Title

A Normative Bayesian Model of Classification for Agents with Bounded Memory

Abbreviated title

Classification Model with Bounded Memory

Names and affiliations

Heeseung Lee[1], Hyang-Jung Lee[1], Kyoung Whan Choe[2,3], and Sang-Hun Lee[1]

[1]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Korea; [2]Department of Psychology, The University of Chicago, Chicago, Illinois, USA; [3]Mansueto Institute for Urban Innovation, The University of Chicago, Illinois, USA

Corresponding author

Sang-Hun Lee; visionsl@snu.ac.kr

There is no financial interests or conflicts of interest

Abstract

Classification, one of the key ingredients for human cognition, entails establishing a criterion that splits a given feature space into mutually exclusive subspaces. In classification tasks performed in daily life, however, a criterion is often not provided explicitly but instead needs to be guessed from past samples of a feature space. For example, we judge today's temperature to be "cold" or "warm" by implicitly comparing it against a "typical" seasonal temperature. In such situations, establishing an optimal criterion is challenging for cognitive agents with bounded memory because it requires retrieving an entire set of past episodes with precision. As a computational account for how humans carry out this challenging operation, we developed a normative

Bayesian model of classification (NBMC), in which Bayesian agents, whose working-memory precision decays as episodes elapse, continuously update their criterion as they perform a binary perceptual classification task on sequentially presented stimuli. We drew a set of specific implications regarding key properties of classification from the NBMC, and demonstrated the correspondence between the NBMC and human observers in classification behavior for each of those implications. Furthermore, in the functional magnetic resonance imaging responses acquired concurrently with behavioral data, we identified an ensemble of brain activities that coherently represent the latent variables, including the inferred criterion, of the NBMC. Given these results, we believe that the NBMC is worth being considered as a useful computational model that guides behavioral and neural studies on perceptual classification, especially for agents with bounded memory representation of past sensory events.

Significant Statement

Although classification—assigning events into mutually exclusive classes—requires a criterion, people often have to perform various classification tasks without explicit criteria. In such situations, forming a criterion based on past experience is quite challenging because people's memory of past events deteriorates quickly over time. Here, we provided a computational model for how a memory-bounded yet normative agent infers the criterion from past episodes to maximally perform a binary perceptual classification task. This model successfully captured several key properties of human classification behavior, and the neural signals representing its latent variables were identified in the classifying human brains. By offering a rational account for memory-bonded agents' classification, our model can guide future behavioral and neural studies on perceptual classification.

Introduction

Classification, an act of organizing entities into mutually exclusive classes according to a criterion, is an essential ingredient of human cognition. Classification is necessary to the emergence of scientific constructs such as taxonomies in biology (Hempel, 1965; Jacob, 2004) and entailed in generating basic linguistic propositions such as predication (Rips and Turnbull, 1980), being frequently exercised when describing daily events (e.g., 'the train arrived earlier/later than the scheduled time').

A criterion is necessary for classification but often not explicitly provided. Previous modelling efforts offered several algorithms for how a criterion is guessed in such situations (Lapid et al., 2008; Raviv et al., 2012; Treisman and Williams, 1984) and described human classification behavior (Lages and Treisman, 1998). However, they did not specify any process of minimizing classification errors, leaving it unclear how their algorithms can be adopted from the optimality perspective. Furthermore, they did not specify how uncertainty—which prevails in the world and is critical for normative information processing (Gorgoraptis et al., 2011; Kersten et al., 2004; Körding and Wolpert, 2004; Zokaei et al., 2015)—is incorporated into their algorithms. Here, to overcome these limitations, we propose a normative Bayesian model of classification (NBMC) with an explicit cost function and probabilistic representation of uncertainty (Griffiths et al., 2012; Ma, 2012).

For model transparency and validation, we developed the NBMC based on a simple perceptual classification task[1]. This task requires classifying sequentially presented stimuli into 'large' or 'small' sizes with respect to their median size, which is not explicitly presented to observers (Fig. 1). Given this statistical structure, the best criterion computation is to infer the median size of the stimuli that have been experienced so far from the noisy sensory evidences

[1] We are aware that 'classification' and 'categorization' are used interchangeably. But we insist that these two tasks, despite their superficial similarity, must be distinguished because they differ in computational architecture. The implications of this distinction will be addressed in Results and Discussion.

instigated by those stimuli. Critically, we assumed that these sensory evidences become increasingly uncertain as trials elapse, following recent reports on the temporal decay of human working memory precision (Gorgoraptis et al., 2011; Zokaei et al., 2015). Accordingly, the NBMC modeled humans as Bayesian observers who infer the criterion from past sensory evidences by incorporating the temporal increase of memory uncertainty into their generative models. Then, by inferring the size of a current stimulus (s) and comparing that against the inferred criterion (c), the Bayesian observers make a decision.

The NBMC has the following useful (Kording et al., 2020) implications regarding the key properties of perceptual classification. Firstly, the NBMC implicates specific history effects: (i) only past stimuli, but not past decisions, bias current decisions because c varies only as a function of past sensory evidences; (ii) this stimulus history effect will diminish as trials elapse because the Bayes-optimal integration weights for sensory evidences decrease as the degree of uncertainty increases (Ernst and Banks, 2002; Körding and Wolpert, 2004). Secondly, because classification outcomes are determined by the interplay between s and c, (i) the trial-to-trial variability of decision and its uncertainty can be dissected into those from two distinct sources, s and c, and (ii) classification outcomes are not governed by absolute quantities of stimuli but by relative quantities between s and c. Lastly, because the decision variable (v) for classification is determined by the difference between s and c, these three latent variables of the NBMC constitute the causal structure of 'common effect' ($s \rightarrow v \leftarrow c$). This implicates that (i) there exist the respective signals of s, c, and v in the human brain engaged in classification and that (ii) the correlations among those brain signals are consistent with the causal structure prescribed by the NBMC. We verified these implications by examining the behavioral and functional-magnetic-resonance-imaging (fMRI) brain responses from human observers performing a perceptual classification task.

Materials and Methods

A total of four data sets was acquired (Table 1). The main data set (DS#1) was acquired from 18 participants (9 females, aged 20–30 years), who performed the task inside an MRI scanner. The current sample size is comparable to that used in previous fMRI studies with methods similar to the current study (Kahnt et al., 2011a; Kahnt et al., 2011b). The remaining three data sets (DS#2~4) were borrowed from previously published (Choe et al., 2014, 2016) or in-preparation (Choe et al., 2014, 2016; Lee et al., 2016) papers in our lab. DS#1 was never published nor used in any previous work. The Research Ethics Committee of Seoul National University approved the experimental procedures. All participants gave informed consent and were naïve to the purpose of the experiments.

Behavioral data acquisition for DS#1

Figure 1A illustrates the experimental procedure and stimuli for DS#1. While fixating at the screen center, observers were instructed to view a brief (0.3s) ring-shape stimulus and classify its size within 1.5s after stimulus onset into either 'small' or 'large' by pressing the left-hand and right-hand keys, respectively. The identity (left/right) and timing of each key press were recorded.

Task conditions for the training (pre-scanning) sessions. Before participating in the main fMRI experiment, observers had practiced on the task intensively over several (3 to 6) training sessions (~1,000 trials for each session) outside the scanner until they reached an asymptotic level in accuracy. On an initial block of trials (~400) of each training session, we presented ring stimuli of 24 different, fine-grained radii (7.65° ~ 10.35°) in the order prescribed by an adaptive staircase (1-up-2-down) method and provided oberservers with trial-to-trial feedback based on a fixed, objective classification criterion (radius of 9°) to help them understand the goal of the task (i.e., to maximize the proportion of correct classification) and to determine the three threshold-level ring sizes tailored for a given observer. Then, on the following block of trials (~600), we

asked observers to perform the task on this tailored triplet of ring stimuli while providing them with run-to-run feedback to help them get used to the task conditions under which they will later perform the task inside the MR scanner. During the training sessions, consecutive trials were apart from one another by 2.7s. Note that we opted to train observers with the stimuli that were much larger from those for the main experiment (mean radius of 2.84°) to avoid any unwanted adaptation or learning effects at low sensory levels and thus to focus training on the task structure of perceptual classification.

Task conditions for the main (scanning) session. The task conditions for the main session were identical to those for the training sessions, except for the following. Consecutive trials were apart from one another by 13.2s. The long inter-stimulus interval in conjunction with the brief stimulus presentation was implemented to minimize possible sensory adaptation to stimuli, and thus to prevent sensory adaptation in previous trials from interfering with decision processes in a current trial (Nakashima and Sugita, 2017; Pavan et al., 2012). Observers were not provided with trial-to-trial feedback but with run-to-run feedback, which was to show the percent correct for a run of 26 trials during each break between scan runs. Before the main fMRI runs, observers inside the MRI scanner repeated what they did during the training sessions by performing 54 practice trials and then 180 threshold-calibration trials. While performing the threshold-calibration trials, which were apart by 2.7s, observers received trial-to-trial feedback based on the classification criterion with 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 size threshold (i.e., the size difference between M ring and L or S ring) associated with 70.7% correct proportion was estimated. We expected that these calibration and practice trials, in addition to the repetitive alternations between the trial-to-trial-feedback calibration trials and the run-to-run-feedback practice trials during the training sessions (see above), would help observers perform the task during the main session in the same way as they did with trial-to-trial feedback. This expectation was supported by the results that the

150 performance levels in the main session (proportion correct = 75±7%) were similar to those
151 targeted by the threshold-estimation procedure (70.7%).
152
153 **Behavioral data acquisition for DS#2~4**
154 DS#2-4 were borrowed from the works conducted with different purposes in our lab (Table 1).
155 The specifics of these data sets were as follows: DS#2 was previously published in our previous
156 studies (Choe et al., 2014, 2016) and contributed by 23 observers, each of whom performed 162
157 trials on the ring stitmuli of 3 different sizes (2.80° ~ 2.88°), with run-to-run and deterministic
158 feedback (see below for definition of deterministic feedback) and inter-trial interval of 13.2s.
159 DS#3 was collected from the same participants of DS#2 but not published before. Each of the
160 subjects, except for one, performed 630 trials on the ring stimuli of 16 different ring sizes (2.72° ~
161 2.95°) presented via a staircase method (1-up-2-down), with trial-to-trial and deterministic
162 feedback and inter-trial interval of 2s. The one exceptional subject performed 315 trials. DS#4 was
163 from our preliminary poster presentation (Lee et al., 2016) and contributed by 30 observers, each
164 of whom performed 1,700 trials on the ring stimuli of 5 different ring sizes (3.84°, 3.92°, 4.00°,
165 4.08 °, 4.16°), with trial-to-trial and stochastic feedback (see below for definition of stochastic
166 feedback) and inter-trial interval of 2.5s. The feedback correctness was determined by a constant
167 criterion for the deterministic feedback condition (DS#1-3) or by a noisy criterion (i.e., the
168 criterion was slightly jittered around a mean, with a gaussian distribution, across trials) for the
169 stochastic feedback condition (DS#4). Stimulus duration was 0.3s for all the data sets. Unlike
170 DS#1, DS#2-4 were acquired without any intensive training sessions.
171
172 **Imaging data acquisition and preprocessing**
173 MRI data were collected using a 3 Tesla Siemens Tim Trio scanner equipped with a 12-channel
174 Head Matrix coil at the Seoul National University Brain Imaging Center. Stimuli were generated
175 using MATLAB (MathWorks) in conjunction with MGL (http://justingardner.net/mgl) on a

176 Macintosh computer. Observers looked through an angled mirror attached to the head coil to
177 view stimuli displayed via an LCD projector (Canon XEED SX60) onto a back-projection screen at
178 the end of the magnet bore at a viewing distance of 87 cm, yielding a field of view of 22×17°.
179 For each observer, we acquired three types of MRI images, as follows: (1) 3D, T1-weighted,
180 whole-brain images (MPRAGE; resolution, 1×1×1 mm; field of view (FOV), 256 mm; repetition time
181 (TR), 1.9 s; time for inversion, 700ms; time to echo (TE), 2.36; and flip angle (FA), 9°), (2) 2D, T1-
182 weighted, in-plane images (MPRAGE; resolution, 1.08×1.08×3.3 mm; TR, 1.5s; T1, 700 ms; TE,
183 2.79ms; and FA, 9°), and (3) 2D, T2*-weighted, functional images (gradient EPI; TR, 2.2s; TE, 40ms;
184 FA, 73°; FOV, 208mm; image matrix, 90×90; slice thickness, 3mm with 10% space gap; slice, 32
185 oblique transfers slices; bandwidth, 790Hz/ps; and effective voxel size, 3.25×3.25×3.3mm).
186 For univariate analysis, the images of individual observers were normalized to the MNI
187 template using the following steps: motion correction, coregistration to whole-brain anatomical
188 images via the in-plane images (Nestares and Heeger, 2000), spike elimination, slice timing
189 correction, normalization using the SPM DARTEL Toolbox (Ashburner, 2007) to 3×3×3mm voxel
190 size, and smoothing with 8×8×8mm full-width half-maximum Gaussian kernel. All the procedures
191 were implemented with SPM8 and SPM12 (http://www.fil.ion.ucl.ac.uk.spm) (Friston et al., 1996;
192 Jenkinson et al., 2002), except for spike elimination for which we used the AFNI toolbox (Cox,
193 1996). The first six frames of each functional scan (the first trial of each run) were discarded to
194 allow hemodynamic responses to reach a steady state. Then, the normalized BOLD time series at
195 each voxel, each run, and each subject were preprocessed using linear detrending and high-pass
196 filtering (132 s cut-off frequency with a Butterworth filter), conversion into percent-change
197 signals, and correction for non-neural nuisance signals by regressing out mean BOLD activity of
198 cerebrospinal fluid (CSF).
199 To define anatomical masks, probability tissue maps for individual participants were
200 generated from T1-weighted images, normalized to the MNI space, and smoothed as were done
201 for the functional images by using SPM12, and then averaged across participants. Finally, the

locations of CSF, white matter, and gray matter were defined as respective groups of voxels in which the probability was more than 0.5.

The preprocessing steps for multivoxel analysis were the same, but only spatial smoothing was omitted to prevent blurring of the pattern of activity. Unfortunately, in a few of the subjects, functional images did not cover entire brain areas. Voxels in which data were derived from fewer than 17 subjects were excluded for further analysis, which included those in the temporal pole, orbitofrontal, and posterior cerebellum.

NBMC: generative model

The generative model is the observers' causal account for noisy sensory measurements, where true ring size, Z, causes a noisy sensory measurement on a current trial, $m_{(t)}$, which becomes noisier as trials elapse, thus turning into a noisy retrieved measurement of the value of Z on trial $t-i$, $r_{(t-i)}$. Hence, the generative model can be specified with the following three probabilistic terms: a prior of Z, $p(Z)$, a likelihood of Z given $m_{(t)}$, $p(m_{(t)}|Z)$, and a likelihood of Z given $r_{(t-i)}$, $p(r_{(t-i)}|Z)$. 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 learnt 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, $\Theta = \{\mu_0, \sigma_0, \sigma_m, \kappa\}$.

We defined the ring size, Z, in a normalized space with an arbitrary unit. Specifically, physical ring sizes in visual angle unit were normalized to Z values by dividing their deviances from the M-ring size with the minimum deviance from the M-ring size. For example, five ring sizes in visual degree $\{1,5,7,9,13\}$ were normalized to Z values of $\{-3,-1,0,1,3\}$ by dividing the deviances from 7 ($\{-6,-2,0,2,6\}$) with the minimum deviance from 7 (2).

NBMC: stimulus inference (s)

A Bayesian estimate of the value of Z on a current trial, $s_{(t)}$, was defined as the most probable value of a posterior function of a given sensory measurement $m_{(t)}$, which minimizes the mean squared error of estimation for Gaussian distributions (Jaynes, 2003; Körding and Wolpert, 2004). The posterior $p(Z|m_{(t)})$ is a conjugate normal distribution of the prior and likelihood of Z given the evidence $m_{(t)}$, whose mean $\mu_{s_{(t)}}$ and standard deviation $\sigma_{s_{(t)}}$ were calculated as follows:

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

Whenever Z takes one of the ring sizes on each trial as $Z_{(t)}$, its generative noise in generating $m_{(t)}$ was assumed to be equivalent to σ_m. Therefore, σ_m propagates through the Bayesian estimates of stimulus $s_{(t)}$, which results in the sampling distribution of estimates whose mean $\mu_{p(s_{(t)})}$ and standard deviation $\sigma_{p(s_{(t)})}$ were calculated as follows:

$$\mu_{p(s_{(t)})} = ({\sigma_0}^2 Z_{(t)} + {\sigma_m}^2 \mu_0)/({\sigma_0}^2 + {\sigma_m}^2);\ \sigma_{p(s_{(t)})} = {\sigma_0}^2 \sigma_m/({\sigma_0}^2 + {\sigma_m}^2).$$

NBMC: criterion inference (c)

The Bayesian observer estimates the value of criterion on a current trial, $c_{(t)}$, by inferring the most probable value of a posterior function of a given set of retrieved sensory measurements $\mathbb{r}_{(t)} = \{r_{(t-1)}, r_{(t-2)}, \ldots - r_{(t-n)}\}$, where the maximum number of measurements that can be retrieved, n, was set to 7. This definition of $c_{(t)}$ allows the NBMC to obtain the optimal estimate

of $c_{(t)}$ by minimizing squared error (Jaynes, 2003; Körding and Wolpert, 2004). Here, $p(Z|\mathbb{r}_{(t)})$ is a conjugate normal distribution of the prior and likelihoods of Z given the evidence $\mathbb{r}_{(t)}$, $p(Z|\mathbb{r}_{(t)}) \propto p(\mathbb{r}_{(t)}|Z)p(Z) = p(r_{(t-1)}|Z)p(r_{(t-2)}|Z) \dots p(r_{(t-7)}|Z)p(Z)$, whose mean and standard deviation were calculated (Bromiley, 2003) based on the knowledge of how the retrieved stimulus becomes noisier as trials elapse:

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

, where $\beta_0 = {\sigma_0}^{-2}/({\sigma_0}^{-2} + \sum_{i=1}^{7} {\sigma_{r_{(t-i)}}}^{-2})$ and $\beta_i = {\sigma_{r_{(t-i)}}}^{-2}/({\sigma_0}^{-2} + \sum_{i=1}^{7} {\sigma_{r_{(t-i)}}}^{-2})$.

Much like stimulus estimates, the sampling distribution of criterion estimates have a mean $\mu_{p(c_{(t)})}$ and a standard deviation $\sigma_{p(c_{(t)})}$ due to generative noise propagation, and they were calculated as follows:

$$\mu_{p(c_{(t)})} = \beta_0\mu_0 + \sum_{i=1}^{7} \beta_i Z_{(t-i)};\ \sigma_{p(c_{(t)})} = \sqrt{\sum_{i=1}^{7} {\beta_i}^2 {\sigma_{r_{(t-i)}}}^2}.$$

We could analytically derive this equation because we assumed that, when the criterion is inferred, the retrived stimulus measurement $r_{(t-i)}$ is randomly resampled from a normal distribution whose mean is equal to the stimulus that instigated it, $Z_{(t-i)}$, and whose standard deviation is $\sigma_{r_{(t-i)}}$. In other words, retrieved sensory measurements ($\{r_{(t-1)}, r_{(t-2)}, \dots - r_{(t-n)}\}$) were assumed to be independent from their initial sensory measurements ($\{m_{(t-1)}, m_{(t-2)}, \dots - m_{(t-n)}\}$) and from one another as well.

