A new formalism for perceptual classification with normative inference of internal criteria

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

For correspondence: visionsl@snu.ac.kr

Abstract Perceptual classification, one canonical form of decision-making, entails assigning stimuli to discrete classes according to internal criteria. Accordingly, the standard formalisms of perceptual decision-making have incorporated both *stimulus* and *criterion* as necessary components, but granted them unequal representational status, *stimulus* a random variable and *criterion* a scalar variable. This representational inconsistency obscures identifying the origins of behavioral or neural variability in perceptual classification. Here, we redress this problem by presenting an alternative formalism in which *criterion*, as a latent random variable, plays causal roles in forming *decision variable* on equal footings with *stimulus*. By implementing this formalism into a Bayes-optimal algorithm, we could predict, simulate, and explain the key human classification behaviors with high fidelity and coherency. Further, by acquiring concurrent fMRI measurements from humans engaged in classification, we demonstrated an ensemble of brain activities that embodies the causal interactions between *stimulus*, *criterion*, and *decision variable* as the algorithm prescribes.

Introduction

Classification, the act of assigning objects or events to discrete classes according to a criterion, is a necessary precursor to the emergence of rigorous scientific constructs (e.g., taxonomies in physics, biology, psychiatry; (*Ghiselin, 1981; Hempel, 1965*). Classification is required to generate and understand basic linguistic propositions such as predication (e.g., 'small/large', 'near/far', 'dark/bright'; *Rips and Turnbull, 1980*). Mirroring these roles in science and language, classification is considered among the most fundamental of all decision-makings (*Ashby, 2001; Ashby and Ell, 2001*) and exercised countlessly in daily life: a weatherperson forecasts whether the upcoming summer will be cool or hot; a sommelier tells us that the wine of our choice is dry or sweet.

Perceptual classification[1], the most basic form of classification, requires comparing a sensed quantity of a stimulus feature (e.g., 'sensed sweetness of a particular wine') against a criterion quantity learnt prior to comparison (e.g., 'a typical sweetness of wines'). This means that a 'classifying' brain needs to form two representations, one for stimulus (s) and the other for criterion (c). Accordingly, these two representations are incorporated into standard formalisms for perceptual decision-making, such as the signal detection theory (*Green and Swets, 1966*).

These standard formalisms and their modern extensions, despite their remarkable successes in guiding behavioral and neural studies on perception (*Gold and Shadlen, 2007*), all share a fundamental inconsistency in formalizing s and c: s is a random variable that causes sensory measurements via a stochastic process whereas c is a scalar variable that is determined on a rather arbitrary basis. Consequentially, while rigorous normative algorithms have been developed for inferring s as a latent variable (*Körding and Wolpert, 2004; Petzschner et al., 2015; Pouget et al., 2013*), c has mostly been treated simply as a constant (*Fründ et al., 2014; Gold and Shadlen, 2007; Green and Swets, 1966; Meyniel et al., 2015; Renart and Machens, 2014; White et al., 2012*). Even when c is assumed to vary by a few descriptive models (*Benjamin et al., 2009; Fründ et al., 2014; Kepecs et al., 2008; Rahnev and Denison, 2018; Treisman and Williams, 1984*), it has never been treated as a latent random variable that can be inferred on a normative basis.

We argue that this 'representational inconsistency' between s and c incurs grave repercussions for the effort of identifying the sources of the neural and behavioral variabilities during perceptual classification, which radically thwarts the understanding of underlying cognitive and neural processes (*Renart and Machens, 2014*). Specifically, with the algorithms in which c is fixed to a constant, any variability in behavior or neural activity must be either attributed to s or left unexplained, but never attributed to c. On the other hand, with those in which c is allowed to vary but is determined rather arbitrarily, any observed behavior or neural variabilities are bound to be over-attributed to c owing to the unprincipled nature of c estimation.

To redress this representational inconsistency, the current work presents an alternative formalism in which c is also defined as a latent random variable, such that c can be inferred on a normative basis. In this new formalism, the inferred s and the inferred c jointly, on computationally equal footings with each other, cause the decision variable (v) for classification. We will show that the new formalism not only shields classification algorithms from the aforementioned repercussions but also guides us to a normative and parsimonious account for three core properties of classification, which are readily intuited from our daily experiences but have never been quantified by previous studies.

To help intuit those core properties of classification, consider the scenario depicted in Figure 1; three apple farms differ in fertility ('B(arren)', 'O(rdinary)', and 'F(ertile)'), and, on each farm, farmers have been picking apples and sorting them into 'small' and 'large' groups. The overall apple size tends to be small, medium, and large in the B, O, and F farms, respectively, and the farmers' criteria for 'small' versus 'large' also increase in that order, which reflects the overall differences in the apple size encountered by the farmers so far (*Figure 1A*). Then, one day, the farmers visit one another's farms and sort the apples there.

[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 Discussion.

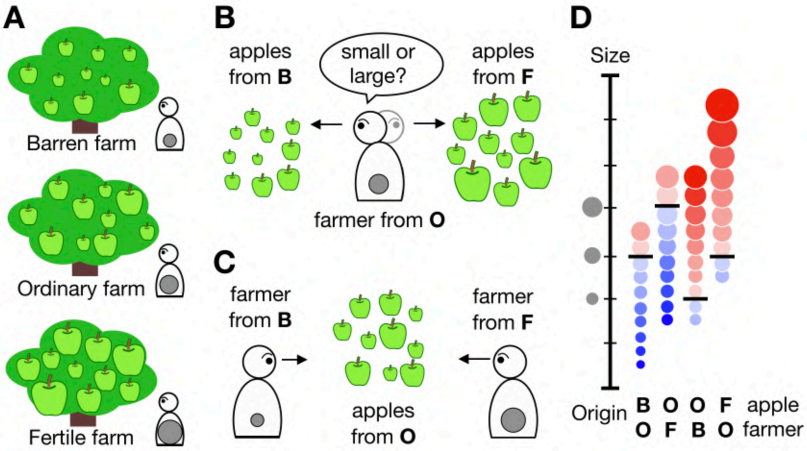


Figure 1. The formation, use, and consequences of criterion in an apple-sorting task (**A**) Farmers working at three apple farms that differ in fertility. As the farms differ in fertility, the three farmers experience apples of different sizes and form different criteria based on the typical apple size experienced (denoted by gray circles). (**B**) Example cases where the apples from different farms are sorted by the same farmer. The farmer from the Ordinary farm judges the apples from the Barren and Fertile farms as mostly 'small' and 'large', respectively, by comparing the apple sizes against her "medium-size" criterion. (**C**) Example cases where the apples from the same farm are sorted by different farmers. Due to the different 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 indicate the criteria, and the size of circles represents the size of apples. The hue and saturation of circles represent decision identity (blue for 'small', red for 'large') and uncertainty (increasing desaturation with increasing uncertainty), respectively.

