working memory recall precision is a more sensitive index than span. J Neuropsychol 9:319–329.