NBMC: deduction of decision variable (v), decision (d) and decision uncertainty (u)

On each trial, the Bayesian observer makes a binary decision $d_{(t)}$ by comparing $s_{(t)}$ and $c_{(t)}$, such that $d_{(t)}$ = 'large' if $s_{(t)} > c_{(t)}$, and $d_{(t)}$ = 'small' if $s_{(t)} < c_{(t)}$. Thus, given the generative noise propagation, the fraction of 'large' choices, $pL_{(t)}$, was defined as the proportion of the bivariate sampling distribution that satisfied the inequality of $s_{(t)} > c_{(t)}$: $pL_{(t)} = \Phi[\frac{\mu_{p(s_{(t)})} - \mu_{p(c_{(t)})}}{\sqrt{\sigma^2_{p(s_{(t)})} + \sigma^2_{p(c_{(t)})}}}]$, where Φ is cumulative normal distribution (Marzban, 2004).

However, decisions are not made deterministically but probabilistically, because $s_{(t)}$ and $c_{(t)}$ have their respective imprecisions parameterized with $\sigma_{s_{(t)}}$ and $\sigma_{c_{(t)}}$. Thus, when committing to a decision with $s_{(t)}$ and $c_{(t)}$, the Bayesian observer calculates the probability of $s_{(t)} > c_{(t)}$, which is called the decision variable, $v_{(t)}$, defined as $v_{(t)} = \Phi[\frac{s_{(t)}-c_{(t)}}{\sqrt{\sigma^2_{s_{(t)}}+\sigma^2_{c_{(t)}}}}]$, and the decision uncertainty, $u_{(t)}$, which represents the odds that the current decision will be incorrect (Sanders et al., 2016), as follows: $u_{(t)} = \Phi[\frac{-|s_{(t)}-c_{(t)}|}{\sqrt{\sigma^2_{s_{(t)}}+\sigma^2_{c_{(t)}}}}]$.

Fitting the parameters of NBMC

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

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

, where $p(D_{(t)}|\Theta) = pL_{(t)}$ for $D_{(t)}$ ='large' and $p(D_{(t)}|\Theta) = 1 - pL_{(t)}$ for $D_{(t)}$ ='small'.

For each subject, estimation was carried out in the following steps: First, we found local minima for parameters using a MATLAB function, 'fminseachbnd.m', with the iterative evaluation number set to 50. We repeated this step by choosing 1,000 different initial parameter sets, which were randomly sampled within uniform prior bounds, and acquired 1,000 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 the first trial of each run, the trials in which RTs were too short (less than 0.3 s), and the trials in which no responses were given for parameter estimation for any further analyses. We discarded the first trial of each run for two reasons: first, the criterion could not be inferred on the first trial since there is no measurement to be retrieved yet; second, the first trial is susceptible to non-specific fMRI signals that are irrelevant to the task. Note that some aspects of the fitting procedure are arbitrary (e.g., the parameter boundaries and the number of initial parameter randomization). To be sure, the outcomes of the fitted model parameters are subect to slight changes depending on how those aspects are arbitrated. However, we confirmed that such changes are sufficiently small (results were not shown here) not to affect the main claims of the current study. The fitting codes and data are available in GitHub (see Methods: Data and code accessibility).

The constant-criterion model

The constant-criterion model has two parameters, bias of classification criterion μ_0 and measurement noise σ_m. Stimulus estimates, $s_{(t)}$, were assumed to be sampled from a normal distribution, $\mathcal{N}(Z_{(t)}, \sigma_m)$. Each stimulus sample has uncertainty $\sigma_{s_{(t)}} = \sigma_m$. Criterion estimate $c_{(t)}$ was assumed to be a constant, μ_0; thus $\sigma_{p(c_{(t)})} = \sigma_{c_{(t)}} = 0$.

Comparing the Bayesian and human observers in decision behavior

Estimating the model parameters separately for each human observer ($\widehat{\Theta} = \{\mu_0, \sigma_0, \sigma_m, \kappa\}$) allowed us to create the same number of Bayesian observers, each tailored to each human individual. By repeating the experiment on these Bayesian observers using the stimulus sequences that were identical to those presented to their human partners, 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 stimulus estimate, $s_{(t)}$, and criterion estimate, $c_{(t)}$, that were randomly sampled from the stimulus ($\mu_{p(s_{(t)})}$with $\sigma_{p(s_{(t)})}$) and criterion

($\mu_{p(c_{(t)})}$ with $\sigma_{p(c_{(t)})}$) estimate distributions, respectively, on each trial for each observer. Then, the averages across those 10^6 simulations were taken as the final outcomes. When predicting $s_{(t)}$, $c_{(t)}$, $v_{(t)}$, and $u_{(t)}$ for the observed choice $D_{(t)}$, we only counted the simulation outcomes in which the relation between $s_{(t)}$ and $c_{(t)}$ (i.e., $d_{(t)}$) matched the observed choice $D_{(t)}$.

To check the correspondence in the impact of previous stimuli and previous choices on current choices, we logistically regressed the human current choices onto stimuli and choices using the following model to obtain regression coefficients $\vec{p} = [p_1, \cdots, p_{11}]$ for each observer: $\vec{D}_{(t)} \sim e^{\vec{K}_{(t)}}/(1 + e^{\vec{K}_{(t)}})$, where $\vec{K}_{(t)} = p_0 + p_1\vec{Z}_{(t)} + \sum_{i=1}^{5}(p_{1+i}\vec{Z}_{(t-i)} + p_{6+i}\vec{D}_{(t-i)})$, the independent variables were each standardized into z-scores for each observer, and $\vec{D}_{(t)} = [D_{(1)}, \cdots D_{(T)}]$, $\vec{K}_{(t)} = [K_{(1)}, \cdots K_{(T)}]$, $\vec{Z}_{(t)} = [Z_{(1)}, \cdots Z_{(T)}]$, where T is the number of trials in a session. The Bayesian observers' choices were also regressed with the logistic regression model by substituting $\vec{d}_{(t)}$ for $\vec{D}_{(t)}$, where $\vec{d}_{(t)} = [d_{(1)}, \cdots d_{(T)}]$. The regression was repeatedly carried out for each simulation, and the coefficients $\vec{\hat{p}}$ that were averaged across simulations were taken as final outcomes. Due to the additional time consumption associated with regression, the number of simulations for regression was compromised to 10^5 repetitions, which is smaller than that for $pL_{(t)}$ prediction (10^6). However, we confirmed that the simulation number was sufficiently large to produce stable simulation outcomes.

When checking the correspondence between the decision uncertainty values $u_{(t)}$ of the Bayesain observers and the RT measurements of the human observers, we converted $u_{(t)}$ values into the values that are comparable to RT measurements by regressing RT measurements onto $u_{(t)}$ using a generalized linear mixed model (GLMM), $\overrightarrow{RT}_{(t)} \sim \beta_0 + \beta_1\vec{u}_{(t)}$, with a random effect of individual observers, where $\overrightarrow{RT}_{(t)} = [RT_{(1)}, \cdots RT_{(T)}]$ and $\vec{u}_{(t)} = [u_{(1)}, \cdots u_{(T)}]$. Then, '$\beta_0 + \beta_1\vec{u}_{(t)}$' were pitted against the RT measurents.

To check the correspondence in the impact of previous stimuli on decision uncertainty, we regressed the decision uncertainty estimates $\vec{u}_{(t)}$ of the Bayesian observers and the RTs of

the human observers $\overrightarrow{RT}_{(t)}$, respectively, on current trials onto the congruency of current and past stimuli with a current choice using the following GLMM with a random effect of individual observers:

$\overrightarrow{RT}_{(t)} \sim q_0 + \sum_{i=0}^{5} q_{i+1} g_{Z(t-i)} + \sum_{i=1}^{5} q_{i+6} g_{D(t-i)}$, where $g_{Z(t-i)} = Z_{(t-i)} \times D_{(t)}$; $g_{D(t-i)} = D_{(t-i)} \times D_{(t)}$; $D_{(t)} = -1$ and 1 for 'large' and 'small' decisions respectively; for the definition of $Z_{(t)}$ see NBMC: generative model. Here, the independent and dependent variables were both standardized into z-scores for each observer. For comparison with the regression coefficients for the human observers, the regression coefficients for the Bayesian observers were calculated as $\vec{u}_{(t)} \sim \hat{q}_0 + \sum_{i=0}^{5} \hat{q}_{i+1} g_{Z(t-i)} + \sum_{i=1}^{5} \hat{q}_{i+6} g_{D(t-i)}$ by GLMM with a random effect of individual observers, and then linearly re-scaled to those for the human observers using the following GLM: $\sum_{i=0}^{11} q_i \sim 1\left(\sum_{i=0}^{11} \hat{q}_i\right)$, where γ is a GLM parameter for scaling.

Regression of fMRI responses onto model estimates of decision uncertainty

To identify the brain regions wherein fMRI responses were correlated with $u_{(t)}$ estimates of the NBMC, we regressed the fMRI responses, $B_{(v,t,i)}$, of each voxel, v, at each of the six time points, i, within a single trial, t, onto $u_{(t)}$ using the following GLMM with random effects of individual observers: $\vec{B}_{(v,t,i)} \sim \beta_{0(v,i)} + \beta_{1(v,i)} \vec{u}_{(t)}$, where $\vec{B}_{(v,t,i)} = [B_{(v,1,i)}, \cdots, B_{(v,T,i)}]$. The independent and dependent variables of the regression were both standardized into z-scores for each observer across trials. We first localized the cortical sites where $\beta_{1(v,i)}$ were statistically significant after correcting for the false discovery rate (FDR) (Benjamini and Hochberg, 1995) over the entire brain voxels tested at each time point, i. Then, the regions of interest (ROIs) were defined as the voxel clusters that cover a region larger than 350 mm^3 (> 12 contiguous voxels) in which the voxels' FDR-corrected p-values were less than 0.05 and raw p-values were less than 10^{-4}. For ROI analysis, fMRI responses were averaged across individual voxels within a given ROI, and their

correspondences with the model estimates of decision uncertainty were calculated using the same procedure as that used for the analysis of RT data.

Bayes factor analysis for statistical judgment of distinguishable and indistinguishable classifications

Because the NBMC specifies each and every trial with a set of expected values of $\{s_{(t)}, c_{(t)}, pL_{(t)}, u_{(t)}\}$, it can deterministically predict which trial pairs (i, j) are distinguishable or indistinguishable in decision probability ($\Delta pL_{(i,j)} = pL_{(i)} - pL_{(j)}$) and decision uncertainty ($\Delta u_{(i,j)} = u_{(i)} - u_{(j)}$) based on the metameric currency for each trial pair ($[s_{(i)} - c_{(i)}] - [s_{(j)} - c_{(j)}]$).

To statistically judge whether a given trial pair is significantly distinguishable or indistinguishable, we applied the following procedure on DS#1. First, we grouped individual trials according to physical ring sizes, and sorted the trials within each stimulus group by $\mu_{p(s_{(t)})} - \mu_{p(c_{(t)})}$ into 10 equal-sized bins, resulting in a total of 30 bins. Second, we made all possible bin pairs out of these bins (900 bin pairs = 30 bin x 30 bins). Then, for each bin pair (i,j), we judged whether the observed $\Delta pL_{(i,j)}$ is significantly "equal to zero (H_I: $\Delta pL_{(i,j)} = 0$)" or "not equal to zero (H_D: $\Delta pL_{(i,j)} \neq i$) by calculating the Bayes factor B (Dienes, 2008, 2014; Good, 1979; Jeffreys, 1961; Kass and Raftery, 1995; Wod, 1985), which represents the ratio of the marginal likelihoods of two competing hypotheses, H_I ("Indistinguishable") and H_D ("Distinguishable"):

$$B_{(i,j)} \equiv \frac{p(\Delta pL_{(i,j)}|H_D)}{p(\Delta pL_{(i,j)}|H_I)} = \frac{\int_{-1}^{1} p(\Delta pL_{(i,j)}|\Delta pL)p(\Delta pL|H_D)d\Delta pL}{\int_{-1}^{1} p(\Delta pL_{(i,j)}|\Delta pL)p(\Delta pL|H_I)d\Delta pL},$$

where, $p(\Delta pL|H_I)$ and $p(\Delta pL|H_D)$ represent how likely H_I and H_D, respectively, are for a given value of ΔpL, and $p(\Delta pL_{(i,j)}|\Delta pL)$ represents how likely ΔpL is for a given value $\Delta pL_{(i,j)}$. These three likelihoods were assumed to be normally distributed, their means (μs) and standard deviation (σs) were estimated as follows: μs of $p(\Delta pL_{(i,j)}|H_I)$ and $p(\Delta pL_{(i,j)}|H_D)$ were both zero; σs of $p(\Delta pL_{(i,j)}|H_I)$ and $p(\Delta pL_{(i,j)}|H_D)$ were estimated by computing the SD of $\Delta pL_{(i,j)}$s for the 30

bin pairs where $i = j$ (i.e., $SD = 0$) and for the 870 bin pairs where $i \neq j$; for $p(\Delta pL_{(i,j)}|\Delta pL)$, μ was $\Delta pL_{(i,j)}$, and σ was estimated by computing the across-subject standard error of $\Delta pL_{(i,j)}$ for each bin pairs (thus, σ slightly differed across the bin pairs). Lastly, for each bin pair, the observed $\Delta pL_{(i,j)}$ was judged to be significantly equal to zero and different from zero when the Bayes factor was smaller than 1/3 and greater than 3, respectively, following the convention (Jeffreys, 1961).

The almost identical analyses were applied to the RT and dACC data, except that the metameric currency for $u_{(t)}$ was defined in an absolute term, $|s_{(t)} - c_{(t)}|$. Therefore, two trials become metamers when they have sufficiently similar values of $|s_{(t)} - c_{(t)}|$ despite their difference in stimulus-decision congruency ($u_{(i)} \approx u_{(j)}|Cong._{(i)} \neq i)g.\, i_{(j)}$) or anti-metamers when they have sufficiently different values of $|s_{(t)} - c_{(t)}|$, despite their similarity in stimulus-decision congruency ($u_{(i)} \neq u_{(j)}|Cong._{(i)} = Cong._{(j)}$).

Searching for multivoxel patterns of activity signaling the latent variables of the NBMC

To decode the latent variables of the NBMC in fMRI responses, the time-resolved support vector regression (SVR) was carried out in conjunction with a searchlight technique (Haynes, 2015; Kahnt et al., 2011b). At each of the first four time points, i, within a single trial, t, $B_{(e,k,t,i)}$ (which represents the preprocessed but unsmoothed fMRI responses of a voxel cluster centered at a gray-matter voxel, e, where k where x denotes an entire set of voxels that comprise the cluster) was selected as a searchlight. The first four time points were chosen given that the latent variables must precede $u_{(t)}$, which was detected at the fourth time point of the univariate fMRI responses in the dACC. Although the searchlight cluster had a radius of 9mm and thus consisted of 123 voxels, the exact number of voxels in each searchlight varied, because the voxels located in CSF or white matter were discarded as non-neural signals. For each observer, the latent variables ($s_{(t)}$, $c_{(t)}$, and $v_{(t)}$) were decoded for each searchlight using the cross-validation method of one-run-leave-out (8-fold cross-validation).

422 We performed the SVR using the LIBSVM (http://www.csie.ntu.edu.tx/~sjlin/libsvm) with
423 a linear kernel and constant regularization parameter of 1. For each voxel and each subject,
424 $B_{(e,k,t,i)}$ and the values of the latent variables were z-scored across trials before decoding. To
425 calculate the significance level of the decoded information, each of the 3D spatial maps of the
426 decoded values of the latent variables ($v_{B(e,t,i)}$, $c_{B(e,t,i)}$, or $s_{B(e,t,i)}$) was smoothed with a 5 mm
427 FWHM Gaussian kernel. We then regressed the smoothed $v_{B(e,t,i)}$, $c_{B(e,t,i)}$, and $s_{B(e,t,i)}$ onto $v_{(t)}$,
428 $c_{(t)}$, and $s_{(t)}$, respectively, using 16, 13, and 13 regression models, respectively. Those regression
429 models were deduced from the causal structure between the latent variables of the NBMC (see
430 the next section). We concluded that a given cluster carries the neural signals of $v_{(t)}$, $c_{(t)}$, or $s_{(t)}$,
431 only when all those regression models are satisfied over 12 contiguous searchlights. For the ROI
432 analysis, the decoded values of a given latent variable were averaged over the all searchlights
433 within each ROI. The brain imaging results were visualized using xjView toolbox for the cross
434 sectional images and Connectome Workbench (Marcus et al., 2011) for the inflated images.
435

436 **Deduction of the regression models for $\mathbf{v_{(t)}}$, $\mathbf{c_{(t)}}$, and $\mathbf{s_{(t)}}$**

437 To identify brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$, we defined three a priori lists of regressions that
438 must be satisfied by the brain signals. We stress that each of these lists consists of the necessary
439 conditions to be satisfied, and the conditions that must not be satisfied as well, because both
440 significant (βof the ne) and non-significant (β a) regressions are deduced from the causal
441 structure of the variables in the NBMC. The lists were as follows.

442 The 13 regressions for the brain signal of c: (c1-3), y_c, c decoded from brain signals, must
443 be regressed onto c—the variable it represents—even when c is orthogonalized to v or d,
444 because it should reflect the variance irreducible to the offspring variables of c; (c4), y_c must not
445 be regressed onto s because c and s are independent; (c5,6), y_c must be regressed onto v but not
446 when v is orthogonalized to c because the influence of c on v is removed; (c7,8) y_c must be
447 regressed onto d but not onto u because u cannot be correlated with c unless d or s is fixed; (c9-

11), y_c must be regressed onto, not the current stimulus, but the past stimuli—strongly onto the 1-back stimulus and more weakly onto the 2-back stimulus (thus, non-significant regression with one-tailed regression in the opposite sign is modeled conservatively); (c12,13), y_c must not be regressed onto previous decisions at all because c is inferred solely from retrieved stimulus measurements.

The 13 regressions for the brain signal of s: (s1-3), y_s, s decoded from brain signals, must be regressed onto s—the variable it represents per se—even when s is orthogonalized to v or d because it should reflect the variance irreducible to the offspring variables of s; (s4), y_s must not be regressed onto c because s and c are independent of each other; (s5,6), y_s must be regressed onto v but not when v is orthogonalized to s because the influence of s on v is removed; (s7,8) y_s must be regressed onto d but not onto u because u cannot be correlated with s unless d or c is fixed (s9-11), y_s must be regressed onto the current stimuli and not the past stimuli because s is inferred solely from the current stimulus measurement; (s12,13), y_s must not be regressed onto previous decisions because s is inferred solely from the current stimulus measurement.