First, imagine that the farmer from O farm sorts the apples from B farm ('B by O') or the apples from F farm ('F by O'; *Figure 1B*). Farmer O is likely 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. The farmer's decision uncertainty also varies from apple to apple, becoming increasingly uncertain as the apple size falls closer to her criterion. This exemplifies the contribution of s to v, decision variable, when c is fixed (*Figure 1D*). Next, imagine that apples from O farm are sorted by the farmer from F farm ('O by F') or by the farmer from B farm ('O by B'; *Figure 1C*). This exemplifies the contribution of c to v: the very same apples will be sorted differently, and with different degrees of uncertainty, depending on who sorts them (*Figure 1D*). Intriguingly, different pairings bespeak the relativity of classification in which decision and its uncertainty are determined not by an absolute quantity of s but by the relative quantity of s to c. For example, the 'B by O' and 'O by F' cases differ both in absolute apple size and in farmer, but are similar in decision fraction and uncertainty distribution (*Figure 1D*). This is because the differences in s are compensated by the differences in c.

The above scenario spells out three intuitions: (i) c varies according to past experiences of s; (ii) s and c have respective causal contributions to the formation of v; (iii) interplays between s and c realize the relativity of classification. To demonstrate how our new formalism offers a normative and quantitative account for these three intuitions, we concurrently acquired behavioral responses and functional magnetic resonance images (fMRI) from human brains engaged in a perceptual classification task that mimics the apple-sorting scenario, and proceeded as follows. First, by implementing our new formalism in the Bayesian framework (*Griffiths et al., 2012; Sheth et al., 2012*), we developed a generative model that people would adopt to perform the task, which incorporates c as a latent variable whose trial-to-trial states are inferred from recent experiences of stimuli. Then, we showed that Bayes-optimal classifiers with that generative model actualize the three intuitions, and that their behaviors well matched the actual classification behaviors of people. We further replicated the predictions of criterion inference in diverse variations of task conditions. Lastly, by searching the entire brain with multivariate pattern analysis on fMRI measurements, we identified a set of brain regions within which activity patterns signal the trial-to-trial states of s, c, and v, respectively, and further showed that each of those brain signals satisfy the causal relations with the remaining variables in our model. We concluded by discussing how conferring of the just formalism on 'internal criterion' can elucidate the cognitive and neural processes underlying perceptual classification.

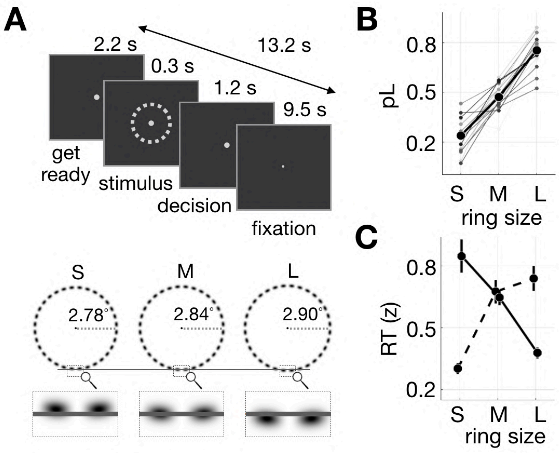


Figure 2. Task and behavioral measures. (**A**) Ring size classification task. On each trial, observers briefly viewed a ring stimulus and classified it as 'small' or 'large' within 1.2 s after stimulus onset while fixating on a spot on the display. Three ring stimuli that were slightly different in size 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. Bold and thin markers, pLs for the mean and individual observers, 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.

Results

Ring-size classification

Over consecutive trials, observers sorted rings by size into two classes, 'small' and 'large', under moderate time pressure (*Figure 2A*). 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. Observers were trained intensively with trial-to-trial feedback until reaching an asymptotic level of performance (~6,000 trials/observer) before the main fMRI experiment. We expected this intensive training to help observers understand the presence of 'veridical criterion' that determines correct and incorrect responses and to minimize possible behavioral and neural variabilities due to perceptual learning (*Kahnt et al., 2011a; Li et al., 2004; Pourtois et al., 2008; Schwartz et al., 2002*). These training sessions also allowed us to define the threshold (70.7% correct) ring sizes 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 levels of task performance (mean=75.7%, SD=6.3%) were consistent with the level of performance aimed by the training sessions, supporting that observers were likely to carry out the task with run-to-run feedback as they did with trial-to-trial feedback.

The contributions of stimuli to decision and uncertainty were summarized by plotting the proportion of 'large' decision (pL) and response time (RT), which is often considered as a behavioral proxy of decision uncertainty (*Palmer et al., 2005; Ratcliff and McKoon, 2008; Urai et al., 2017*), against ring size (*Figure 2B,C*). The pL and RT of 'large' decisions decreased as the ring size increased, and vice versa (pL, $\beta = 0.99$ ($P = 6.6\times10^{-44}$) by logistic regression; RT, $\beta = -0.17$ ($P = 8.8\times10^{-13}$) for 'large' $\beta = 0.19$ ($P = 2.4\times10^{-16}$) for 'small' decisions by linear regression).