The 16 regressions for the brain signal of v. (v1-4), y_v, v decoded from brain signals, must be regressed onto v—the variable it represents—even when v is orthogonalized to c, s, or d, because it should reflect the variance irreducible to the offspring variables of v; (v5,6), y_v must be regressed onto one of its parents c, but not when c is orthogonalized to v, because the influence of c on v is removed; (v7,8), y_v must be regressed onto one of another parent s, but not when s is orthogonalized to v, because the influence of s on v is removed; (v9,10), 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; (v11-13), y_v must be positively regressed onto the current stimulus because the influence of the current stimulus on v is propagated via s, and negatively regressed onto the past stimuli because the influence of the past stimuli on v is propagated via c —strongly onto the 1-back stimulus and more weakly onto the 2-back stimulus (thus, non-significant regression with one-tailed regression in the opposite sign is modeled moderately);

(v14-16), y_v must be regressed onto the current decision and not the past decisions because the current decision is a dichotomous translation of v, whereas past decisions have nothing to do with the current state of v.

Bayesian network analysis

For the data-driven Bayesian network analysis, we derived an exhaustive set of causal structures between the brain signals of the latent variables of the NBMC and calculated BIC for each structure (Scutari, 2009). Specifically, to see whether the c, s and v signals are consistent with the causal graph structure that is prescribed a priori by the NBMC, we searched for the causal graph (G) whose likelihood is maximal given the time series of three brain signals, one for each of the latent variables ($\{\overleftarrow{X}_c, \overleftarrow{X}_s, \overleftarrow{X}_v\}$): $\arg\max_{\{\overleftarrow{X}_c, \overleftarrow{X}_s, \overleftarrow{X}_v\}} p(G|\{c, s, v\} = \{\overleftarrow{X}_c, \overleftarrow{X}_s, \overleftarrow{X}_v\})$. A total of 27 causal graphs can be created out of three variables nodes (3 edges x 3 edges x 3 edges between the nodes); a total of 6 triplets of brain signals can be used for $\{\overleftarrow{X}_c, \overleftarrow{X}_s, \overleftarrow{X}_v\}$ since we have three (IPL_{c1}, STG_{c3}, STG_{c5}), two ($DLPFC_{s3}$, $Cereb_{s5}$), and single (STG_{v5}) candidate brain signals for c, s, and v, respectively (3 x 2 x 1). Thus, we evaluated which of the 162 (27 edges structures x 6 nodes combinations) possible Gs is most likely. Since these Gs differ in complexity, i.e., the number of parameters, we used the Bayesian Information Criterion (BIC) values for comparison (the smaller a BIC value is, the more likely a graph is).

The categorization model

To demonstrate the history effects in the categorization model, we adapted the "exponentially weighted moving average" model that was developed by Norton and her colleagues to the sequences of stimuli in DS#1 (see Norton et al., 2017 for model description). The likelihoods of two categories ('small' and 'large' categories) over ring sizes were assumed to be gaussian distributions. A model parameter σ_m determined the trial-to-trial noise of sensory measurement, $m_{(t)}$. As σ_m becomes greater, a sensory measurement becomes noisier (more deviated from the

true stimulus). The model used the arithmetic mean of the two category likelihoods ($(\mu_{\prime large\prime(t)} + \mu_{\prime small\prime(t)})/2$), where the two category likelihoods are equal, as the decision boundary, such that 'small' and 'large' decisions were made when a sensory measurement $m_{(t)}$ was smaller and larger, respectively, than the boundary. After the trial t, the model updates only the mean of likelihood for one category that corresponds with a past decision by shifting the old mean toward a past stimulus with an exponential decaying weight, which is determined by a parameter, α. Note that trial-to-trial feedback determined the category membership in the orginal study whereas a decision did so in the current study since there was no trial-to-trial feedback in DS#1. For example, when a 'large' decision has been made, only the mean of the 'large' category is updated such that $\mu_{\prime large\prime(t+1)} = r_{(t)} + (1-\alpha)\mu_{\prime large\prime(t)}$ while the mean of the 'small' category remains unchanged ($\mu_{\prime small\prime(t+1)} = \mu_{\prime small\prime(t)}$), where $r_{(t)}$ is a reinstated sensory measurement, the noisiness of which is also governed by σ_m. As α becomes greater, the mean of the category likelihood becomes more attracted toward the past stimulus.

Statistical tests

To calculate confidence intervals, a set of bootstrap-sampled data was obtained by resampling 10^5 times with repetition, and the mean and interval size of the threshold (e.g., 95%) were then computed for the bootstrap data set. For all statistics of DS#1, 18 individuals were used except the whole brain analysis, in which statistics at some regions of the ventral lobe were calculated with 17 individuals. For DS#2-4, the numbers of subjects for statistical test are provided in Table 1. Statistical significances were calculated using two-tailed tests, except for the regression models for decoding the latent variables of the NBMC from fMRI responses.

Data and code accessibility

Codes and data for fitting NBMC to behavior data can be found in https://github.com/Heeseung-Lee/LeeLeeChoeLee2020_JN.

Results

Core properties of classification

To gain intuitions about the core properties of classification that the NBMC aims to capture, consider the scenario depicted in Figure 2. There are three apple farms that differ in fertility ('B(arren)', 'O(rdinary)', and 'F(ertile)'), and thus the overall apple size tends to be small, medium, and large in the B, O, and F farms, respectively. Farmers' job is to pick apples and sort them in size into 'small' and 'large.' In the first year, three farmers were all working in the same O farm (Fig. 2*A*, left side). In the second year, one farmer continued to work in O farm, but the remaining two moved to B and F farms respectively (Fig. 2*A*, right side). Since the apple-sorting criterion for a given farmer reflects the typical size of the apples in the farm , we can readily intuit that the criteria were similar among the farmers in the first year but have become different in the second year. This intuition relates to two historical properties of classification: (P1) c varies as a function of past experiences of stimulus, and (P2) c is more strongly attracted toward recent experiences than old ones.

Let us say that, one day in the second year, the farmers visit one another's farms and classify the apples there. First, imagine that the farmer from O farm sorts the apples from B farm ('O sorting B') or the apples from F farm ('O sorting F'; Fig. 2*B*, the first and the fourth columns in Fig. 2*D*). We can intuit that farmer O tends to classify apples as 'small' in B farm and 'large' in F farm, because apple sizes mostly fall below and above the farmer's criterion, respectively. We can also intuit that the farmer's decision uncertainty also varies from apple to apple, becoming increasingly uncertain as the apple size falls closer to the classification criterion. This exemplifies the contribution of stimulus to decision variable, when c is fixed. Next, imagine that apples from O farm are sorted by the farmer from F farm ('F sorting O') or by the farmer from B farm ('B sorting O'; Fig. 2*C*, the second and the third columns in Fig. 2*D*). Here we can intuit the

contribution of c to decision variable: the very same apples will be sorted differently, and with different degrees of uncertainty, depending on which farmer sorts them.

Intriguingly, different pairings bespeak the relativity of classification in which decision and its uncertainty are determined not by an absolute quantity of stimulus but by the relative quantity of stimulus to c. For example, the 'O sorting B' and 'F sorting O' cases differ both in absolute apple size and in farmer, but are similar in decision fraction and uncertainty distribution (the first and the second columns in Fig. 2*D*). This is because the differences in stimulus are compensated by the differences in c. Put together, the intuitions earned from the farmers' cross-visit day relate to the two remaining properties of classification: (P3) s and c have respective contributions to the variability of the decision variable for classification; (P4) interplays between s and c realize the relativity of classification.

Ring-size classification task

Over consecutive trials, observers sorted rings by size into two classes, 'small' and 'large', under moderate time pressure (Fig. 1*A*). For the main data set, in which behavioral and fMRI responses were concurrently acquired (DS#1, Table 1), rings of three marginally different radii were used to ensure that decisions were made with uncertainty. Observers were not informed about the actual number of ring sizes. The threshold (70.7% correct) ring sizes were tailored for individuals (see Methods for details). Considering the possibility that 'correct' and 'incorrect' feedback events evoke the brain activities associated with rewards (Carlson et al., 2011; Marco-Pallarés et al., 2007) or errors (Carter et al., 1998; Cavanagh and Frank, 2014; Holroyd et al., 2004), we did not provide observers with trial-to-trial feedback during the fMRI runs. Instead, we provided observers with run-to-run feedback by showing the proportion of correct responses at the end of each run of 26 trials. The actual level of task performance (mean=75.7%, SD=6.3%) was close to that aimed by the training sessions, which suggests that a negligible amount of perceptual learning occurred during data collection.

The contributions of current stimuli to classification were summarized by plotting the proportion of 'large' decisions (pL) and response time (RT) against ring size (Fig. 1*B*,*C*). As the ring size increased, the pL increased (logistic regression of decisions onto ring size, $\beta = 0.99, \mathrm{P} = 6.6\times10^{-44}$), the RT of 'large' decisions decreased (linear regression of RT onto ring size, $\beta = -0.17, \mathrm{P} = 8.8\times10^{-13}$), and the RT of 'small' decisions increased (linear regression of RT onto ring size, $\beta = 0.19, \mathrm{P} = 2.4\times10^{-16}$).

When acquiring the main data, we had to adopt specific experimental procedures, including the run-to-run feedback (instead of trial-to-trial feedback), the small (3) number of ring sizes, and the long (13.2 s) inter-trial interval, due to the concurrent collection of behavioral and fMRI data. To show that the validity of the NBMC is not limited to these procedures, we also included additional data sets (DS#2-4; Table 1; (Choe et al., 2014, 2016; Lee et al., 2016)), which were collected outside the MRI scanner with procedures different from those of the main data sets.

The normative Bayesian model of classification (NBMC)

The generative model, an observer's account for how their sensory measurements are generated, was assumed as follows: on a trial t, the true size $\mathrm{Z}_{(t)}$ is randomly sampled from a prior distribution $\mathrm{p}(\mathrm{Z})$, and the observer's sensory measurement $\mathrm{m}_{(t)}$ is another noisy sample from a likelihood distribution $\mathrm{p}(\mathrm{m}|\mathrm{Z})$ centered around $\mathrm{Z}_{(t)}$ (Fig. 3*A*). The size classification task, splitting population size values into small and large halves, can be defined as judging whether $\mathrm{Z}_{(t)}$ is larger or smaller than $\tilde{\mathrm{Z}}$, the median of the true ring sizes shown over all trials. Because the observer can access only the measured sizes of stimuli encountered so far, the Bayesian solution is to infer $\mathrm{Z}_{(t)}$ and $\tilde{\mathrm{Z}}_{(t)}$, which correspond to inferring a stimulus and a criterion, respectively, from a limited set of noisy measured sizes $\{\mathrm{m}_{(1)}, \mathrm{m}_{(2)}, \ldots, \mathrm{m}_{(t)}\}$. We modeled both Z and m as Gaussian random variables, and each was parameterized by the mean and standard deviation (SD).

Critically, we assumed that there exists an additional source of noise arising from imperfect memory retrieval; the likelihood distribution becomes less reliable as trials elapse (Fig. 3*B*). To distinguish from the likelihood of Z for the current-trial measurement $p(m_{(t)}|Z)$, we denoted the likelihood of Z for the measurement 'retrieved' from an elapsed trial $t-i$ by $p(r_{(t-i)}|Z)$ and assumed that its uncertainty increases as trials elapse (Gorgoraptis et al., 2011; Zokaei et al., 2015), as represented by the widening curves as trials elapse shown in Fig. 3*B*.

On a trial t, the Bayesian observer infers the size of a current stimulus ($Z_{(t)}$) and the criterion ($\tilde{Z}_{(t)}$) by inversely propagating $\{\ldots, r_{(t-2)}, r_{(t-1)}, m_{(t)}\}$ over the generative model. The inferred stimulus size, $s_{(t)}$, is the most probable value of Z given the measurement on a current trial, as captured by the following equation:

$$s_{(t)} = \hat{Z}_{(t)} = \arg\max_{Z} p(Z|\, m_{(t)})$$

$$\propto \arg\max_{Z} p(\, m_{(t)}|Z)p(Z) \qquad \text{(Equation 1)}$$

, where the variance of the posterior reflects the precision of $s_{(t)}$ (Fig. 3*C*).

On the other hand, the inferred criterion, $c_{(t)}$, which corresponds to the inferred value of $\tilde{Z}$, is the most probable value of Z given the measurements over elapsed trials, as captured by the following equation:

$$c_{(t)} = \hat{\tilde{Z}}_{(\ldots,t-2,t-1)} = \arg\max_{Z} p(Z|\, \mathbb{r}_{(t)})$$

$$\propto \arg\max_{Z} p(\, \mathbb{r}_{(t)}|Z)p(Z) \qquad \text{(Equation 2)}$$

, where $\mathbb{r}_{(t)} = \{\ldots, r_{(t-2)}, r_{(t-1)}\}$. Here, the variance of the posterior, which reflects the precision of $c_{(t)}$, is the optimally weighted sum of variances of the retrieved measurements and the prior size (Fig. 3*D*). Equation 2 implies that $c_{(t)}$ is more attracted towards a recent stimulus than to

older ones because the representational uncertainty of working memory increases as trials elapse (Fig. 3*D*). Note that $s_{(t)}$ and $c_{(t)}$ are the optimal inferences of the current stimuli and criterion given the generative model of the Bayesian observer because taking the maximum value of a posterior minimizes estimation errors when the posterior follows a Gaussian distribution (Jaynes, 2003; Körding and Wolpert, 2004).

Having estimated $s_{(t)}$ and $c_{(t)}$, the Bayesian observer performs the task by deducing a decision variable $v_{(t)}$ from $s_{(t)}$ and $c_{(t)}$ and translating it into a binary decision $d_{(t)}$ with a degree of uncertainty $u_{(t)}$ (Fig. 3*E*); $v_{(t)}$ is the probability that $s_{(t)}$ will be greater than $c_{(t)}$ ($p(s_{(t)} > c_{(t)})$); $d_{(t)}$ 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 ($p(s_{(t)} < c_{(t)}|d_{(t)} = \text{'large'})$ or $p(s_{(t)} > c_{(t)}|d_{(t)} = \text{'small'})$) (Sanders et al., 2016).

Figure 3F and 3G graphically summarize how the NBMC works. Figure 3F illustrates how c is updated as a function of past stimuli, and takes part, with s, in classification by determining v and u over an example stream of trials: c tends to run after s, as implied by Equation 2; when $s_{(t)}$ is larger (smaller) than $c_{(t)}$, $v_{(t)}$ was larger (smaller) than 0.5 and thus $d_{(t)} = \text{'large'}$ ('small'); as $s_{(t)}$ becomes closer to $c_{(t)}$, $u_{(t)}$ becomes higher. Also, note that the congruence between a ring size and $d_{(t)}$ can descriptively capture the dynamic of $u_{(t)}$: $u_{(t)}$ is lower when a stimulus and a decision are incongruent. The bivariate decision space shown in Figure 3G shows a snapshot of how the variables comprising decision outcomes ($v_{(t)}$, $d_{(t)}$, and $u_{(t)}$) are deduced from the posterior probability distributions of $s_{(t)}$ and $c_{(t)}$ on an example (t=3) trial.

In summary, according to the NBMC, as the ring size (Z) varies over trials, the Bayesian observer continues to (i) make the inferences of current stimuli and criteria (s and c), (ii) deduce v from s and c, and (iii) make decisions (d) with varying degrees of uncertainty (u). To verify the implications of the NBMC, we took the following steps. First, for each human observer, we created a Bayesian partner by fitting the NBMC to the decisions of that observer. Second, we made these Bayesian observers simulate decisions (d) and decision uncertainties (u) on the very

651 sequences of stimuli that were encountered by their corresponding human partners. Then, we
652 examined the correspondences between the human and the Bayesian observers regarding each
653 of the aforementioned four core properties of classification (P1~4).
654
655 **History effects of past stimuli on classification (P1 and P2)**
656 According to the c-inference algorithm of the NBMC (Equation 2; Fig. 3*D*), c is affected only by
657 past stimuli but not by past decisions, and is attracted toward past stimuli more strongly toward
658 recent ones than toward old ones (Fig. 3*D,F*). Then, according to the deduction algorithm of the
659 NBMC (Fig. 3*E*), such attractive shifts of c toward past stimuli will be translated into the decisions
660 that are repelled away from past stimuli, more strongly from recent ones than from old ones.
661 This model behavior realizes the two historical properties of classification that were intuited in
662 the apple-classification scenario (P1 and P2).
663 To check the correspondences between the Bayesian observers and the humans
664 regarding P1 and P2, we regressed current decisions that were made by the Bayesian observers
665 and by the humans onto past stimuli, current stimuli and past decisions (Fig. 4). As implied by the
666 NBMC, for both of the Bayesian observers (Fig. 4*A*) and the humans (Fig. 4*B*), current decisions
667 were more strongly repelled from more recent stimuli (P2), while effects of past decisions on
668 current decisions are negligible (P1). The regression coefficients were well matched between the
669 Bayesian observers and the humans. The above correspondences were also replicated in the
670 other three data sets (Fig. 4*C-E*), which supports the generalizability of the NBMC's account for
671 the history effects in classification. For all of the data sets, the repulsive history effects of past
672 stimuli on current decisions could not be simulated by an alternative model ('constant-c model';
673 Fig. 4*F*), in which c was fixed at a constant instead of being inferred from past sensory
674 measurements (see Methods for detailed model description).
675 The c-inference and deduction algorithms of the NBMC also imply specific history effects
676 of past stimuli on decision uncertainty ($u_{(t)}$): the decision uncertainty on a current trial varies as a

677 function of the congruence of a current decision with only past stimuli, but not with past
678 decisions. This is because, as $c_{(t)}$ is attracted towards a past stimulus, $s_{(t)}$ that would lead to the
679 decision congruent with that past stimulus (e.g., making a 'large' decision on a current trial after
680 viewing an L ring on a previous trial) tends to be closer to $c_{(t)}$ and thus be more uncertain
681 compared to $s_{(t)}$ that would lead to the decision incongruent with that past stimulus (e.g.,
682 making a 'small' decision on a current trial after viewing an L ring in a previous trial). In the case
683 of 'F sorting O' in the apple-classification scenario, which is depicted in the second column of
684 Figure 2*D*, the congruent and incongruent conditions correspond to the red and blue circles,
685 respectively. As a result, $u_{(t)}$ is high when $d_{(t)}$ is congruent with past stimuli (P1). And, again
686 because c is more attracted towards recent stimuli than towards old ones, this positive history
687 effect of decision-congruent past stimuli on $u_{(t)}$ should be stronger for recent stimuli than for old
688 ones (P2).

689 Unlike decisions, decision uncertainty was not directly measured from the human
690 observers. Instead, to compare with the Bayesian observers regarding history effects on decision
691 uncertainty, we used RT as a proxy of decision uncertainty of the human observers based on
692 their tight linkage reported by previous studies (Palmer et al., 2005; Ratcliff and McKoon, 2008;
693 Urai et al., 2017). Then, the simulated values of $u_{(t)}$ and the RT measurements, for the Bayesian
694 observers and the humans respectively, were regressed onto both the congruence of $d_{(t)}$ with
695 past stimuli ('D-S congruence') and the congruence of $d_{(t)}$ with past decisions ('D-D congruence')
696 (Fig. 5). The patterns of regression coefficients were consistent between the Bayesian observers
697 and the humans while satisfying the history effects on decision uncertainty that are implied by
698 the NBMC. As in the decision data, these results were replicated for all the data sets and could
699 not be simulated with the constant-c model.

700 As another means of testing the model implications regarding history effects on decision
701 uncertainty, we examined the fMRI responses from the human observers in search of cortical
702 activities that are correlated with $u_{(t)}$. We found such activities in the dorsal anterior cingulate

703 cortex (dACC) and bilateral anterior insula (Ins), which were known to be associated with
704 decision uncertainty in the literature (Grinband et al., 2006; Shenhav et al., 2014; Sheth et al.,
705 2012). In these regions, the fMRI responses at 5.5s from stimulus onset, which are likely to reflect
706 the neural activities at around the moment of decision making considering the typical
707 hemodynamic delay of fMRI signal, were significantly correlated with $u_{(t)}$ (Table 2; Fig. 6*A*,*C*).
708 Importantly, when we carried out the regression analysis on those fMRI responses as we did on
709 the RT data, the dACC responses were significantly regressed onto the congruence of $d_{(t)}$ with
710 past stimuli but not onto the congruence of $d_{(t)}$ with past decisions (Fig. 6*D*), which corresponds
711 with the history effects of past stimuli on $u_{(t)}$ that are implied by the NBMC. However, the history
712 effect was not statistically significant in the Ins, which indicates that the dACC's activity most
713 faithfully reflects the changes in decision uncertainty as predicted by the NBMC.
714 Furthermore, when the values of $u_{(t)}$ that were estimated by the constant-c model were
715 used instead of those estimated by the NBMC, we could not identify any cortical activities that
716 are significantly correlated with $u_{(t)}$, including the dACC and Ins (Fig. 6*B*). This suggests the
717 critical usefulness of the c-inference algorithm in searching for the decision-uncertainty signal in
718 the brain activity of humans engaged in a classification task.
719
720 **Dissecting stimulus and criterion contributions to the trial-to-trial variability of classification (P3)**
721 According to the s-inference (Equation 1) and c-inference (Equation 2) algorithms, $s_{(t)}$ and $c_{(t)}$ are
722 inferred based on independent evidences (current and past sensory measurements, respectively)
723 and thus vary independently across trials. According to the deduction algorithms (Fig. 3*E*), the
724 variabilities of $d_{(t)}$ that originate from $c_{(t)}$ and from $s_{(t)}$ can be dissected by regressing the
725 variability of $d_{(t)}$ under a fixed value of $s_{(t)}$ onto the variability of $c_{(t)}$ (left panel of Fig. 7*A*) and by
726 regressing the variability of $d_{(t)}$ under a fixed value of $c_{(t)}$ onto the variability of $s_{(t)}$ (right panel
727 of Fig. 7*A*). This implication of the NBMC captures the third property of classification (P3).
728 Rephrasing in terms of the apple-classification scenario, the NBMC specifies how much of the

variability of classification is ascribed to the farm from which an apple came (Fig. 2*B*) and to the farm from which a farmer came (Fig. 2*C*).

To check how well the NBMC's dissection of $s_{(t)}$ and $c_{(t)}$ contributions to the variability of $d_{(t)}$ captures the corresponding dissection in the human data, we proceeded as follows. For each stimulus condition, we first binned the trials according to the simulated values of $c_{(t)}$ and calculated the fractions of 'large' decision (pL) of the Bayesian observers at those bins (green symbols in Fig. 7*B-E*). Then, we calculated the pL of the human observers at each of all the above bins (blue or red symbols in Fig. 7*B-E*). At all of the bins, the pLs of the humans matched those of the Bayesian observers. Qualitatively, we regressed the human decisions onto current stimulus and the criterion inferred by the NBMC, as follows: $\mathrm{D}_{(t)} \sim \beta_0 + \beta_c \mu_{c_{(t)}} + \beta_Z \mathrm{Z}_{(t)}$. As predicted by the NBMC, the decisions of the humans were concurrently regressed onto $c_{(t)}$ negatively (multiple logistic regressions; $\beta_c(\mathrm{P}) = -0.59\ (2.3\times10^{-20})$, $-0.54\ (2.5\times10^{-29})$, $-0.25\ (5.4\times10^{-19})$, and $-0.35\ (3.9\times10^{-46})$ for DS#1~4) and onto the current stimulus positively ($\beta_Z(\mathrm{P}) = 1.08\ (1.8\times10^{-35})$, $1.36\ (9.9\times10^{-87})$, $1.46\ (10^{-323})$, and $1.44\ (6.9\times10^{-71})$ for DS#1~4) for all of the data sets. Quantitatively, for all of the data sets, the Bayesian pLs fell within the 95% confidence intervals of the human pLs defined at most of the bins inspected.

Similar to the variability of $d_{(t)}$, the variability of $u_{(t)}$ that originates from $c_{(t)}$ and from $s_{(t)}$ can be dissected by regressing the variability of $u_{(t)}$ under a fixed value of $s_{(t)}$ onto the variability of $c_{(t)}$ (left panel of Fig. 8*A*) and by regressing the variability of $u_{(t)}$ under a fixed value of $c_{(t)}$ onto the variability of $s_{(t)}$ (right panel of Fig. 8*A*). However, unlike $d_{(t)}$, which has a monotonic relationship with $v_{(t)}$, the relationship of $u_{(t)}$ with $v_{(t)}$ is contingent on decisions, being positive for 'small' decisions and negative for 'large' decisions. For this reason, to check the correspondence in $s_{(t)}$ and $c_{(t)}$ contributions to the variability of $u_{(t)}$ between the Bayesian observers and the humans, we followed the procedure identical to that used for $d_{(t)}$ except that data bins were further conditioned on current decisions. At most of the bins, the RTs and dACC responses of the humans, which can be considered as the behavioral and neural correlates of $u_{(t)}$

as mentioned earlier, matched the $u_{(t)}$ values of the Bayesian observers. Qualitatively, we regressed the RTs in all the data sets, and the dACC responses in DS#1 as well, onto $c_{(t)}$ with stimulus sizes, as follows: $\mathrm{RT(or\ dACC)} \sim \beta_0 + \beta_c \mu_{c_{(t)}} + \beta_Z Z_{(t)}$. As predicted by the NBMC (Fig. 8*B-F*), β_c was positive for 'large' decisions ($\beta_c(\mathrm{P}) = 0.12\ (2.5\times10^{-5})$, $0.16\ (2.5\times10^{-10})$, $0.085\ (5.1\times10^{-8})$, and $0.051\ (1.7\times10^{-7})$ for RT in DS#1~4 and $0.084\ (5.7\times10^{-4})$ for dACC in DS#1) and negative for 'small' decisions ($\beta_c(\mathrm{P}) = -0.055\ (0.053)$, $-0.21\ (2.4\times10^{-15})$, $-0.038\ (0.013)$, and $-0.058\ (9.5\times10^{-6})$ for RT in DS#1~4 and $-0.059\ (0.020)$ for dACC in DS#1). Also, β_Z was negative for 'large' decisions ($\beta_Z(\mathrm{P}) = -0.19\ (6.9\times10^{-15})$, $-0.21\ (2.4\times10^{-15})$, $-0.23\ (6.0\times10^{-50})$, and $-0.097\ (2.8\times10^{-7})$ for RT in DS#1~4 and $-0.054\ (0.053)$ for dACC in DS#1) and positive for 'small' decisions ($\beta_Z(\mathrm{P}) = 0.20\ (3.5\times10^{-16})$, $0.28\ (7.9\times10^{-28})$, $0.21\ (3.5\times10^{-42})$, and $0.12\ (1.0\times10^{-12})$ for RT in DS#1~4 and $0.11\ (5.8\times10^{-5})$ for dACC in DS#1). Quantitatively, for all of the data sets, the Bayesian simulations of $u_{(t)}$ fell within the 95% confidence intervals of the human RT and dACC data defined at most of the bins inspected.

The relativity of classification (P4)

According to the NBMC, the decision variable $v_{(t)}$, which determines $d_{(t)}$ and $u_{(t)}$, is defined as the difference between $s_{(t)}$ and $c_{(t)}$. This implication realizes the relativity of classification, the last property that the NBMC aims to capture (P4): classification is not governed by the absolute magnitude of stimuli but by the relative magnitude of s to c, as intuited with the different pairings of the farms from which apples and farmers come in the apple-classification scenario. Figure 9A and 10A illustrate how the NBMC captures the relativity of classification in terms of $pL_{(t)}$ and $u_{(t)}$, respectively: $pL_{(t)}$ and $u_{(t)}$ should both be distinguishable as the differences between $s_{(t)}$ and $c_{(t)}$ are unmatched even though physical stimuli are the same (the left two panels of Fig. 9*A* and Fig. 10*A*) and indistinguishable as long as the differences are matched even though physical stimuli are different (the right two panels of Fig. 9*A* and Fig. 10*A*).

To evaluate how well the NBMC captures the relativity of human classification in $pL_{(t)}$, we checked the correspondence between the Bayesian observers and the humans in the following two model implications. First, because what governs classification is 's relative to c', $pL_{(t)}$ should vary only as a single, monotonic function of the difference between $s_{(t)}$ and $c_{(t)}$. When trials were sorted by the differences between the simulated values of $s_{(t)}$ and $c_{(t)}$, the $pL_{(t)}$ values of the humans constituted single, seamless psychometric curves for all of the data sets (red or blue symbols in Fig. 9*B-E*), and most of the $pL_{(t)}$ values of the Bayesian observers (green symbols of Fig. 9*B*-E) fell within the 95% confidence intervals of the human $pL_{(t)}$.

Second, again because what governs classification is 's relative to c', $pL_{(t)}$ should become 'metameric' between the trials in which a physical difference between stimuli is compensated by a counteracting difference in $c_{(t)}$ (e.g., dotted box in Fig. 9*B*) or 'anti-metameric' between the trials in which a physical sameness between stimuli is accompanied by a substantial difference in $c_{(t)}$ (e.g., dashed boxes in Fig. 9*B*). These predictions were statistically confirmed by the Bayes factor analysis across all the possible pairs of trial bins (see Methods for the procedure for the Bayes factor analysis). The distributions of the "significantly indistinguishable" trials ('o' markers in the left columns of Fig. 9*F*; Bayes factor $< 1/3$) and the "significantly distinguishable" trials ('x' markers in the right columns of Fig. 9*F*; Bayes factor > 3) coincided with the metameric relations predicted by the NBMC. Figure 9G, which puts together the panels in Figure 9F, demonstrates that the NBMC readily captures the relativity of classification in a single currency, the difference between $s_{(t)}$ and $c_{(t)}$.

The RT and dACC data, though being noisier than the decision fraction data, also supported the Bayesian-human correspondences in $u_{(t)}$ regarding the relativity of classification. When trials were sorted by the differences between the simulated values of $s_{(t)}$ and $c_{(t)}$, the human RT (black symbols in Fig. 10*B*, *D-F*) and dACC (black symbols in Fig. 10*C*) responses constituted single, seamless psychometric curves for all of the data sets, and most of the $u_{(t)}$

values of the Bayesian observers (green symbols of Fig. 10*B*-F) fell within the 95% confidence intervals of the human RT (black symbols in Fig. 10*B-E*). Furthermore, the pairs of trials in which the human RT and dACC responses became significantly metameric ('o' markers in the upper right panels of Fig. 10*B* and Fig. 10*C*; Bayes factor $< 1/3$) and significantly anti-metameric ('x' markers in the bottom right panels of Fig. 10*B* and Fig. 10*C*; Bayes factor > 3) were found.

Brain signals of inferred stimulus, criterion, and decision variable

The Bayesian observer of the NBMC forms the five variables that are causally related with one another, $s_{(t)}$, $c_{(t)}$, $v_{(t)}$, $d_{(t)}$, and $u_{(t)}$. So far, to verify the implications of the NBMC, we have shown that its simulation of $d_{(t)}$ and $u_{(t)}$ qualitatively and quantitatively captures the core properties of classification (P1~4) that are expressed in the pL and RT (or dACC activity) measurements acquired from the human observers. As indicated by the causal structure of the variables (Fig. 3*E*), the simulated values of $d_{(t)}$ and $u_{(t)}$ were derived from their parent variables, $s_{(t)}$, $c_{(t)}$, and $v_{(t)}$, the trial-to-trial latent states of which were predicted by the NBMC. This allows us to examine the fMRI responses of the human observers in search of the brain activities that represent $s_{(t)}$, $c_{(t)}$, and $v_{(t)}$. Such brain activities should satisfy the following two requirements: first ('correlation' requirement), for each brain signal, its trial-to-trial states must be correlated with those of its target variable; second ('causality' requirement), its correlations with the other remaining (non-target) variables in the NBMC must be consistent with the causal relationships of its target variable with the remaining variables.

Initially, we identified the candidate brain regions within which activity patterns satisfy the first 'correlation' requirement using the multivoxel pattern analysis (MVPA) in conjunction with a searchlight technique, which is known to be highly sensitive to detect brain signals in local patterns of population fMRI responses (Kriegeskorte et al., 2006). However, this approach is limited in specificity: successful decoding of a variable from certain patterns of brain activity does not necessarily indicate the representation of that variable per se because such decoding is also

831 possible even when those patterns of activity indeed represent any other variables that are
832 correlated with the variable of interest.

833 To address this specificity problem, we imposed the second 'causality' requirement on
834 the candidate brain regions. Specifically, we deduced an exhaustive set of regressions (Fig. 11*D-F*)
835 that are implied from the causal relationships of a target variable with other variables (Fig. 11*A-C*),
836 and tested whether a candidate brain signal is consistent with the entire set of regressions. For
837 example, to claim that a candidate brain signal of c satisfies the 'causality' relationship between c
838 and v, we must show that that the c decoded from the brain signal is (i) negatively regressed
839 onto v (the 5th regression in Fig. 11*D*) but (ii) not negatively regressed onto v any more when the
840 causal influence of c is removed in v (the 6th regression in Fig. 11*D*), which are implied by the
841 negative causality between c and v (the top left arrow in Fig. 11*A*; the histogram of correlation
842 coefficients between c and v in Fig. 11*B*). Additional regressions can be further deduced from the
843 causal relationship of c with the remaining variables (s, u, d, Z), resulting in a total of 13
844 regressions for c (Fig. 11*D*). Likewise, we deduced 13 and 16 regressions for s and v, respectively
845 (Fig. 11*D-F*; see Methods for detailed definitions of the regressions).

846 Six brain regions satisfied both of the two requirements (Table 3; Fig. 12*A*). The signal of c
847 appeared in three different regions at different time points relative to stimulus onset, as follows:
848 a region in the inferior parietal lobe at 1.1 s ($\mathrm{IPL_{c1}}$), followed by two regions in the superior
849 temporal gyrus at 3.3 and 5.5 s ($\mathrm{STG_{c3}}$, $\mathrm{STG_{c5}}$). The signal of s appeared both in the dorsolateral
850 prefrontal cortex at 3.3 s ($\mathrm{DLPFC_{s3}}$) and in the cerebellum at 5.5 s ($\mathrm{Cereb_{s5}}$). The signal of v
851 appeared in the left superior temporal gyrus at 5.5 s ($\mathrm{STG_{v5}}$).

852 We also confirmed that the concurrent trial-to-trial variabilities of the decoded variables
853 from the brain signals were consistent with the causal structure defined by the NBMC, as follows.
854 First, the interplays between the decoded c, s, and v were well captured by a multiple regression
855 model, $\mathrm{STG_{v5}} \sim \beta_{s3}\mathrm{DLPFC_{s3}} + \beta_{s5}\mathrm{Cereb_{s5}} + \beta_{c1}\mathrm{IPL_{c1}} + \beta_{c3}\mathrm{STG_{c3}} + \beta_{c5}\mathrm{STG_{c5}}$, which was derived
856 from the 'common-effect' structure of causal relationships between the variables ($\mathrm{c} \rightarrow \mathrm{v} \leftarrow \mathrm{s}$;

the top edges of Fig. 11*A*). The regression coefficients matched the causal functions in sign for all of the five brain signals and were all significant except for $\mathrm{DLPFC_{s3}}$ ($\beta_{s3} = 0.011, \mathrm{P} = 0.55$; $\beta_{s5} = 0.047, \mathrm{P} = 0.013$; $\beta_{c1} = -0.046, \mathrm{P} = 0.0088$; $\beta_{c3} = -0.056, \mathrm{P} = 0.0037$; $\beta_{c5} = -0.15, \mathrm{P} = 1.9\times10^{-17}$; Fig. 12*B*). Second, the decision-dependent interplay between v and u signals was successfully captured by an interactive regression model, $\mathrm{dACC_{u5}} \sim \beta_{\mathrm{v}}\mathrm{STG_{v5}} + \beta_{\mathrm{d}}\mathrm{d} + \beta_{\mathrm{d\times v}}\mathrm{d}\times \mathrm{STG_{v5}}$, which was derived from the 'common-cause' structure of causal relationships between the variables ($\mathrm{d} \leftarrow \mathrm{v} \rightarrow \mathrm{u}$; the bottom edges of Fig. 11*A*). Critically, because the sign of the causal function from v to u depends on the value of d, the NBMC predicts the interaction between d and $\mathrm{STG_{v5}}$. As predicted, $\mathrm{dACC_{u5}}$ was significantly regressed onto $\mathrm{d\times STG_{v5}}$ ($\beta_{\mathrm{d\times v}} = -0.040, \mathrm{P} = 0.04$; the left bar in Fig. 12*C*). The causal structure of '$\mathrm{d} \leftarrow \mathrm{v} \rightarrow \mathrm{u}$' was further supported when the RT data ($\beta_{\mathrm{d\times v}} = -0.045, \mathrm{P} = 0.013$) or the model estimates of decision uncertainty $\mathrm{u_{(t)}}$($\beta_{\mathrm{d\times v}} = -0.054, \mathrm{P} = 0.0012$), instead of $\mathrm{dACC_{u5}}$, were regressed onto d and $\mathrm{STG_{v5}}$ (the middle and right bars in Fig. 12*C*).

Finally, as a data-driven approach that complements the model-driven regression analyses performed above, we calculated the Bayesian Information Criterion (BIC) for each of the possible causal structures between the brain signals of c, s, and v (Scutari, 2009); see Methods for BIC calculation). The outcomes of BIC evaluation were consistent with the NBMC in the following two aspects. First, out of the 162 possible causal graph structures, the smallest (best) BIC value was found for '$\mathrm{c} \rightarrow \mathrm{v} \leftarrow \mathrm{s}$' (Fig. 12*D*). This causal graph of 'common effect' is exactly the one prescribed by the NBMC (Fig. 11*A*). Second, note that one critical outcome that falsifies the NBMC is the 'significant' existence of any graph that includes causal arrows between c and s because the NBMC is built upon the assumption that c and s are the random variables that are independent from one another. To evaluate the statistical significance of those graphs, we followed the statistical convention that any pairs of hypotheses with the BIC differences greater than 2 are considered to be sufficient to conclude significant differences (Kass and Raftery, 1995). With this critical value of statistical significance (BIC>2), we found that none of the graphs with

the causal arrows between c and s could pass the criterion. The closest group of the graphs (shown at the bottom of Fig. 12*D*) was well apart from the winner graph, by more than the BIC value of 3.