Criterion-inference model

The generative model, an observer's causal account for how their sensory measurements are generated, is as follows: on a trial t, the true size $Z_{(t)}$ is randomly sampled from a prior distribution $p(Z)$, and the observer's sensory measurement $m_{(t)}$ is another noisy sample from a likelihood distribution $p(m|Z)$ centered around $Z_{(t)}$ (*Figure 3A*; see Methods). The size classification task, splitting population size values into small and large halves, can be defined as judging whether $Z_{(t)}$ is larger or smaller than $\tilde{Z}$, the median of the true ring sizes shown over all trials. Because the observer can access only

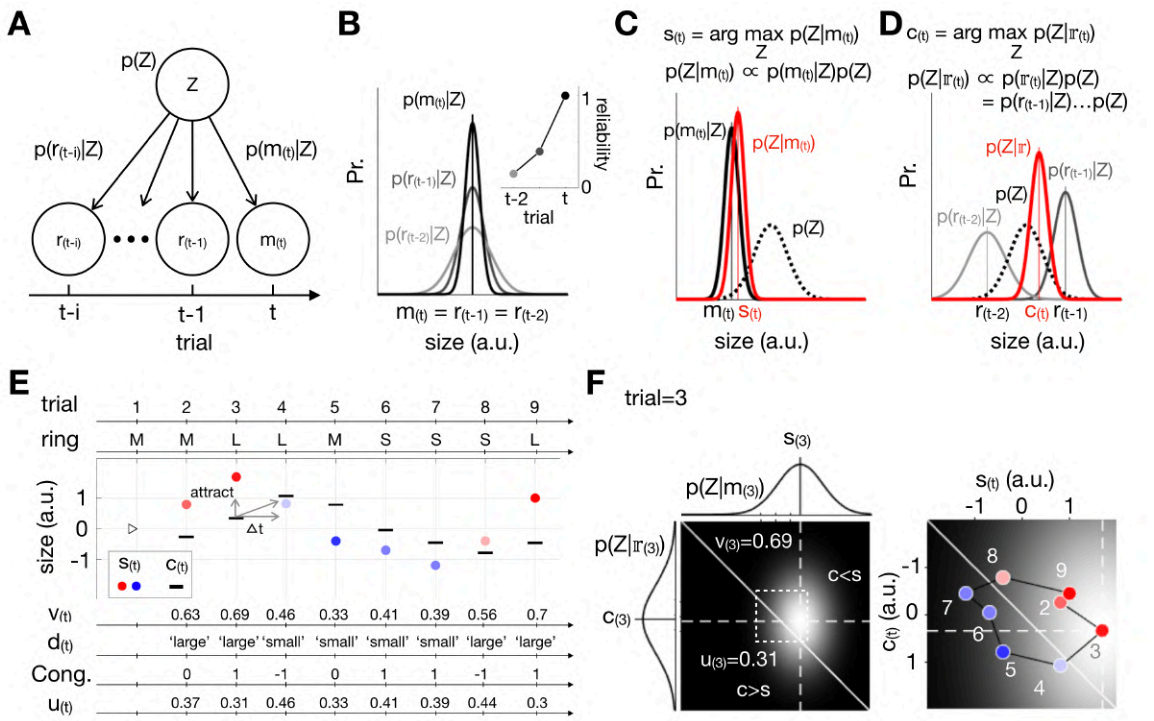


Figure 3. Criterion-inference model. (**A**) Generative model. Arrows indicate conditional dependencies. Z, stimulus size. $\mathrm{m}_{(\mathrm{t})}$, noisy observation on a current trial t. $\mathrm{r}_{(\mathrm{t}-\mathrm{i})}$, noisy observation retrieved from a past trial $\mathrm{t}-\mathrm{i}$. (**B**) Increases in measurement noise over trials. A likelihood function over ring size encodes the most likely value of $\mathrm{Z}_{(\mathrm{t})}$ and its uncertainty, which increases as trials elapse. Reliability is the reciprocal of variance of the likelihood. (**C**) Bayesian inference for a current stimulus, *s*. A posterior function (red) represents $\mathrm{s}_{(\mathrm{t})}$ and its uncertainty. (**D**) Bayesian inference for criterion, *c*. A posterior function (red) represents $\mathrm{c}_{(\mathrm{t})}$ and its uncertainty. (**E**) Illustration of decision episodes over trials. On the stimuli presented by an experimenter, an observer makes decisions by comparing *s* (colored dots) against *c* (black bars), which is updated continuously over trials. $\mathrm{v}_{(\mathrm{t})}$, decision variable, represents $\mathrm{p}(\mathrm{s}_{(\mathrm{t})} > \mathrm{c}_{(\mathrm{t})})$. The sign of '$\mathrm{v}_{(\mathrm{t})} - 0.5$' determines $\mathrm{d}_{(\mathrm{t})}$, binary decision (color hue of dots): 'large' when positive and 'small' when negative. The absolute value of '$\mathrm{v}_{(\mathrm{t})} - 0.5$' inversely reflects $\mathrm{u}_{(\mathrm{t})}$, decision uncertainty (color desaturation of dots): the greater $\mathrm{u}_{(\mathrm{t})}$, the more uncertain a decision. Cong., congruence of $\mathrm{d}_{(\mathrm{t})}$ with ring size; 1, -1: correct/incorrect; 0: M-ring. Note that $\mathrm{c}_{(\mathrm{t})}$ appears to chase after $\mathrm{s}_{(\mathrm{t}-1)}$ as indicated by small arrows. (**F**) Decision processes shown in a bivariate decision space. Left, a decision episode captured on trial 3 in (E). The bivariate conjugate of $\mathrm{p}(\mathrm{Z}|\mathrm{m}_{(3)})$ and $\mathrm{p}(\mathrm{Z}|\mathbb{r}_{(3)})$ determines $\mathrm{v}_{(3)}$ with its fraction above the identity line and $\mathrm{u}_{(\mathrm{t})}$ with its fraction in the opposite side of the identity line to where $\mathrm{s}_{(\mathrm{t})}$ and $\mathrm{c}_{(\mathrm{t})}$ are located. The light intensity corresponds to probability density. Right, pairs of $\mathrm{s}_{(\mathrm{t})}$ and $\mathrm{c}_{(\mathrm{t})}$ shown in (E) are replotted to illustrate that $\mathrm{s}_{(\mathrm{t}-1)}$ attracts $\mathrm{c}_{(\mathrm{t})}$ in the zoomed-in decision space, which is demarcated by the dotted box in the left panel. The numbers in the panel indicate the trial numbers of $\mathrm{s}_{(\mathrm{t})}$ and $\mathrm{c}_{(\mathrm{t})}$ pair shown in (E). *Figure 3-figure supplement 1* and Methods illustrate the procedure of model fitting to human observers. See also *Figure 3-figure supplement 1*.