In sum, the outcomes of the fMRI data analyses indicate that there exist the brain activities that respectively represent the key latent variables of the NBMC (c, s, and v) and concurrently vary in the ways consistent with the causal relationships of those variables with themselves and the other variables in the NBMC.

Different history effects in classification and categorization

Although "classification" and "categorization" have been used interchangeably, as footnoted in Introduction, the two tasks should not be confused with each other (Jacob, 2004) since they fundamentally differ in statistical structure. Whereas classification requires judging which side of a single distribution (i.e., 'which class') a given entity belongs to relative to a criterion (Fig. 13*A*), categorization requires judging which of two distributions (i.e., 'which category') a given entity belongs to (Fig. 13*B*). In the context of the apple-sorting scenario, classification is to decide whether a given apple is smaller or larger than the typical size of apples while categorization is to decide whether a given apple is from the B farm or from the F farm.

Due to this difference in statistical structure, classification and categorization predict different history effects. The history effects in classification occur as an observer updates the single, entire distribution of stimuli in her generative model through recent experience of stimuli, and thus the classification criterion (the median of the distribution). Critically, past decisions are not involved at all in the history effects in classification because past decisions do not provide any information about the distribution of stimuli (Fig. 13*C*)—only past experience of stimuli can attractively shift the distribution of stimuli and thus the criterion (Fig. 13*E*). As results, only past stimuli, but not past decisions, bring about negative history effects on current decisions as observed in the current study (Fig. 4).

By contrast, the history effects in categorization occur as an observer selectively updates the two category likelihoods over stimuli in his generative model. Importantly, unlike in classification, 'past decisions' contribute to the history effects in categorization by determining which of the two category likelihoods is updated because assigning a stimulus to a certain decision means that the stimulus belongs to the category corresponding to that decision (Fig. 13*D*). On the other hand, like in classification, 'past stimuli' attractively shift only that selected category likelihood toward themselves (Fig. 13*F*). As results, in categorization, 'past decisions' tend to be repeated on current trials while 'past stimuli' tend to repel current decisions. Alternatively, the history effects in categorization can also occur as an observer updates the prior probabilities of the categories in her generative model, as suggested recently (Glaze et al., 2015). Even in this alternative updating, past decisions still contribute to the history effects by determining which of the two category priors is increased (or decreased, depending on other assumptions in the generative model such as 'hazard' (Glaze et al., 2015)). In sum, due to the difference in statistical structure, classification and categorization critically differ in history effect in that 'past decision' is a critical factor in categorization but not in classification.

To concretely demonstrate the history effects in categorization, we adapted one of recent models of categorization (Norton et al., 2017) to our experiment (see Methods for model specifications) and examined the history effects through model simulations. This particular model, which updates the likelihoods of categories over stimuli and was called the "exponentially weighted moving average" model in Norton et al. (2017), was chosen because it best captured human categorization while exhibiting the history effect of past stimuli similar to that in the current study. When we carried out a regression analysis on the simulated data, as was done to probe the history effects on decisions for the NBMC (Fig. 4), the regression coefficients of past stimuli and past decisions were opposite in sign (negative for past stimuli and positive for past decision) and similar in magnitude (green symbols in Fig. 13*G*). These history effects differed from those observed for the humans in our experiments (black and gray symbols in Fig. 13*G*). This

difference could not be resolved by changing the model parameters: the parameter set that best captures the stimulus effect fails to capture the decision effect (top panel of Fig. 13G), and vice versa (bottom panel of Fig. 13G).

Discussion

By specifying the statistical structure of classification, we developed the NBMC, in which a normative observer infers the stimulus and criterion from current and past sensory evidences, respectively, and commits to a decision with a varying degree of uncertainty. The NBMC well captured human classification behavior and neural activity by precisely defining the latent process of criterion inference. Furthermore, we demonstrated how the difference in statistical structure between classification and categorization leads to the difference in history effect.

Normative inference of classification criterion from sensory evidences with memory decay

By characterizing our model as "normative", we mean that, under the assumption that sensory traces of past stimuli become noisier with memory decay, the optimal way of inferring the center of past stimuli is to probabilistically integrate the likelihoods by taking into account their increasing uncertainties associated with memory decay since such integration minimizes the squared error of inferred estimates (Jaynes, 2003; Körding and Wolpert, 2004). To be sure, for an agent without memory decay, this algorithm is suboptimal because the inferred criterion keeps digressing from the true, fixed criterion by being swayed with recent stimuli. In this sense, the c-inference algorithm proposed here can be considered as an instantiation of the bounded rationality of an agent with limited cognitive resources (Simon, 1955).

Classification versus categorization

Classification requires learning about a criterion, which is arbitrarily set and divides a distribution of entities into mutually exclusive classes. For example, classifying the roasting degrees of coffee beans into 'dark-medium' or 'light-medium' requires setting an arbitrary criterion that divides a given distribution of roasting degrees into two exclusive parts. Here, 'dark-medium' or 'light-

963 medium' is a relative adjective that is assigned to a bean via comparison against an arbitrary
964 criterion.

965 On the other hand, categorization requires a knowledge about how feature values are
966 distributed for given categories (Glaze et al., 2015; Gold and Shadlen, 2007; Green and Swets,
967 1966; Maddox, 2002; Norton et al., 2017; Qamar et al., 2013). For example, categorizing the origin
968 of coffee beans into 'Ethiopia' or 'Brazil' based on color requires knowing about the respective
969 color distributions of the Ethiopian and Brazilian beans. Here, the categorical membership of a
970 given bean is determined by the distributions of color feature intrinsic to the geographical origin
971 of the bean, not by any arbitrary criterion value of color. Therefore, unlike in classification, the
972 color distributions may be overlapped between 'Ethiopian' and 'Brazilian' beans.

973 Due to this difference in statistical structure, classification and categorization entail
974 different generative models and inferential processes. As prescribed by the NBMC, the ideal
975 classification is to infer both the criterion (c) that divides the 'dark-medium' and 'light-medium'
976 classes and the amount of roasting (s) by propagating the information gleaned from past (r) and
977 current sensory (m) measurements of roasting, respectively, and then to commit to a decision by
978 computing the posterior probability that the amount of roasting is greater than the criterion,
979 $p(s > c|m, r)$. By contrast, the ideal categorization is to infer the posterior probabilities of
980 'Ethiopia' and 'Brazil' given sensory measurements as $p(\mathrm{Ethoipia}|m)$ and $p(\mathrm{Brazil}|m)$ and then to
981 judge which category has the higher posterior.

982 Although the size classification task is performed ideally by the NBMC, human observers
983 could have performed the task by adopting the categorization model because doing so leads to a
984 "satisfactory" level of performance though not the ideal level. However, this seems unlikely
985 because the observers in our study did not exhibit history effects of past decisions on current
986 decisions whereas an observer who adopts the generative model of categorization would exhibit
987 substantive history effects of past decisions irrespective of whether the observer updates the
988 likelihoods or priors of categories. Our results suggest that researchers who develop or use any

binary decision-making tasks need to specify whether their task entails classification or categorization and to apply an appropriate model accordingly. We stress that the NBMC should be considered as a computational model for binary classification tasks in which a criterion is not explicitly provided.

History effects in classification via dynamic criterion update

The history effects in sequential perceptual decision-making have recently received considerable attention. The NBMC offers a normative account for the repulsive effect of past stimuli on classification of a current stimulus by specifying how the Bayesian inference of the criterion is dynamically updated from past sensory evidences. This account is distinct from those offered by previous studies in the following aspects.

First, in categorization tasks, history effects occur as the likelihoods (Norton et al., 2017) or priors (Glaze et al., 2015; Glaze et al., 2018) of categories are updated. As shown in the last section of Results, these categorization-based accounts are different from our classification-based account in that they predict the systematic influences of past decisions on current decisions.

Second, unlike in many previous accounts (Burr and Cicchini, 2014; Dyjas et al., 2012; Fritsche et al., 2017; Glaze et al., 2015; Raviv et al., 2012; Treisman and Williams, 1984), any specific assumption regarding environmental volatility is not required in our account for history effects. Instead, our account is grounded in the assumption that the uncertainty of a past sensory event increases as trials elapse (Gorgoraptis et al., 2011; Zokaei et al., 2015).

Third, as proposed previously (Fritsche et al., 2017; Schwartz et al., 2007; Stocker and Simoncelli, 2006), sensory adaptation might be considered as an alternative origin for the repulsive history effect of past stimuli because an exposure to a specific stimulus would induce a repulsive bias on a population neural representation. However, this account seems unlikely because stimuli were presented very briefly (0.3s) and the history effect of past stimuli sustained

more than 26.4s in the current study whereas the neural bias due to sensory adaptation requires a long duration of exposure and lasts for a short period of time (Nakashima and Sugita, 2017; Pavan et al., 2012).

Brain signals of inferred criterion

The brain signals of c appeared initially at the IPL and then at the STG, in the vicinity of which the signal of v resides. We conjecture that working-memory representations of past stimuli are likely to be formed in the IPL and then are transferred to the STG, wherein they take part in classification of current stimuli. This conjecture is congruous with the literature in several aspects.

Optical inactivation of the posterior parietal cortex (PPC) of rat – a region presumed to be a rat homologue of the primate IPL (Goard et al., 2016; Hanks et al., 2015; Roitman and Shadlen, 2002) – selectively diminished the influence of previous stimuli but left the task performance on current stimuli intact (Akrami et al., 2018), suggesting a critical role of the IPL in encoding previous stimulus information. In addition, perturbing the PPC after current stimulus onset failed to diminish history effects, suggesting that the information of previous stimuli in PPC was likely to be transferred to some brain regions outside the PPC (Hwang et al., 2017). On the other hand, the c signal at the STG seems consistent with its suggested roles in coordinating the spatial reference frame (Karnath, 2001; Karnath et al., 2001; Karnath et al., 2002), in that c in our task works as a spatial reference against which s is compared.

One previous fMRI study suggested the inferior temporal pole region as the locus of c (White et al., 2012). Unfortunately, we could not confirm this finding, because this region was not fully imaged for all participants (see Methods). Nonetheless, their failure of finding neural correlates of c in the IPL or STG could be due to the fact that the effect of past stimuli was not considered in defining c.

Brain signals of inferred stimulus and decision variable

1041 The s signal in the DLPFC is compatible with its suggested roles in maintaining current sensory
1042 information during decision-making (Curtis and Lee, 2010; Haller et al., 2018). The s signal in the
1043 cerebellum is also in line with previous studies indicating the involvement of the cerebellum in
1044 sensory perception (Baumann et al., 2015; Gao et al., 1996). By contrast, we failed to find any
1045 signal of s in the early retinotopic visual areas. Is this failure due to the relatively poor spatial
1046 resolution of the fMRI imaging protocol used in the current study? To check this possibility, we
1047 revisited one of our previous imaging studies (Choe et al., 2014), where high-resolution fMRI
1048 responses were acquired in V1 while observers performed the same task as in the current study,
1049 and examined the V1 population activity with a fine retinotopy-based population decoding
1050 analysis. However, unlike the DLPFC and cerebellum, where the s signals were regressed onto
1051 both external stimuli (Z) and decisions (d) (Fig. 14*A*,*B*), the V1 activity was regressed only onto
1052 external stimuli (Z) but not onto decisions (d) (Fig. 14*C*). This implies, as concluded previously
1053 (Choe et al., 2014; Jasper et al., 2019; Lee et al., 2007; Nienborg and Cumming, 2009), that the V1
1054 population activity represents physical stimuli (Z) robustly but its trial-to-trial variability does not
1055 seem to be strongly involved in determining the trial-to-trial variability in choice (d).

1056 **References**

1057

1058 Akrami, A., Kopec, C.D., Diamond, M.E., and Brody, C.D. (2018). Posterior parietal cortex
1059 represents sensory history and mediates its effects on behaviour. Nature 554, 368.
1060 Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. Neuroimage 38,
1061 95-113.
1062 Baumann, O., Borra, R.J., Bower, J.M., Cullen, K.E., Habas, C., Ivry, R.B., Leggio, M.,
1063 Mattingley, J.B., Molinari, M., and Moulton, E.A. (2015). Consensus paper: the role of
1064 the cerebellum in perceptual processes. The Cerebellum 14, 197-220.
1065 Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical
1066 and powerful approach to multiple testing. Journal of the royal statistical society
1067 Series B (Methodological), 289-300.
1068 Bromiley, P. (2003). Products and convolutions of Gaussian probability density functions.
1069 Tina-Vision Memo 3, 1.
1070 Burr, D., and Cicchini, G.M. (2014). Vision: efficient adaptive coding. Current Biology 24,
1071 R1096-R1098.
1072 Carlson, J.M., Foti, D., Mujica-Parodi, L.R., Harmon-Jones, E., and Hajcak, G. (2011). Ventral
1073 striatal and medial prefrontal BOLD activation is correlated with reward-related
1074 electrocortical activity: a combined ERP and fMRI study. Neuroimage 57, 1608-1616.
1075 Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., and Cohen, J.D. (1998).
1076 Anterior cingulate cortex, error detection, and the online monitoring of performance.
1077 Science 280, 747-749.
1078 Cavanagh, J.F., and Frank, M.J. (2014). Frontal theta as a mechanism for cognitive
1079 control. Trends in cognitive sciences 18, 414-421.

1080 Choe, K.W., Blake, R., and Lee, S.-H. (2014). Dissociation between neural signatures of
1081 stimulus and choice in population activity of human V1 during perceptual decision-
1082 making. Journal of Neuroscience 34, 2725-2743.
1083 Choe, K.W., Blake, R., and Lee, S.-H. (2016). Pupil size dynamics during fixation impact the
1084 accuracy and precision of video-based gaze estimation. Vision research 118, 48-59.
1085 Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic
1086 resonance neuroimages. Computers and Biomedical research 29, 162-173.
1087 Curtis, C.E., and Lee, D. (2010). Beyond working memory: the role of persistent activity in
1088 decision making. Trends in cognitive sciences 14, 216-222.
1089 Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific
1090 and statistical inference (Macmillan International Higher Education).
1091 Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in
1092 psychology 5, 781.
1093 Dyjas, O., Bausenhart, K.M., and Ulrich, R. (2012). Trial-by-trial updating of an internal
1094 reference in discrimination tasks: Evidence from effects of stimulus order and trial
1095 sequence. Attention, Perception, & Psychophysics 74, 1819-1841.
1096 Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a
1097 statistically optimal fashion. Nature 415, 429.
1098 Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., and Turner, R. (1996). Movement-
1099 related effects in fMRI time-series. Magnetic resonance in medicine 35, 346-355.
1100 Fritsche, M., Mostert, P., and de Lange, F.P. (2017). Opposite effects of recent history on
1101 perception and decision. Current Biology 27, 590-595.

1102 Gao, J.-H., Parsons, L.M., Bower, J.M., Xiong, J., Li, J., and Fox, P.T. (1996). Cerebellum
1103 implicated in sensory acquisition and discrimination rather than motor control.
1104 Science 272, 545-547.
1105 Glaze, C.M., Kable, J.W., and Gold, J.I. (2015). Normative evidence accumulation in
1106 unpredictable environments. Elife 4, e08825.
1107 Goard, M.J., Pho, G.N., Woodson, J., and Sur, M. (2016). Distinct roles of visual, parietal,
1108 and frontal motor cortices in memory-guided sensorimotor decisions. Elife 5, e13764.
1109 Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annual review
1110 of neuroscience 30.
1111 Good, I.J. (1979). Studies in the history of probability and statistics. XXXVII AM Turing's
1112 statistical work in World War II. Biometrika, 393-396.
1113 Gorgoraptis, N., Catalao, R.F., Bays, P.M., and Husain, M. (2011). Dynamic updating of
1114 working memory resources for visual objects. Journal of Neuroscience 31, 8502-8511.
1115 Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics (Oxford,
1116 England: John Wiley).
1117 Griffiths, T.L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their
1118 beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012).
1119 Grinband, J., Hirsch, J., and Ferrera, V.P. (2006). A neural representation of
1120 categorization uncertainty in the human brain. Neuron 49, 757-763.
1121 Haller, M., Case, J., Crone, N.E., Chang, E.F., King-Stephens, D., Laxer, K.D., Weber, P.B.,
1122 Parvizi, J., Knight, R.T., and Shestyuk, A.Y. (2018). Persistent neuronal activity in
1123 human prefrontal cortex links perception and action. Nature Human Behaviour 2, 80.

1124 Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015).
1125 Distinct relationships of parietal and prefrontal cortices to evidence accumulation.
1126 Nature 520, 220.
1127 Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls,
1128 and perspectives. Neuron 87, 257-270.
1129 Hempel, C.G. (1965). Aspects of scientific explanation.
1130 Holroyd, C.B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R.B., Coles, M.G., and
1131 Cohen, J.D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal
1132 and external error signals. Nature neuroscience 7, 497.
1133 Hwang, E.J., Dahlen, J.E., Mukundan, M., and Komiyama, T. (2017). History-based action
1134 selection bias in posterior parietal cortex. Nature communications 8, 1242.
1135 Jacob, E.K. (2004). Classification and categorization: a difference that makes a
1136 difference.
1137 Jasper, A.I., Tanabe, S., and Kohn, A. (2019). Predicting perceptual decisions using visual
1138 cortical population responses and choice history. Journal of Neuroscience, 0035-0019.
1139 Jaynes, E.T. (2003). Probability theory: The logic of science (Cambridge university press).
1140 Jeffreys, H. (1961). Theory of probability, Clarendon. (Oxford).
1141 Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for
1142 the robust and accurate linear registration and motion correction of brain images.
1143 Neuroimage 17, 825-841.
1144 Kahnt, T., Grueschow, M., Speck, O., and Haynes, J.-D. (2011a). Perceptual learning and
1145 decision-making in human medial frontal cortex. Neuron 70, 549-559.
1146 Kahnt, T., Heinzle, J., Park, S.Q., and Haynes, J.-D. (2011b). Decoding different roles for
1147 vmPFC and dlPFC in multi-attribute decision making. Neuroimage 56, 709-715.