Figure 3-figure supplement 1: Estimation of model parameters and confidence intervals.

the measured sizes of stimuli encountered so far, the Bayes-optimal solution is to infer $\mathrm{Z}_{(\mathrm{t})}$ and $\tilde{\mathrm{Z}}_{(\mathrm{t})}$, which corresponds to inferring *s* and *c* from a limited set of noisy measured sizes $\{\mathrm{m}_{(1)}, \mathrm{m}_{(2)}, \dots, \mathrm{m}_{(\mathrm{t})}\}$. 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 (*Figure 3B*). To distinguish from the likelihood of Z for the current-trial measurement $\mathrm{p}(\mathrm{m}_{(\mathrm{t})}|\mathrm{Z})$, we denoted the likelihood of Z for the measurement 'retrieved' from an elapsed trial $\mathrm{t}-\mathrm{i}$ by $\mathrm{p}(\mathrm{r}_{(\mathrm{t}-\mathrm{i})}|\mathrm{Z})$ and assumed that its precision increases as trials elapse (*Gorgoraptis et al., 2011; Zokaei et al., 2015*), as represented by the widening curves shown in *Figure 3B*.

On each trial, the Bayes-optimal classifier infers *s* and *c* by inversely propagating $\{\dots, \mathrm{r}_{(\mathrm{t}-2)}, \mathrm{r}_{(\mathrm{t}-1)}, \mathrm{m}_{(\mathrm{t})}\}$ over the generative model. The inferred size $\mathrm{s}_{(\mathrm{t})}$ is the most probable value of Z given the measurement on a current trial, as captured by the following equation :

$$\begin{aligned} \mathrm{s}_{(\mathrm{t})} = \hat{\mathrm{Z}}_{(\mathrm{t})} &= \arg\max_{\mathrm{Z}} \mathrm{p}(\mathrm{Z}|\,\mathrm{m}_{(\mathrm{t})}) \\ &\propto \arg\max\nolimits_{\mathrm{Z}} \mathrm{p}(\,\mathrm{m}_{(\mathrm{t})}|\mathrm{Z})\mathrm{p}(\mathrm{Z}) \qquad (\textit{Equation 1}) \end{aligned}$$

, where the variance of the posterior reflects the precision of $s_{(t)}$ (*Figure 3C*; see Methods). On the other hand, the inferred $c_{(t)}$, 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:

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

, where $\mathbb{r}_{(t)} = \{\ldots, r_{(t-2)}, r_{(t-1)}\}$ (see Methods). 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 (*Figure 3D*). *Equation 2* implies that $c_{(t)}$ is more attracted towards a recent stimulus than to older ones, because of the decay of representational reliability of working memory as trials elapse (*Figure 3D*). On each trial, the Bayes-optimal classifier performs the task by deducing $v_{(t)}$ from $s_{(t)}$ and $c_{(t)}$ and translating it into a binary decision $d_{(t)}$ with a degree of uncertainty $u_{(t)}$ (*Figure 3E*; see Methods); $v_{(t)}$ 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 (*Sanders et al., 2016*) $(p(s_{(t)} < c_{(t)}|d_{(t)} = \text{'large'})$ or $p(s_{(t)} > c_{(t)}|d_{(t)} = \text{'small'}))$. Because all of the decision outcomes are deduced from $s_{(t)}$ and $c_{(t)}$, a bivariate decision space of $p(Z|m_{(t)})$ and $p(Z|\mathbb{r}_{(t)})$ effectively captures both of the present decision outcomes and the attraction of c toward the past stimuli (*Figure 3F*). In summary, we propose that, as the ring size (Z) varies over trials, the classifier (i) infers s on a current trial and the median values of the ring sizes encountered over 'past' trials as c, (ii) deduces the values of v from s and c, and (iii) makes decisions (d) with varying degrees of uncertainty (u).

To verify our proposal, we created eighteen Bayes-optimal classifiers by fitting the model to eighteen human observers' decisions respectively (*Figure 3-figure supplement 1*; see Methods). We simulated decisions and decision uncertainties made of the Bayes-optimal classifiers on the stimuli sequence encountered by their human partners, and examined the correspondences between the human and Bayes-optimal classifiers regarding each of the aforementioned three intuitions.

Intuition I: criterion relates past stimuli to current decisions

The model makes testable predictions of how past stimuli relate to current decisions through multiple regression analysis (see Methods). As exemplified in the apple-sorting scenario, whereby different farmers had formed different c according to recent experiences of apples (*Figure 1A*), the model predicts that c is more attracted to more recent measurements (*Figure 3D,E*). This entails that current decisions are more strongly repelled from more recent stimuli (*Figure 4A*), and that decision uncertainties are high when current decisions are congruent with recent stimuli (*Figure 4B*), as in the case of 'O by F', whereby decisions are repelled from 'large' and their uncertainties are high on 'large' decisions (*Figure 1D*). By contrast, our model predicts negligible effects of past decisions on current decisions and uncertainties.

Observers' decisions and RTs confirmed the model predictions when they are regressed onto past stimuli, as well as current stimuli and past decisions (*Figure 4C,D*). In addition, based on previous studies suggesting that the dorsal anterior cingulate cortex (dACC) is a locus of decision uncertainty (*Behrens et al., 2007; Shenhav et al., 2014; Sheth et al., 2012*), we also examined whether the blood-oxygen-level-dependent (BOLD) reponses in the dACC are consistent with the model predictions of $u_{(t)}$. We found that the dACC activity at 5.5 s from stimulus onset ($dACC_{u5}$) was not just correlated with $u_{(t)}$ $(\beta = 0.10, P = 2.4\times10^{-7})$ (*Figure 4F; Figure 4-figure supplement 1, 2*) but also significantly regressed onto the congruencies between past stimuli and decisions as predicted by the model (*Figure 4E*).

The variances explained by the constant-c model, whereby c was fixed at a constant and thus effects of past stimuli were neglected, increased by 29%, 34% and 49% for human decisions, RT, and dACC activity, respectively, compared to the c-inference model ($\frac{Var_{inferred} - Var_{constant}}{Var_{constant}} \times 100$; *Figure 4-figure supplement 3A-F*; see Methods). Furthermore, the uncertainty estimates of the constant- c model were no longer correlated with $dACC_{u5}$ $(P_{FDR} > 0.05)$ (*Figure 4F; Figure 4-figure supplement 3G*), which indicates that the inferred c has an essential contribution to the cortical representation of decision uncertainty.

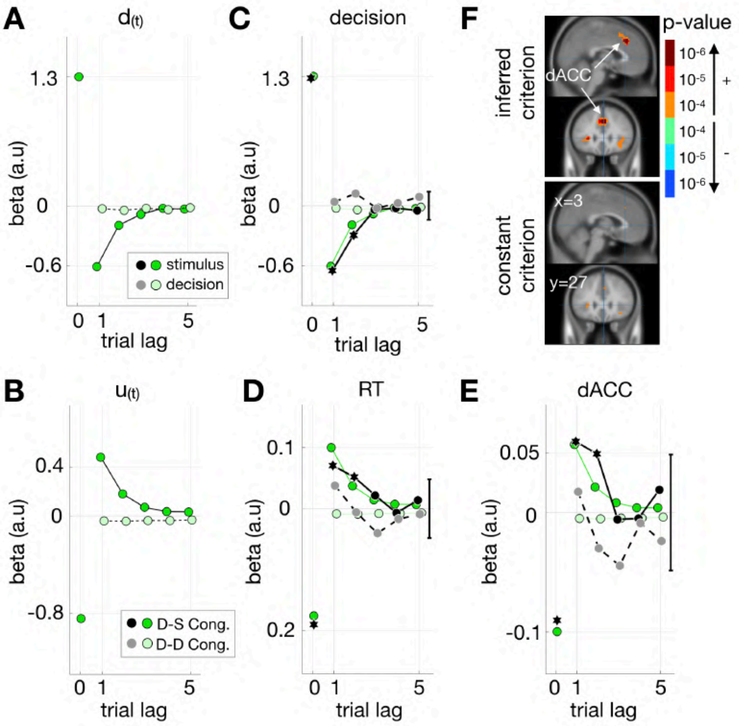


Figure 4. Criterion relates past stimuli to current decisions. (**A, B**) Model prediction for impacts of past stimuli and decisions on present decisions. (**A**) Model predictions of multiple logistic regressions of current decisions onto stimuli and past decisions. (**B**) Model predictions of multiple linear regressions of current decision uncertainties onto the congruence of a current decision with stimuli and past decisions. (**C**) Multiple logistic regressions of observed current decisions onto stimuli and past decisions. (**D, E**) Multiple linear regressions of RT and dACC activity onto the congruence of current decisions with stimuli and past decisions. (A-E) Black and green symbols represent the coefficients for observed and simulated data, respectively. (C-E) Observed coefficients that are significantly deviated from zero (C: pairwise t-test, P < 0.05; D, E: GLMM, P < 0.05) are indicated by black hexagons. For clarity, only the average of 95% confidence intervals (CIs) for the mean of observed coefficients (C: bootstrap CI; D, E: GLMM CI) are indicated by vertical black bars. None of the observed coefficients significantly deviates from the simulated coefficients (C: pairwise t-test, P > 0.05; (D, E) the simulated mean was located within GLMM 95% CI of the observed mean). (**F**) Statistical significances (p-values) of linear regression (GLMM) of dACC activity 5.5 s after stimulus onset onto decision uncertainty estimated by the inferred-criterion model (top) or by the constant-criterion model (bottom). The signs of regression coefficients are indicated by '+' and '—'. dACC, dorsal anterior cingulate cortex. See also *Figure 4-figure supplement 1, 2, 3, 4, 5*.

Figure 4-figure supplement 1: Specifications of the ROIs in which activity was correlated with decision uncertainty.

Figure 4-figure supplement 2: Univariate BOLD signals correlated with decision uncertainty.

Figure 4-figure supplement 3: Limitations of the constant-criterion model

Figure 4-figure supplement 4. Verifying the model predictions of the impacts of past stimuli on current decisions with different sets of stimuli, feedback types, and inter-trial-interval lengths.

Figure 4-figure supplement 5. Verifying the model predictions of the impacts of past stimuli on current decision RTs with different sets of stimuli, feedback types, and inter-trial-interval lengths.

Intuition II: dissecting stimulus and criterion contributions to classification

Our model decomposes the trial-to-trial variability in decision and uncertainty into that originating from s and that from c. Speaking in terms of the apple-sorting scenario, our model specifies why and how much a current decision is ascribed to the farm from which an apple came (*Figure 1B*) and to the farm from which a farmer came (*Figure 1C*).

The model's dissection of the s and c effects is illustrated in the bivariate decision space (*Figure 5A*). The probability of making the 'large' decision ($\mathrm{pL}_{(t)}$), which corresponds to the truncated bivariate distribution that satisfies $s_{(t)} > c_{(t)}$, is greater for 'L' rings than for 'S' rings, but decreases as $c_{(t)}$ increases, regardless of ring size (*Figure 5A,B*). The decision uncertainty ($u_{(t)}$) of 'large' decisions, which corresponds to how close the mean of the truncated bivariate distribution that satisfies $s_{(t)} > c_{(t)}$ is to the equality of $s_{(t)}$ and $c_{(t)}$, is higher for 'S' rings than for 'L' rings, but increases as $c_{(t)}$ increases, regardless of ring size (*Figure 5A,C*). For 'small' decisions, the s and c effects on decision uncertainty can be described in similar ways, based on the truncated bivariate distributions that satisfy $s_{(t)} < c_{(t)}$ (*Figure 5A,C*).