1148 Karnath, H.-O. (2001). New insights into the functions of the superior temporal cortex.
1149 Nature Reviews Neuroscience 2, 568.
1150 Karnath, H.-O., Ferber, S., and Himmelbach, M. (2001). Spatial awareness is a function of
1151 the temporal not the posterior parietal lobe. Nature 411, 950.
1152 Karnath, H.-O., Milner, A.D., and Vallar, G. (2002). The cognitive and neural bases of
1153 spatial neglect (Oxford University Press Oxford:).
1154 Kass, R.E., and Raftery, A.E. (1995). Bayes factors. Journal of the american statistical
1155 association 90, 773-795.
1156 Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian
1157 inference. Annu Rev Psychol 55, 271-304.
1158 Kording, K.P., Blohm, G., Schrater, P., and Kay, K. (2020). Appreciating the variety of
1159 goals in computational neuroscience. arXiv preprint arXiv:200203211.
1160 Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning.
1161 Nature 427, 244.
1162 Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional
1163 brain mapping. Proceedings of the National Academy of Sciences 103, 3863-3868.
1164 Lages, M., and Treisman, M. (1998). Spatial frequency discrimination: visual long-term
1165 memory or criterion setting? Vision research 38, 557-572.
1166 Lapid, E., Ulrich, R., and Rammsayer, T. (2008). On estimating the difference limen in
1167 duration discrimination tasks: A comparison of the 2AFC and the reminder task.
1168 Perception & psychophysics 70, 291-305.
1169 Lee, H.-J., Rhim, I., and Lee, S.-H. (2016). Optimal and suboptimal integration of sensory
1170 and value information in perceptual decision-making. CoSyNe.

1171 Lee, S.-H., Blake, R., and Heeger, D.J. (2007). Hierarchy of cortical responses underlying
1172 binocular rivalry. Nature neuroscience 10, 1048.
1173 Ma, W.J. (2012). Organizing probabilistic models of perception. Trends in cognitive
1174 sciences 16, 511-518.
1175 Maddox, W.T. (2002). Toward a unified theory of decision criterion learning in perceptual
1176 categorization. Journal of the experimental analysis of behavior 78, 567-595.
1177 Marco-Pallarés, J., Müller, S.V., and Münte, T.F. (2007). Learning by doing: an fMRI study
1178 of feedback-related brain activations. Neuroreport 18, 1423-1426.
1179 Marcus, D., Harwell, J., Olsen, T., Hodge, M., Glasser, M., Prior, F., Jenkinson, M.,
1180 Laumann, T., Curtiss, S., and Van Essen, D. (2011). Informatics and data mining tools
1181 and strategies for the human connectome project. Frontiers in neuroinformatics 5, 4.
1182 Marzban, C. (2004). The ROC curve and the area under it as performance measures.
1183 Weather and Forecasting 19, 1106-1114.
1184 Nakashima, Y., and Sugita, Y. (2017). The reference frame of the tilt aftereffect measured
1185 by differential Pavlovian conditioning. Scientific reports 7, 40525.
1186 Nestares, O., and Heeger, D.J. (2000). Robust multiresolution alignment of MRI brain
1187 volumes. Magnetic Resonance in Medicine: An Official Journal of the International
1188 Society for Magnetic Resonance in Medicine 43, 705-715.
1189 Nienborg, H., and Cumming, B.G. (2009). Decision-related activity in sensory neurons
1190 reflects more than a neuron's causal effect. Nature 459, 89.
1191 Norton, E.H., Fleming, S.M., Daw, N.D., and Landy, M.S. (2017). Suboptimal criterion
1192 learning in static and dynamic environments. PLoS computational biology 13,
1193 e1005304.

1194 Palmer, J., Huk, A.C., and Shadlen, M.N. (2005). The effect of stimulus strength on the
1195 speed and accuracy of a perceptual decision. Journal of vision 5, 1-1.
1196 Pavan, A., Marotti, R.B., and Campana, G. (2012). The temporal course of recovery from
1197 brief (sub-second) adaptations to spatial contrast. Vision research 62, 116-124.
1198 Qamar, A.T., Cotton, R.J., George, R.G., Beck, J.M., Prezhdo, E., Laudano, A., Tolias, A.S.,
1199 and Ma, W.J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries
1200 in visual categorization. Proceedings of the National Academy of Sciences 110, 20332-
1201 20337.
1202 Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for
1203 two-choice decision tasks. Neural computation 20, 873-922.
1204 Raviv, O., Ahissar, M., and Loewenstein, Y. (2012). How recent history affects perception:
1205 the normative approach and its heuristic approximation. PLoS computational biology
1206 8, e1002731.
1207 Rips, L.J., and Turnbull, W. (1980). How big is big? Relative and absolute properties in
1208 memory. Cognition 8, 145-174.
1209 Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal
1210 area during a combined visual discrimination reaction time task. Journal of
1211 neuroscience 22, 9475-9489.
1212 Sanders, J.I., Hangya, B., and Kepecs, A. (2016). Signatures of a statistical computation in
1213 the human sense of confidence. Neuron 90, 499-506.
1214 Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. Nature
1215 Reviews Neuroscience 8, 522.
1216 Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. arXiv
1217 preprint arXiv:09083817.

1218 Shenhav, A., Straccia, M.A., Cohen, J.D., and Botvinick, M.M. (2014). Anterior cingulate
1219 engagement in a foraging context reflects choice difficulty, not foraging value. Nature
1220 neuroscience 17, 1249.
1221 Sheth, S.A., Mian, M.K., Patel, S.R., Asaad, W.F., Williams, Z.M., Dougherty, D.D., Bush, G.,
1222 and Eskandar, E.N. (2012). Human dorsal anterior cingulate cortex neurons mediate
1223 ongoing behavioural adaptation. Nature 488, 218.
1224 Simon, H.A. (1955). A behavioral model of rational choice. The quarterly journal of
1225 economics 69, 99-118.
1226 Stocker, A.A., and Simoncelli, E.P. (2006). Sensory adaptation within a Bayesian
1227 framework for perception. In Advances in neural information processing systems, pp.
1228 1289-1296.
1229 Treisman, M., and Williams, T.C. (1984). A theory of criterion setting with an application
1230 to sequential dependencies. Psychological Review 91, 68.
1231 Urai, A.E., Braun, A., and Donner, T.H. (2017). Pupil-linked arousal is driven by decision
1232 uncertainty and alters serial choice bias. Nature communications 8, 14637.
1233 White, C.N., Mumford, J.A., and Poldrack, R.A. (2012). Perceptual criteria in the human
1234 brain. Journal of Neuroscience 32, 16716-16724.
1235 Wod, I. (1985). Weight of evidence: A brief survey. Bayesian statistics, 249-270.
1236 Zokaei, N., Burnett Heyes, S., Gorgoraptis, N., Budhdeo, S., and Husain, M. (2015).
1237 Working memory recall precision is a more sensitive index than span. Journal of
1238 Neuropsychology 9, 319-329.

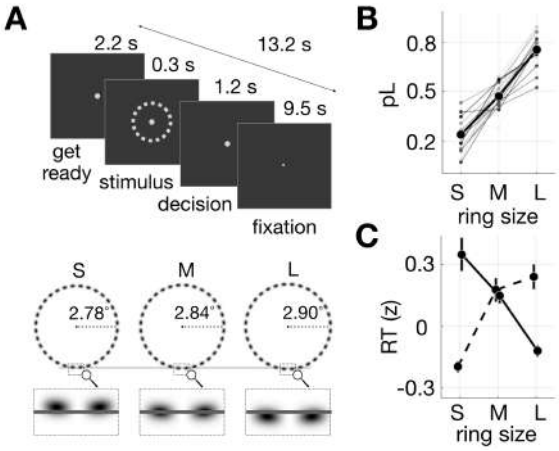

Figure 1. Task and behavioral measures. ***A,*** Ring-size classification task. On each trial, observers briefly viewed a ring stimulus and classified it as 'small' or 'large' within 1.2 s after stimulus onset while fixating on a spot on the display (top). Three ring stimuli that were slightly different in size (bottom) were presented in a random order. The luminance polarity of the ring images at the bottom is reversed here for illustrative purposes. ***B,*** Proportion of 'large' decisions (pL) as a function of ring size. Thin and bold markers, pLs for the individual observers and across-observer average, respectively. ***C,*** Z-scored RT plotted against ring size. Dashed and solid lines represent data from trials on which 'small' and 'large' decisions were made, respectively. *B*, *C*, Circles and error bars indicate the mean and standard error of the mean across observers, respectively. The stimuli, procedure, and data from the main data set (see Methods and Table 1) are shown here.

Data set #	DS#1 (Main data)	DS#2	DS#3	DS#4
Presented previously	No	Yes	No	Yes
Source	-	Choe et al., 2014; paper; Choe et al., 2016; paper	Choe et al., 2014; paper; Choe et al., 2016; paper	Lee et al., 2016; Conference poster
Feedback interval	Run-to-run	Run-to-run	Trial-to-trial	Trial-to-trial
Feedback manipulation	Deterministic	Deterministic	Deterministic	Stochastic
Inter-trial interval (s)	13.2	13.2	2	2.5
The number of stimuli	3	3	16	5
Participants number	18	23	23	30
With fMRI imaging	Yes	No	No	No
Training sessions	Yes	No	No	No

Table 1. Specifications of the data sets for their origins and experimental procedures.

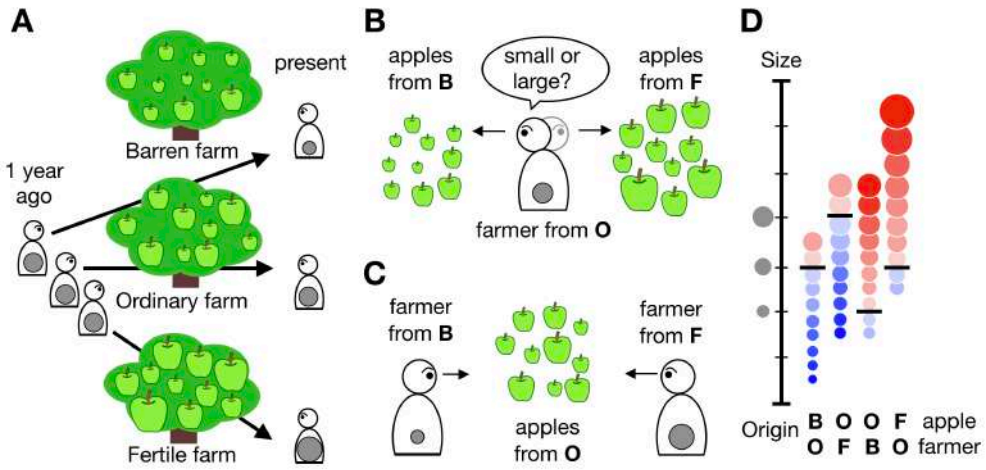

Figure 2. The formation, use, and consequences of a criterion in an apple-classification scenario. ***A***, Experience-dependent formation of criteria. In the first year, three farmers all worked at Ordinary farm and classified the apples there into 'large' or 'small' sizes, forming a criterion that is close to the typical size of the apples of Ordinary farm (gray circles inside the bodies in the left side). In the second year, two of the farmers moved to Barren and Fertile farms respectively. As the farms differ in fertility and thus as their apples differ in size, the three farmers experienced apples of different sizes and form different criteria (gray circles inside the bodies in the right side). ***B,*** Example cases where the apples from different farms are sorted by the same farmer. The farmer from the Ordinary farm classifies the apples from the Barren and Fertile farms as mostly 'small' and 'large', respectively, by comparing the apple sizes against the medium apple size of Ordinary farm, the criterion. ***C,*** Example cases where the apples from the same farm are sorted by different farmers. Due to the different classification criteria, apples from the Ordinary farm are mostly judged as 'large' by the farmer from the Barren farm and 'small' by the farmer from the Fertile farm. ***D,*** Distributions of decisions and uncertainties for the four cases shown in *B* and *C*. The black horizontal bars represent a criterion, and the size of circles represents apple size. The hue and saturation of circles represent decision identity (blue for 'small', red for 'large') and uncertainty (increasing desaturation with increasing uncertainty), respectively.

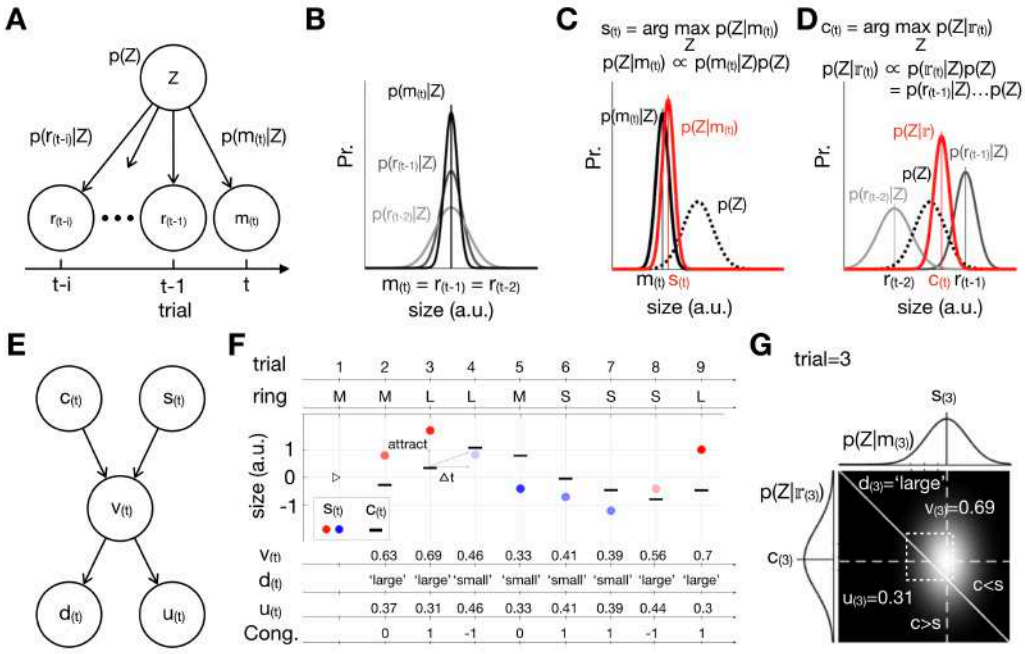

Figure 3. Descriptions of NBMC. *A,* Generative model. Arrows indicate conditional dependencies. Z, true stimulus size. $m_{(t)}$, noisy observation on a current trial t. $r_{(t-i)}$, noisy observation retrieved from a past trial $t-i$. ***B,*** Increases in measurement noise over trials. A likelihood function over ring-size encodes the most likely value of $Z_{(t)}$ and its uncertainty, which increases as trials elapse. ***C,*** Bayesian inference for a current stimulus, s. A posterior function (red) represents $s_{(t)}$ and its uncertainty. ***D,*** Bayesian inference of a criterion, c. A posterior function (red) represents $c_{(t)}$ and its uncertainty. ***E,*** Causal structure of variables. $s_{(t)}$ and $c_{(t)}$ construct $v_{(t)}$, a continuous decision variable. Then, $v_{(t)}$ determines $d_{(t)}$, binary decision, and $u_{(t)}$, decision uncertainty. ***F,*** Illustration of decision episodes over example trials. On the stimuli presented by experimenters, the Bayesian observer makes decisions by comparing s (colored dots) against c (black bars), which is updated continuously over trials. As $s_{(t)}$ and $c_{(t)}$ are inferred, $v_{(t)}$, $d_{(t)}$ and $u_{(t)}$ are deduced accordingly. 'Cong.' indicates a congruence between $d_{(t)}$ and ring size and is defined as $d_{(t)} \times Z_{(t)}$, where $d_{(t)} \in \{-1,1\}$ for 'small' and 'large', respectively, and $Z_{(t)} \in \{-1,0,1\}$ for S, M, L rings, respectively. Note that $c_{(t)}$ appears to chase after $s_{(t-1)}$ as indicated by the small arrows. ***G,*** Illustration of the model variables in a bivariate decision space for trial 3 in the panel *F*. The light

intensity corresponds to the probability density of the bivariate conjugate of $p(Z|m_{(3)})$ and $p(Z|r_{(3)})$. $v_{(3)}$ corresponds to the fraction of the bivariate distribution above the identity line, which is greater than 0.5 and thus leads to the 'large' decision ($d_{(3)} = \text{'large'}$). $u_{(3)}$ corresponds to the fraction of the bivariate distribution that is against the decision, i.e., the probability that the decision was incorrect.

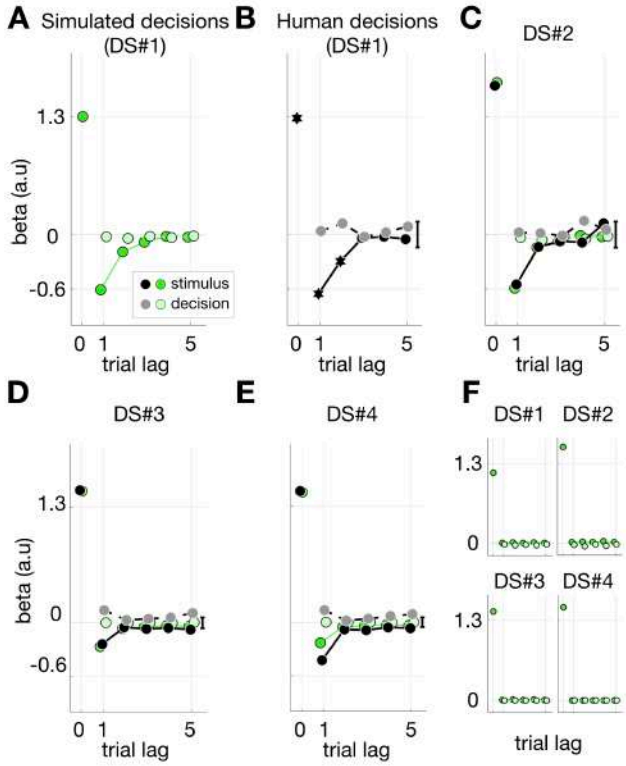

Figure 4. Regression of current decisions onto past stimuli. *A-B,* Multiple logistic regressions of the simulated decisions of the NBMC Bayesian observers (A) or the decisions of the human observers (B) onto stimuli and past decisions for DS#1. The regression coefficients are plotted against trial lags, where zero means a current trial. The green and black symbols represent the coefficients for the Bayesian and human observers, respectively. The dark and light symbols represent the coefficients for stimuli and decisions, respectively. For clarity, only the average of 95% bootstrap confidence intervals of the mean coefficients is shown as vertical black bar for the human data. The hexagons indicate the coefficients that are significantly deviated from zero (pairwise t-test, $P < 0.05$). ***C-E,*** Multiple logistic regressions of the simulated decisions of the NBMC Bayesian observers and the decisions of the human observers onto stimuli and past decisions for DS#2 (C), DS#3 (D), and DS#4 (E), respectively. The color and lightness schemes are identical to those used in A and B. ***F,*** Multiple logistic regressions of the simulated decisions of the constant-c model observers onto stimuli and past decisions for all the data sets. The lightness schemes are identical to those used in A.