The human data matched the Bayes-optimal classifier's $d_{(t)}$ and $u_{(t)}$ in both qualitative and quantitative aspects. Qualitatively, the decision fractions were concurrently regressed onto the current

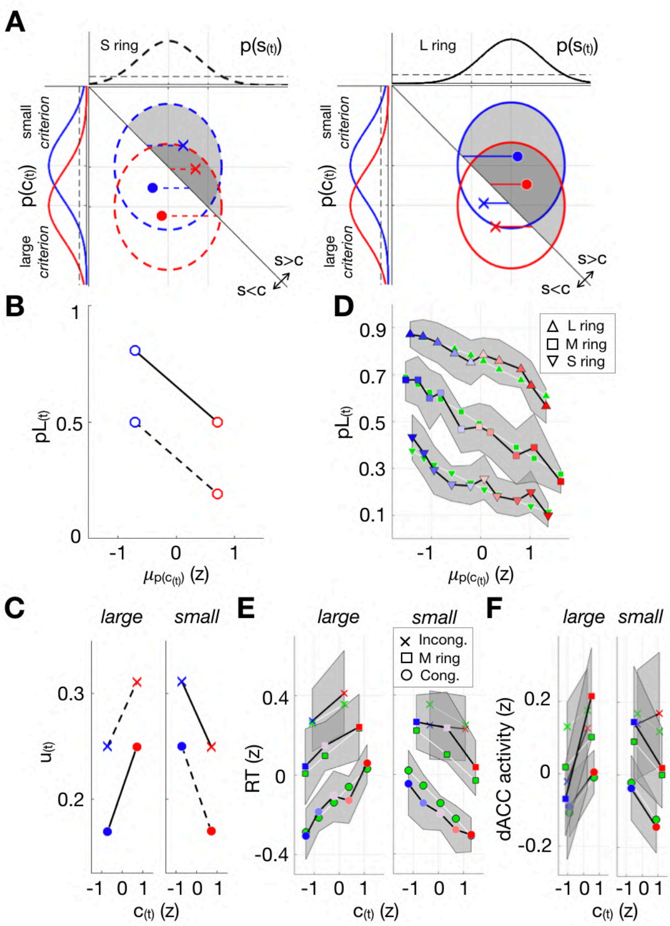


Figure 5. Dissecting stimulus and criterion contributions to classification. (**A**) Bivariate distributions of $s_{(t)}$ and $c_{(t)}$ for different combinations of ring size and criterion size. Ellipsoids demarcate the distribution contours defined at 20% of the peak, as indicated by the dashed gray lines traversing the univariate distributions. Bivariate distributions shift horizontally depending on ring size, leftward for S ring (dashed) and rightward for L ring (solid), or vertically depending on criterion size, downward for large criterion (red) and upward for small criterion (blue), which allow us to dissect and predict the respective contributions of $s_{(t)}$ and $c_{(t)}$ to decision fraction and uncertainty. Fractions of trials with 'large' decisions ($pL_{(t)}$) are indicated by the shaded areas. Decision uncertainty ($u_{(t)}$) is represented by how close the means of the truncated bivariate distributions conditional on decisions ($s_{(t)} > c_{(t)}$ on 'large' decision and $s_{(t)} < c_{(t)}$ on 'small' decision) are to the identity ($s_{(t)} = c_{(t)}$) line, as indicated by the lengths of horizontal red and blue lines. The 'o' and 'x' symbols represent trials on which decisions are congruent and incongruent with stimuli, respectively (see *Figure 3E* for the definition of congruence). (**B, C**) $pL_{(t)}$ (**B**) and $u_{(t)}$ (**C**) plotted against mean criteria for the four bivariate distributions shown in (A). Colors and line styles match those in (A). (**D-F**) $pL_{(t)}$ (**D**), RTs (**E**), and dACC activity (**F**) are plotted against the means of inferred criteria for each stimulus (D) or decision-stimulus congruency (E, F) conditions (blue or red symbols), juxtaposed with model simulation results (green symbols). Trials were sorted and binned by $c_{(t)}$, as indicated by the symbols' color and hue that vary in a scheme similar to that in (A-C). Shaded areas represent 95% bootstrap confidence intervals of data means. None of the observed data significantly deviated from the model predictions (pairwise t-test, P > 0.05) . See also *Figure 5-figure supplement 1*.

Figure 5-figure supplement 1: Verifying the model's ability to account for respective contributions of stimulus and criterion with different sets of stimuli, feedback types, and inter-trial-interval lengths.

stimulus and the inferred criterion (to be exact, the expected value of model *c* estimates $\mu_{p(c_{(t)})}$; see Methods for its definition and derivation) as the model predicts (multiple logistic regression, $\beta_{Z_{(t)}} = 1.07$ ($P = 1.8\times10^{-35}$), $\beta_{\mu_{p(c_{(t)})}} = -0.59$ ($P = 2.3\times10^{-20}$); *Figure 5D*). Likewise, the RTs and dACC responses were, when decisions were controlled, both regressed onto the current stimuli and the inferred criterion as the model predicts (multiple regression, RT: $\beta_{Z_{(t)}} = -0.19$ ($P = 6.9\times10^{-15}$) , $\beta_{\mu_{p(c_{(t)})}} = 0.12$ ($P = 2.5\times10^{-5}$), for 'large' decision; $\beta_{Z_{(t)}} = 0.20$ ($P = 3.5\times10^{-16}$) , $\beta_{\mu_{p(c_{(t)})}} = -0.081$ ($P = 9.4\times10^{-4}$), for 'small' decision; dACC: $\beta_{Z_{(t)}} = -0.055$ ($P = 0.053$) , $\beta_{\mu_{p(c_{(t)})}} = 0.084$ ($P = 5.7\times10^{-4}$) ,

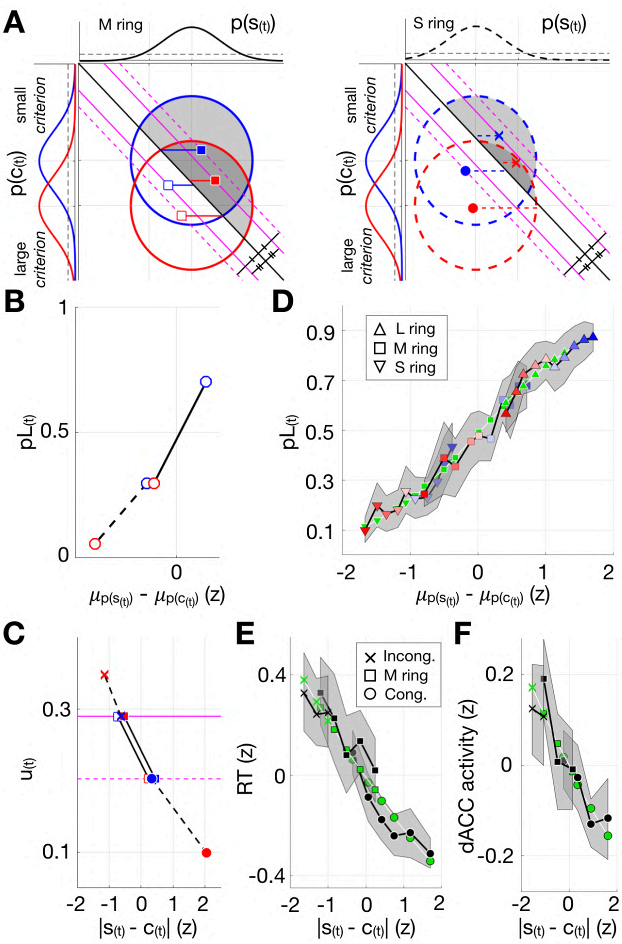


Figure 6. Relativity of classification. (**A**) Bivariate distributions of $s_{(t)}$ and $c_{(t)}$ for different combinations of ring size and criterion size. Formats, symbols, and lines are identical to those in *Figure 5A*, unless stated otherwise. Symbols falling on the same diagonal lines represent metameric decisions. Symbols falling on the diagonal lines of the same style (either solid or dashed), which are equivalent to each other in distance from the identity line, represent decisions with the same degree of decision uncertainty. The stimulus-decision congruence is indicated by markers of 'o' (congruent), 'x' (incongruent), and '□' (M ring). (**B**) Fractions of 'large' decisions ($\mathrm{pL}_{(t)}$) plotted against the difference between the mean stimulus and criterion estimates. Colors and line styles match those in (A). (**C**) Decision uncertainty estimates plotted against the absolute difference between the stimulus and criterion estimates. Symbols and colors match those in (A). Dashed and solid purple lines in (A) are replotted here to further clarify the relationship between (A) and (C). (**D**) Observed $\mathrm{pL}_{(t)}$ are plotted against the binned difference between the mean stimulus and criterion estimates for each stimulus condition. The color spectrum from blue to red represents the increasing size of criterion estimates within a given condition. (**E, F**) Observed RTs (**E**) and dACC activity (**F**) plotted against the binned absolute difference between the stimulus and criterion estimates for each stimulus-decision congruence condition. (D-F) Observed data juxtaposed with model simulation results (green symbols). Shaded areas, 95% bootstrap confidence intervals of mean of data. None of the observed data significantly deviated from the model predictions (pairwise t-test, P > 0.05). (*Figure 6-figure supplement 1* specifies whether each of the observed pLs, RT and dACC activity between a bin pairs are indistinguishable or distinguishable). See also *Figure 6-figure supplement 1, 2*.

Figure 6-figure supplement 1: Definitions and predictions of decision-probability (pL) metamers and anti-metamers by the criterion-inference model.

Figure 6-figure supplement 2: Verifying the model's ability to account for the relativity of classification with different sets of stimuli, feedback types, and inter-trial-interval lengths.

for 'large' decision; $\beta_{\mathrm{z}_{(t)}} = 0.11\ (\mathrm{P} = 5.8\times10^{-5})$, $\beta_{\mu_{\mathrm{p}(\mathrm{c}_{(t)})}} = -0.059\ (\mathrm{P} = 0.020)$, for 'small' decision; *Figure 5E,F*). Quantitatively, the Bayesian simulations fell within the 95% confidence intervals defined at all data bins inspected (*Figure 5D–F*).

Intuition III: relativity of classification

Our model explains the relativity of classification at both computational and algorithmic levels (*Marr and Poggio, 1976*). At the computational level, what to be computed in the classification task is the inequality between s and c. This inequality computation, at the algorithmic level, is accomplished by forming a probabilistic representation of the difference between $\mathrm{s}_{(t)}$ and $\mathrm{c}_{(t)}$, which is the decision variable, $\mathrm{v}_{(t)}$. As inuited in the apple-sorting scenario, trials are identical in $v_{(t)}$ as long as they are matched in relative difference between $s_{(t)}$ and $c_{(t)}$, and increasingly differ as the relative difference increases. In the bivariate decision space (*Figure 6A*), this means that decisions are (i) indiscernible as long as the expected values of $s_{(t)}$ and $c_{(t)}$ fall on any single lines parallel to the equality line and (ii) discernible only to the extent to which the joint coordinates of $s_{(t)}$ and $c_{(t)}$ are distant from one another along the dimension perpendicular to the equality. Consequentially, the model predicts that the decision fraction and uncertainty vary only as single, monotonic functions of the signed (*Figure 6B*) and unsigned (*Figure 6C*) differences between $s_{(t)}$ and $c_{(t)}$, respectively.

When trials were sorted by the differences between $s_{(t)}$ and $c_{(t)}$, human pLs, RTs, and dACC responses all constituted single, seamless psychometric, chronometric, and neurometric curves, respectively, which indicates that what governs classification is 's relative to c' (*Figure 6D–F*). As consequences, pLs, RTs, and dACC responses became 'metameric' (significantly indiscernible) between the trials in which a physical difference between stimuli was compensated by a counteracting difference in $c_{(t)}$, or 'anti-metameric' (significantly discernible) between the trials in which a physical sameness between stimuli was accompanied by a substantial difference in $c_{(t)}$ (*Figure 6-figure supplement 1*), which supports the model's prediction of metameric classifications. The model's simulations fell within the 95% confidence intervals defined at all data bins inspected (*Figure 6D–F*).