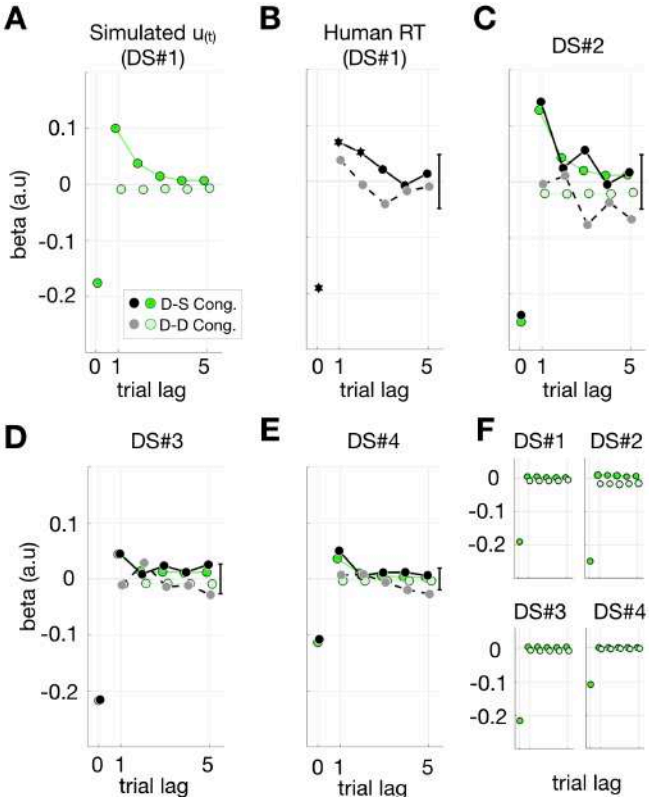

Figure 5. Regression of current decision uncertainty or RT onto the congruence of current decision with past stimuli. *A-B,* Multiple linear regressions of the simulated decision uncertainty measures ($u_{(t)}$) of the NBMC Bayesian observers (*A*) or the decision RTs of the human observers (*B*) onto the congruence of current decisions with stimuli or with past decisions for DS#1. The regression coefficients are plotted against trial lags, where zero means a current trial. The green and black symbols represent the coefficients for the Bayesian and human observers, respectively. The dark and light symbols represent the coefficients for the congruence of current decisions with stimuli and past decisions, respectively. For clarity, only the average of 95% bootstrap confidence intervals of the mean coefficients is shown as vertical black bar for the human data. The hexagons indicate the coefficients that are significantly deviated from zero (Generalized multiple linear mixed effect model (GLMM), $P < 0.05$). ***C-E,*** Multiple linear regressions of the simulated $u_{(t)}$ of the NBMC Bayesian observers and the decision RTs of the human observers onto the congruence of current decisions with stimuli or with past decisions for DS#2 (C), DS#3 (D),

and DS#4 (E), respectively. The color and lightness schemes are identical to those used in A and B.

F, Multiple linear regressions of the simulated $u_{(t)}$ of the constant-c model observers onto the congruence of current decisions with stimuli or with past decisions for all the data sets. The lightness schemes are identical to those used in A.

Cortical area	Correlated time point (s)	Volume (mm^3)	Peak voxel		
			MNI coordinate	GLMM p-value of $u_{(t)}$	GLMM coefficient of $u_{(t)}$
Dorsal anterior cingulate cortex	5.5	3564	[3, 27, 36]	1.43×10^{-8}	0.095
Left anterior insula	5.5	2268	[-27, 24, 6]	4.6×10^{-8}	0.095
Right anterior insula	5.5	2619	[33, 18, 6]	8.3×10^{-7}	0.082

1331

1332 **Table2.** Specifications of the brain clusters in which fMRI responses were correlated with
1333 decision uncertainty. When a set of constraints (uncorrected $P < 10^{-4}$, FDR-corrected $P < 0.05$,
1334 cluster size > 324 mm^3 (or contiguous voxels > 12)) was applied, three regions of interest (ROIs)
1335 were significantly correlated with $u_{(t)}$ at the fourth fMRI time point (5.5s after stimulus onset).
1336 The two-tailed tests were applied for evaluating the significance of GLMM.

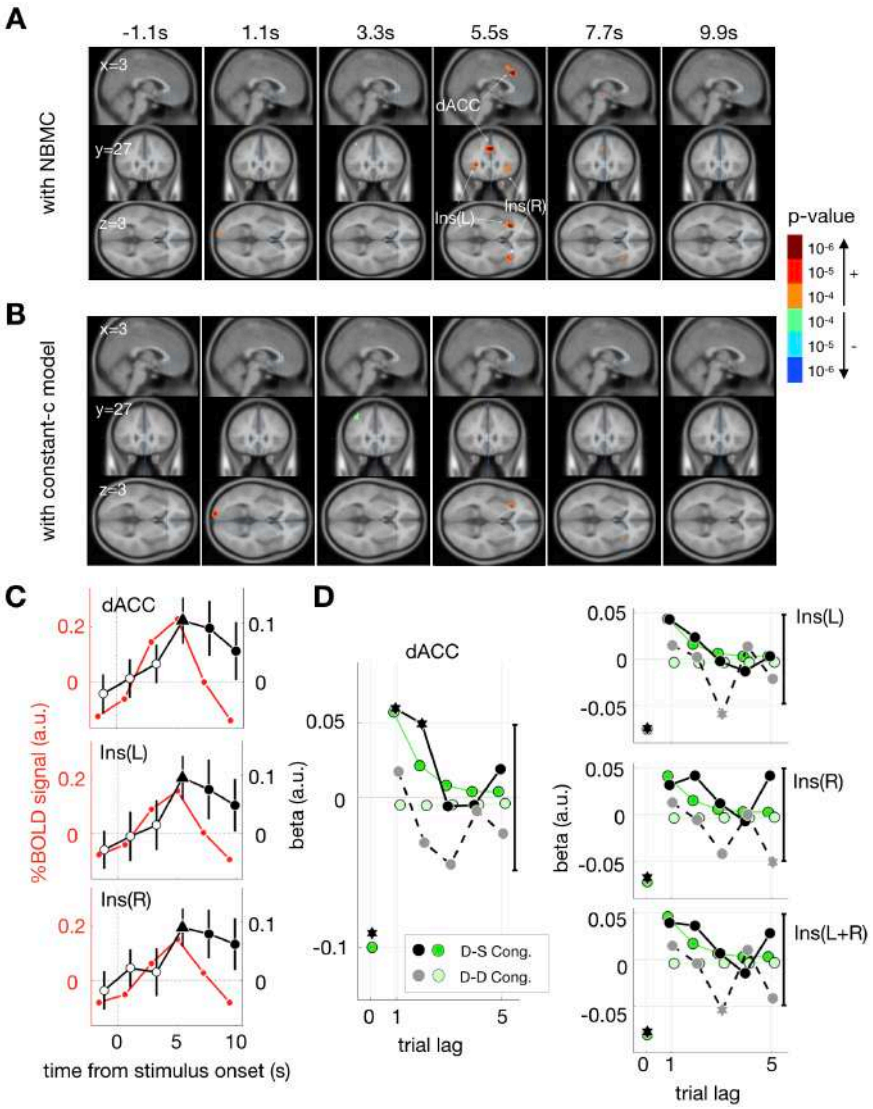

Figure 6. Univariate fMRI responses correlated with decision uncertainty. ***A-B,*** Regressions of fMRI responses at individual voxels onto $u_{(t)}$ for each of the 6 time points that comprise single trials (the numbers at the top indicate the time points from stimulus onset). For each of those time points, the cortical sites where the p-values of regression coefficients were smaller than a threshold (GLMM, $P < 10^{-4}$, uncorrected) were marked in color and overlaid on the sagittal (x=3), coronal (y=27), and axial (z=3) slices of a template brain. ***A,*** Regression results with $u_{(t)}$ of the NBMC. ***B,*** Regression results with $u_{(t)}$ of the constant-c model. ***C,*** Trial-locked time courses of the brain signals of decision uncertainty in the dACC and the insula. The red and black symbols are the fMRI responses averaged within an ROI and the regression coefficients of $u_{(t)}$ that is

estimated by the NBMC, respectively. The significant regression coefficients are marked by the filled circles (GLMM, $P < 0.05$) and the filled triangles (GLMM, $P < 10^{-5}$). Error bars are 95% GLMM CIs of regression coefficients. ***D,*** Results of the multiple linear regressions of the fMRI responses in the dACC and the insula onto the congruencies of current decisions with stimuli and past decisions. The regression coefficients for the brain signals associated with $u_{(t)}$ (black symbols) are juxtaposed with the regression coefficients for the model simulation of $u_{(t)}$ (green symbols). The significant regression coefficients (GLMM, $P < 0.05$) for the brain signals are marked by the hexagons, the 95% CIs of which did not include zero. For clarity, only the averages of 95% CIs for the mean of the regression coefficients for the brain signals are shown as the vertical black error bars.

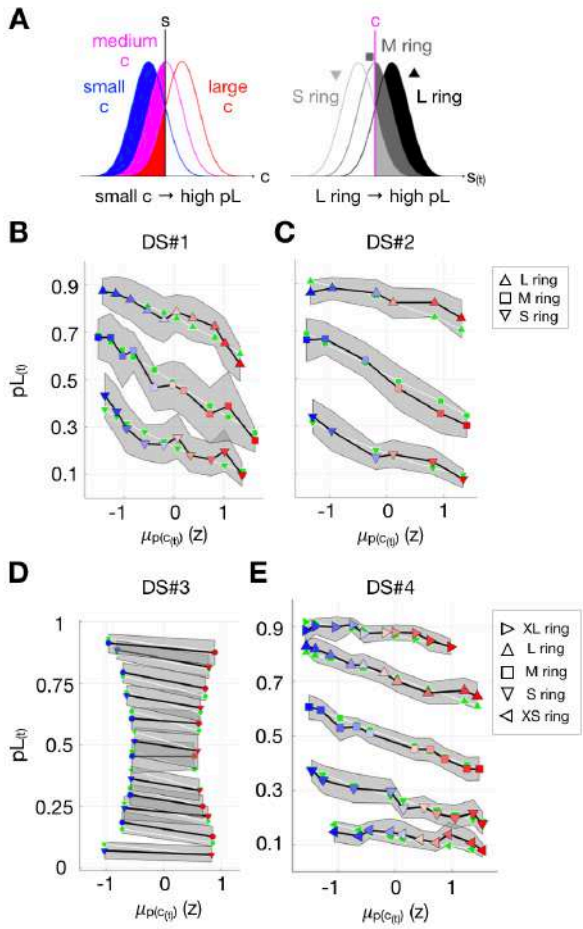

Figure 7. Dissecting stimulus and classification criterion contributions to the variability of decision fraction. ***A,*** A schematic illustration of the respective contributions of c and s to the variability of pL. Left, c varies while s is fixed. Right, s varies while c is fixed. The shaded areas represent the fractions of s being larger than c (i.e., pL). ***B-E,*** Comparisons of pL between the Bayesian and human observers. The average pLs of the Bayesian observers (green symbols) and the humans (blue to red symbols) are plotted against the binned model estimates of c. The data falling on each of the continuous lines are the pLs from the same ring size. Trials were binned by the mean of the inferred criteria $\mu_{p(c_{(t)})}$, as indicated by the symbols' color that varies in a scheme similar to that in the left panel of *A*. The shaded gray areas demarcate the 95% bootstrap confidence intervals of data means.

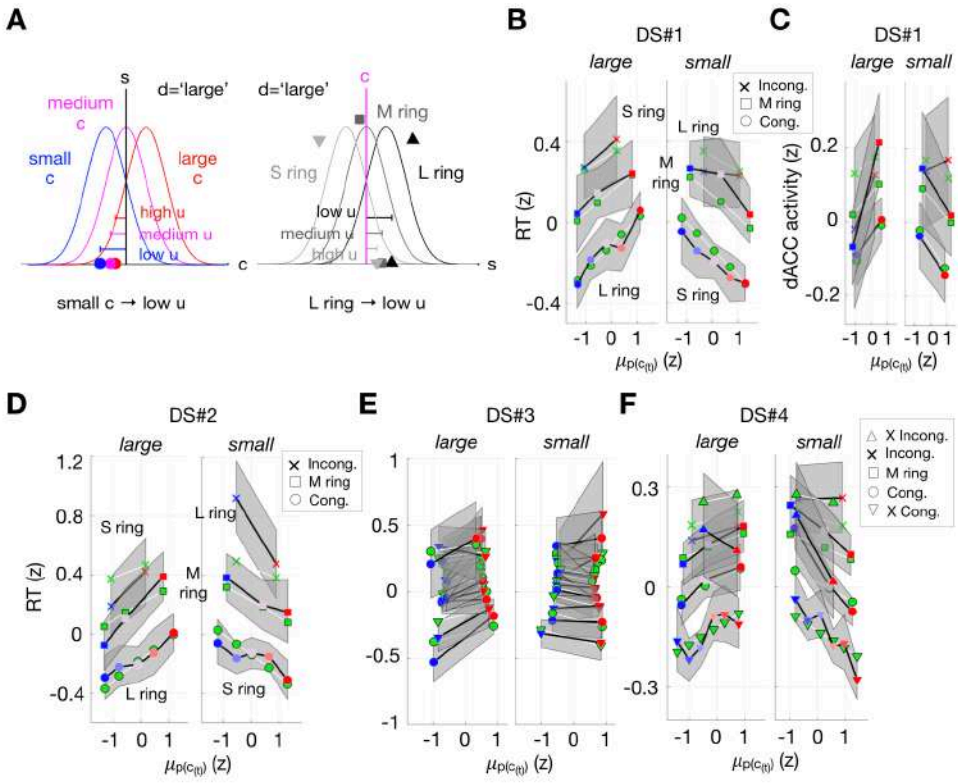

Figure 8. Dissecting stimulus and classification criterion contributions to the variability of decision uncertainty. ***A,*** A schematic illustration of the respective contributions of c and s to the variability of u for 'large' decision. Left, c varies while s is fixed. Right, s varies while c is fixed. The symbols on the horizontal axes demarcate the means of c (left panel) and s (right panel) values that yield the 'large' decision. For a given condition, the distance between c and s, which is indicated by the length of the horizontal bars, determines the degree of decision uncertainty. ***B-E,*** Comparisons of the decision uncertainty $u_{(t)}$ of the Bayesian observers and the RTs (*B*, *D-F*) or the dACC responses (*C*) of the human observers. The average $u_{(t)}$ of the Bayesian observers (green symbols) and the average RTs or dACC responses of the humans (blue to red symbols) are plotted against the binned model estimates of c. The data falling on each of the continuous lines are the data from the same ring size. Trials were binned by the mean of inferred criteria $\mu_{p(c_{(t)})}$, as indicated by the symbols' color that varies in a scheme similar to that in the left panel of *A*. The shaded gray areas demarcate the 95% bootstrap confidence intervals of data means.

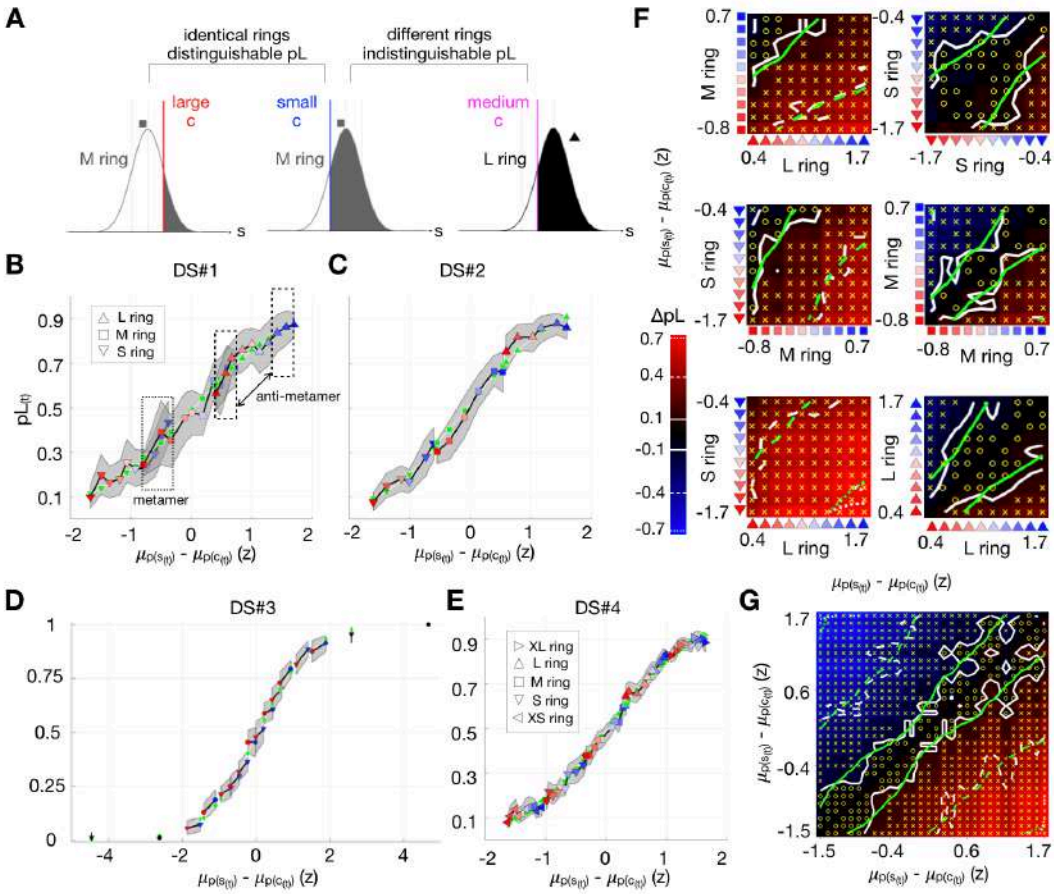

Figure 9. Relativity of classification in decision fraction. ***A,*** A schematic illustration of 'anti-metameric' and 'metameric' trial pairs in decision fraction. The curves represent the distributions of s for given ring sizes. The light gray vertical lines demarcate the means of s for the S, M, and L-size rings. The vertical lines in color represent the locations of c. The shaded areas represent pLs, where s is greater than c. In the 'anti-metameric' pair (left two panels), the pLs differ substantively despite identical stimuli due to a large difference in c. In the 'metameric' pair (right two panels), the pLs do not differ despite different stimuli because the locations of c relative to the distribution of s do not differ. ***B-E,*** Continuous and monotonic increases in $pL_{(t)}$ with increasing differences between s and c over different stimulus conditions. The pLs of the Bayesian observers (green symbols) and the humans (non-green symbols) are plotted against the differences between s and c. The color spectrum from blue to red represents the increasing size of c within each stimulus condition. The symbol shapes represent different stimulus conditions

(but only two different shapes were used alternatingly to distinguish many different ring sizes for DS#3). Shaded areas, 95% bootstrap confidence intervals of the means of data. ***F-G,*** Statistical demonstration of the relativity in DS#1 by using Bayes factor. ***F,*** For all of the possible pairs of the trial bins in DS#1 (shown in *B*), the differences in $\mathrm{pL}_{(t)}$ were calculated and represented by color gradients within 6 separate grid panels. For each grid panel, the columns and rows of trial bins are marked by the same symbols and colors used in *B*. 'x' and 'o' markers indicate $\mathrm{pL}_{(t)}$ of the bin pairs that are significantly distinguishable ($\text{Bayes factor} > 3$) and indistinguishable ($\text{Bayes factor} < 1/3$), respectively. The dotted, dashed and solid lines are the iso-$\Delta\mathrm{pL}$ lines with $\Delta\mathrm{pL} = \pm 0.7, \pm 0.4, \pm 0.1$, respectively. The green and white lines are the iso-$\Delta\mathrm{pL}$ lines defined by the Bayesian simulations and the human data, respectively. ***G,*** A unified grid of $\Delta\mathrm{pL}$ merged across stimuli. The format is identical to that used in *F*.

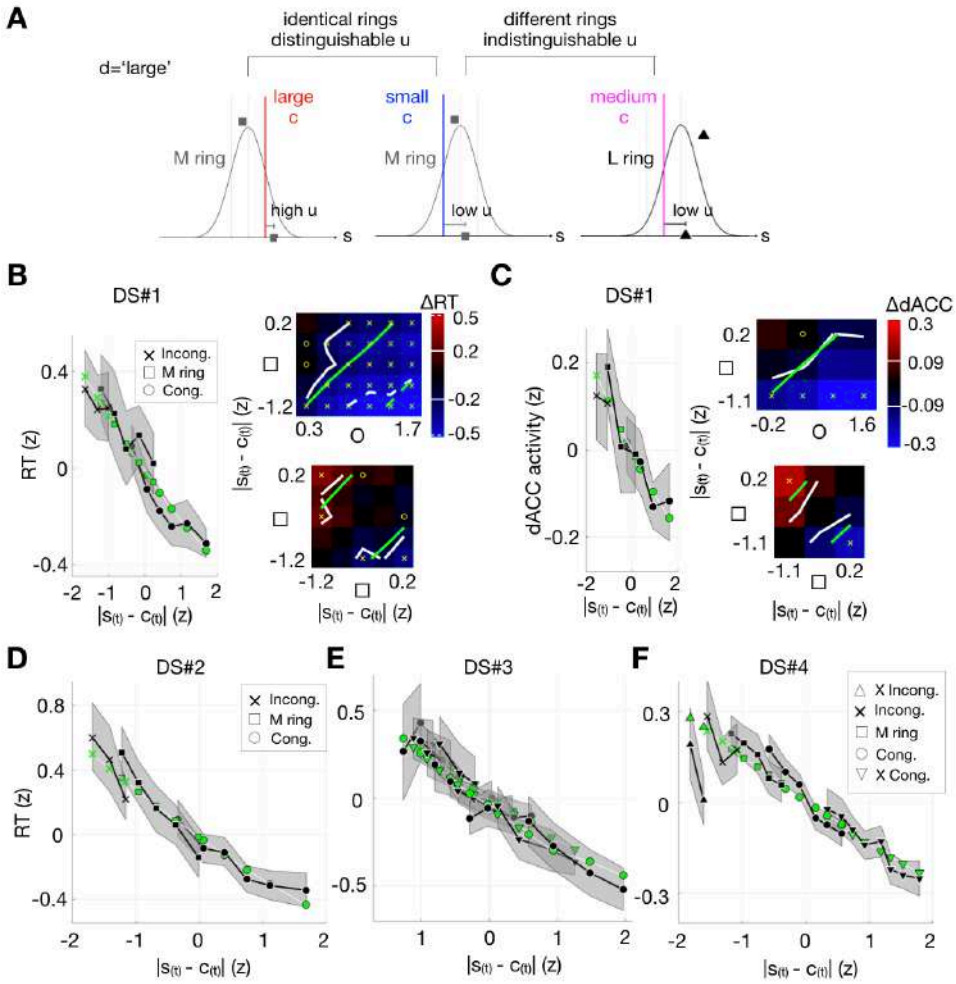

Figure 10. Relativity of classification in decision uncertainty. ***A,*** A schematic illustration of 'anti-metameric' and 'metameric' trial pairs in decision uncertainty. The curves represent the distributions of s for given ring sizes. The light gray vertical lines demarcate the means of s for the S, M, and L-size rings. The vertical lines in color represent the locations of c. The length of the horizontal bars represents the mean degree of decision uncertainty ($u_{(t)}$), when a current decision is 'large'. In the 'anti-metameric' pair (left two panels), the $u_{(t)}$s differ substantively despite identical stimuli due to a large difference in c. In the 'metameric' pair (right two panels), the $u_{(t)}$s do not differ despite different stimuli because the locations of c relative to the distribution of s do not differ. ***B-F,*** Continuous and monotonic increases in $u_{(t)}$, RT, or dACC activity with increasing differences between s and c over different stimulus conditions. The $u_{(t)}$ of

the Bayesian observers (green symbols) and the RTs or dACC responses of the humans (black symbols) are plotted against the differences between s and c. The symbol shapes represent different stimulus conditions (but only two different shapes were used alternatingly to distinguish many different ring sizes for DS#3). Shaded areas, 95% bootstrap confidence intervals of the means of data. ***B-C,*** The right panels show the results of the Bayes factor analysis on the RT (*B*) and dACC (*C*) data for DS#1. For the possible pairs of the trial bins in Cong.-M ring condition (top) or M ring-M ring condition (bottom), the differences in RT (*B*) or in dACC activity (*C*) were calculated and represented by color gradients. For each grid panel, the columns and rows of trial bins are marked by the same symbols and colors used in the left panels. 'x' and 'o' markers indicate RTs or dACC responses of the bin pairs that are significantly distinguishable (Bayes factor > 3) and indistinguishable (Bayes factor $< 1/3$), respectively. The dashed and solid lines in *B* are the iso-$\Delta\mathrm{RT}$ lines with $\Delta\mathrm{RT}= \pm 0.5, \pm 0.2$, respectively. The solid lines in *C* are the iso-$\Delta\mathrm{dACC}$ lines with $\Delta\mathrm{dACC} = \pm 0.09$. The green and white lines are the iso-$\Delta\mathrm{RT}$ or iso-$\Delta\mathrm{dACC}$ lines defined by the Bayesian simulations and the human data, respectively.

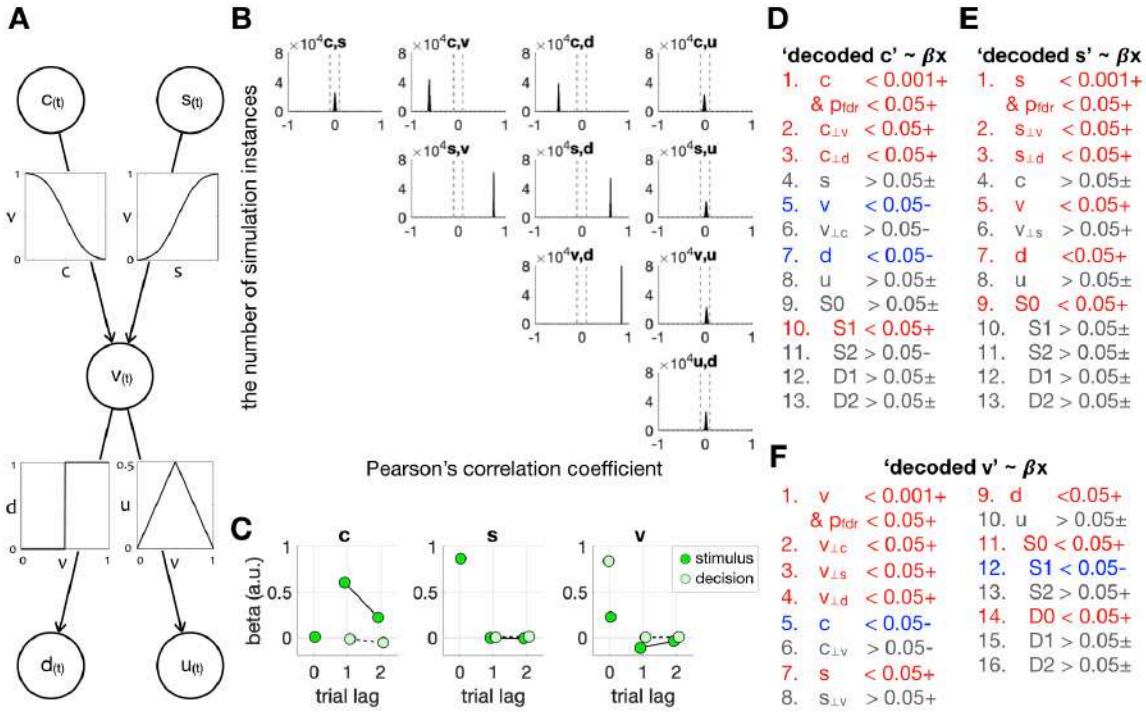

Figure 11. Deduction of regressions for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ from the causal structure of variables in the NBMC. ***A,*** A graphic illustration of the causal structure of variables. Each arrow and an accompanying inset plot represent the direction and function of the causality between two variables, respectively. ***B,*** Histograms of simulated correlation coefficients between the variables. For each possible pair of the variables, the distribution of Pearson's correlation coefficients across many repeated simulations (see Methods for the simulation procedure) was plotted as a histogram. The gray dashed lines demarcate ± 0.1 Pearson's correlation coefficients. ***C,*** Multiple linear regressions of the simulated values of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ onto the stimuli and decisions on current and past trials. ***D-F,*** The lists of regressions that should be satisfied by $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$. The regressions onto the other latent variables (1~8th, 1~8th, 1~10th regressions for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$, respectively) were derived from the simulated correlation analyses depicted in *B* while the regressions onto stimuli and decisions (9-13th, 9~13th, 11~16th regressions for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$, respectively) were derived from the simulated multiple regression analyses depicted in *C*. The meanings of symbols, numbers, and colors used for describing the regressions are as follows: '+',

‘–’, and ‘±’, indicate that the significance test is one-tailed right, one-tailed left, and two-tailed, respectively; numbers with decimal points, threshold p-values for GLMM regression (p_{fdr}, FDR-corrected p-values); ‘<’ and ‘>’, significant and non-significant regressions, respectively; ‘$A_{\perp B}$’, residuals of A from the linear regression onto B; ‘D#’ and ‘S#’, decision and stimulus on a #-back trial; red, blue, and gray are positive, negative, and non-significant regressions, respectively.

Cortical area	Decoded variable	Informative time from stimulus onset (s)	Abbreviation of the decoded variable	Contiguous searchlight centers (N)	Peak searchlight	
					MNI coordinate	GLMM p-value of
Left inferior parietal lobe	$c_{(t)}$	1.1	IPL_{c1}	15	[-54, -27, 48]	8.8×10^{-8}
Left superior temporal gyrus	$c_{(t)}$	3.3	STG_{c3}	13	[-45, -30, 9]	5.7×10^{-7}
Left superior temporal gyrus	$c_{(t)}$	5.5	STG_{c5}	18	[-66, -21, 9]	2.3×10^{-7}
Left dorsolateral prefrontal cortex	$s_{(t)}$	3.3	$DLPFC_{s3}$	33	[-51, 27, 24]	1.9×10^{-7}
Right cerebellum	$s_{(t)}$	5.5	$Cereb_{s5}$	36	[36, -63, -21]	1.4×10^{-6}
Left superior temporal gyrus	$v_{(t)}$	5.5	STG_{v5}	15	[-60, -9, 12]	4.9×10^{-8}

Table 3. Specifications of the regions in which the brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ were identified.

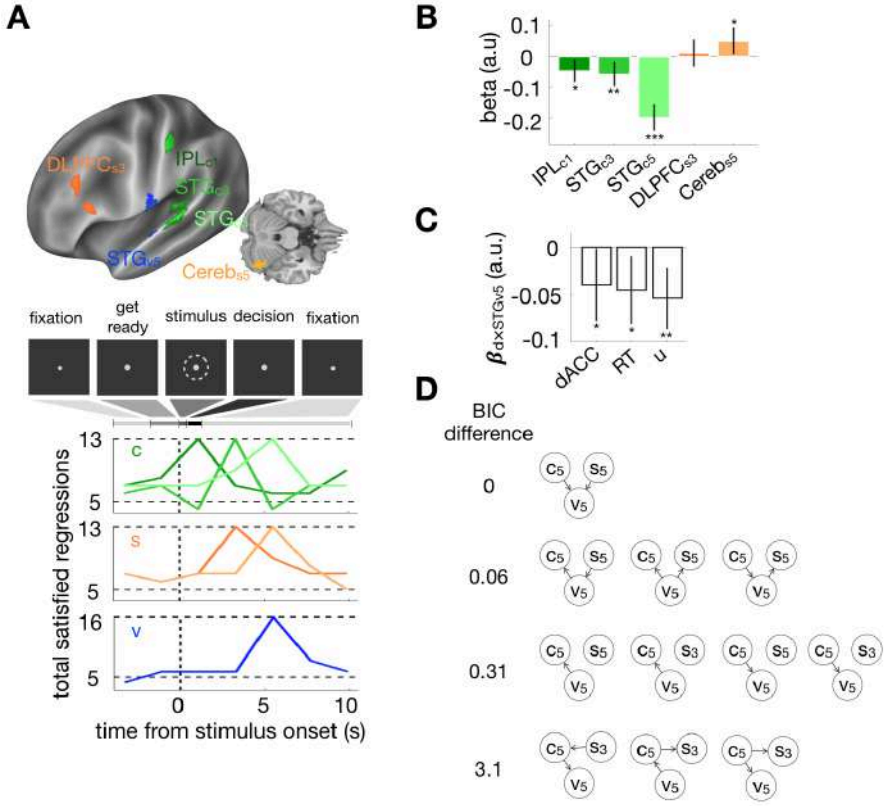

Figure 12. The brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$. ***A,*** Top, the six loci of the brain signals that satisfied all the regressions required by the causal structure of variables in the NBMC are overlaid on the template brain. See Table 3 for the detailed specifications for these loci. Bottom, the numbers of the satisfied regressions are plotted as a function of time relative to stimulus onset for the c, s, and v signals. ***B,*** Coefficients of multiple regressions of the brain signal of v onto the brain signals of c and s. ***C,*** Coefficients of the 'decision x brain signal of v' term in the regression models, in each of which either the brain signal of u ($dACC_{u5}$), the RT data, or the model estimates of u, was regressed onto the brain signal of v (STG_{v5}) and decisions. ***B-C,*** *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Error bars, 95% GLMM confidence intervals of the mean coefficients. ***D,*** The causal networks sorted by their BIC scores relative to the network that best explains the concurrent variabilities between the brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$.

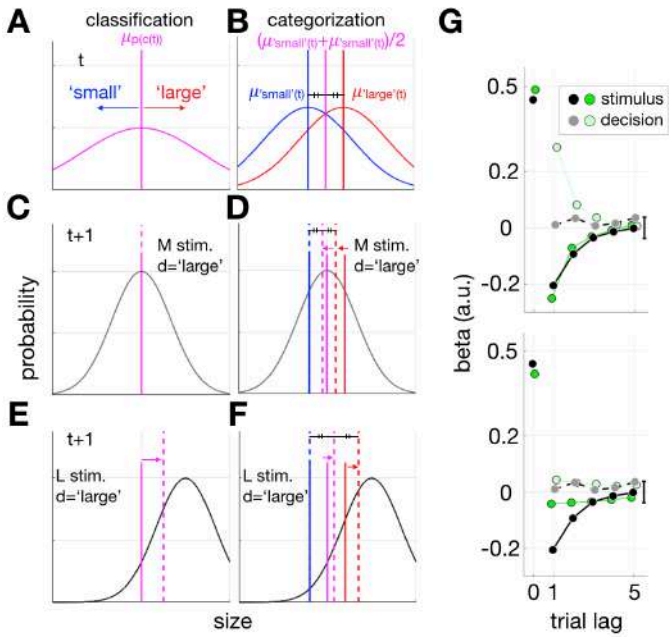

Figure 13. History effects predicted by the categorization model. ***A-F,*** Schematic illustrations of how the criterion and decision boundary are updated in the classification (*A*, *C*, *E*) and categorization (*B*, *D*, *F*) models, respectively. ***A-B,*** Distributions of stimuli assumed in the classification (*A*) and categorization (*B*) models. In the classification model, a stimulus is classified into 'small' or 'large' according to their size relative to the criterion (dashed line) within a single distribution of stimuli (purple curve). In the categorization model, a stimulus is categorized into 'small' or 'large' according to which 'category' distributions of stimuli (blue and red curves for 'small' and 'large' categories, respectively) it belongs to. Thus, when the two distributions are equal in shape, the effective decision boundary (dashed purple line) is prescribed as the arithmetic mean of the means of the two category distributions (dashed blue and red lines). ***C-D,*** Effects of past decisions on classification criterion (*C*) and categorization boundary (*D*). Gray curves represent distributions of noisy sensory measurements associated with a medium-size stimulus. When a 'large' decision was made on a medium-size stimulus in a past trial, the classification criterion (solid purple line in *C*) remains the same whereas the categorization boundary (solid purple line in *D*) is shifted toward the small side as the 'large' category distribution (solid red line in *D*) is shifted toward the small side. ***E-F,*** Effects of past stimuli on

classification criterion (*E*) and categorization boundary (*F*). Gray curves represent distributions of noisy sensory measurements associated with a large-size stimulus. When a large-size stimulus was presented in a past trial, the classification criterion (solid purple line in *E*) and the categorization boundary (solid purple line in *F*) are both shifted toward the small side.

G, Regressions of the simulated decisions of an observer with the categorization-model onto stimuli and past decisions for DS#1. The regression coefficients for the simulated data (green symbols) are juxtaposed with those for the human data (black and gray symbols). The top and bottom panels show the results associated with the two different set of model parameters, the former being best fit to stimulus history effects and the latter to decision history effects.

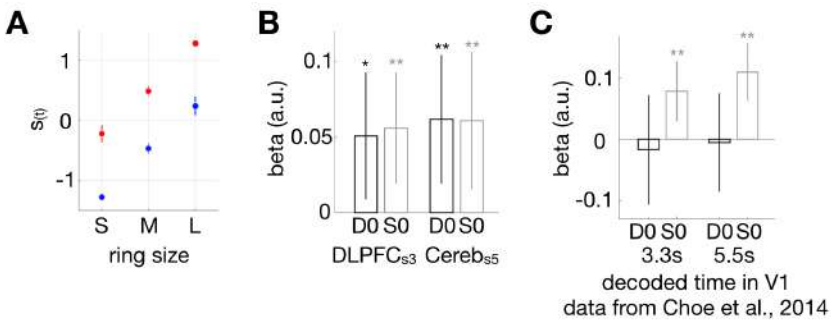

Figure 14. Distinction between brain signal of s and physical ring size. ***A,*** Presence of decision information in the latent variable s of the NBMC. The mean values of s are plotted against the physical ring size, separately for 'small (blue)' an 'large (red)' decisions. The error bars are 95% bootstrap CIs. ***B,*** GLMM of the brain signal of s in the DLPFC and cerebellum onto current decisions (D0) and current stimuli (S0). ***C,*** GLMM of the brain signal of s in the primary visual cortex onto D0 and S0. The error bars are 95% GLMM CIs, and asterisks indicate p values: *< 0.05; **< 0.01.