The immunity of criterion inference to changes in participants, stimulus range, feedback type, and inter-trial interval

Note that some of specific conditions in the main experiment – such as the absence of trial-to-trial feedback, the limited number of ring sizes, and the long inter-trial interval – had to be adopted for the concurrent collections of behavioral and fMRI data. To check the immunity of the criterion-inference model to changes in such conditions, we subjected the model to further data sets that were collected outside the MRI scanner (with the purposes different from those of the current work) from different populations of observers with diverse variations in experimental conditions (*Choe et al., 2014, 2016; Lee et al., 2016*). These variations included: (1) different population of observers (the auxiliary data sets (AD) #1,2,3) (2) presence of trial-to-trial feedback (instead of run-to-run feedback; AD#2,3); (3) presentation of larger numbers and wide ranges of ring sizes (instead of only 3 threshold-level ring sizes; AD#2,3); (4) separation of trials with shorter inter-trial intervals (instead of the long interval of 13.2 s; AD#2,3) (see *Figure 4-figure supplement 4A* for details). Regardless of these variations, a total of 53 new oberservers exhibited all the key behavioral signatures that are predicted by the model and were observed in the main fMRI experiment: the impacts of past stimuli on current decision and uncertainty (*Figure 4-figure supplement 4,5*); respective contributions of stimulus and criterion (*Figure 5-figure supplement 1*); relativity of classification (*Figure 6-figure supplement 2*). When fit to the new data sets, the criterion-inference model readily captured the actual classification behaviors of people and predominated over the constant-criterion model in data accountability (*Figure 4-figure supplement 4D,5C*). These results support the generalizability of the criterion-inference model by suggesting that 'criterion inference' is an essential component of perceptual classification.

Causality between c, s, and v in brain activity

So far, we have validated the model by verifying its predictions for d and u in the behavioral data and dACC activity. Next, we set out to verify the model's predictions for its core latent variables, c, s, and v, from which d and u were deduced (*Figure 3E*), in the patterns of brain activity. These predictions are two fold: first, the model predicts the presence of brain signals that are correlated with the trial-to-trial states of c, s, and v; second, if such brain signals exist, their trial-to-trial variabilities must satisfy their causal relations with all the remaining variables in the model, including the manipulated (Z), latent (c, s, v, u) and observed (d) variables. Verifying these predictions will lend strong support to our formalism by endowing its latent variables and their causal relations with neural presence.

For comprehensive search of candidate brain regions within which activity was correlated with c, s, and v, we opted to use 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*). Then, for each target variable (*Figure 7A*), we deduced a set of regressions from the causal relationships of the target variable with the other variables. For example, the decoded y_c, by

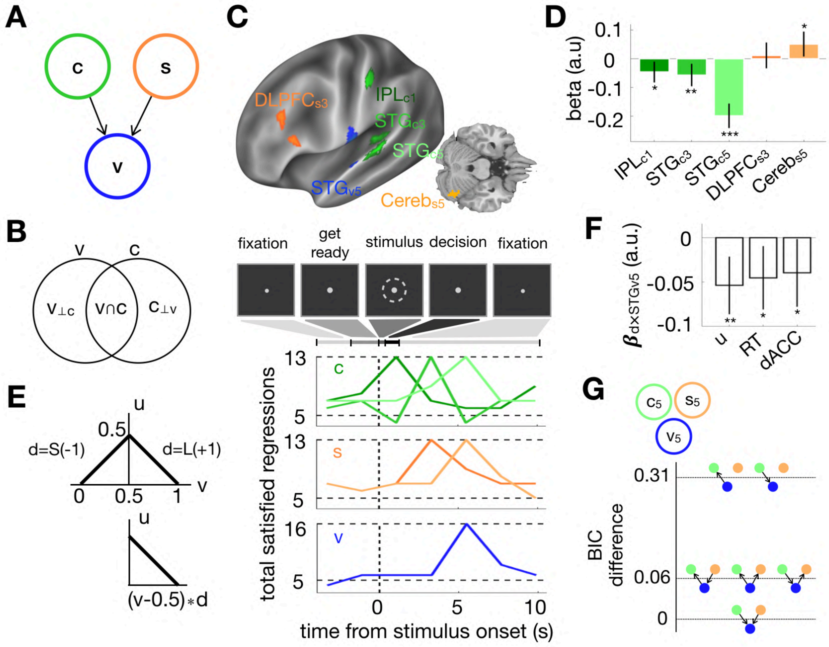


Figure 7. Causality in brain activity. (**A**) Graphical representation of the causal relationships between c, s, and v. (**B**) Venn-diagram representation of the structure of variances between c and v: $v \cap c$, shared variance; $v_{\perp c}$ and $c_{\perp v}$, unexplained variances of c and v by each other. (**C**) Brain signals of c, s, and v that satisfied all the regressions deduced from the causal structure of variables in the model (*Figure 7-figure supplement 1* specifies all the regressions). Top, six brain loci for c (green), s (orange), and v (blue) signals. Bottom, the numbers of the regressions satisfied are plotted as a function of time relative to stimulus onset for the c, s, and v signals (*Figure 7-figure supplement 2, 3* specifies the ROIs correlated with each latent variable). (**D**) Coefficients of multiple regressions of the v signals onto the c and s signals. (**E**) Functional representation of the causal relationships between v, d, and u. Top, v and u are associated but uncorrelated. Bottom, v and u become correlated when controlled for d. (**F**) Coefficients of the 'decision x v signal' interactive term in the regression of the model estimate of u, observed RTs, and brain signal of u, respectively. (**G**) Bayesian networks and their relative BIC scores that explain the causal relation between the brain signals of STG_{c5}, $Cereb_{s5}$, and STG_{v5}. For other candidate structures, see *Figure 7-figure supplement 4*. *P < 0.05, **P < 0.01, ***P < 0.001. Error bars represent 95% GLMM confidence intervals. IPL_{c1}: c decoded from inferior parietal lobe at 1.1 s; $STG_{c3\,(or\,s5)}$: c decoded from superior temporal gyrus at 3.3 s (or 5.5 s); $DLPFC_{s3}$: s decoded from dorsal lateral prefrontal cortex at 3.3 s; $Cereb_{s5}$: s decoded from cerebellum at 5.5 s; STG_{v5}: v decoded from superior temporal gyrus at 5.5 s. The time represents time elapse from stimulus onset. See also *Figure 7-figure supplement 1, 2, 3, 4*.

Figure 7-figure supplement 1: Deduction of the regressions for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ from the causal structure of manipulated (stimuli), observed (decisions), and latent model variables (c, s, v, u, and d).

Figure 7-figure supplement 2: Maps of the numbers of regression tests satisfied for the latent variables in the model (s, c, and v).

Figure 7-figure supplement 3: Specifications of the ROIs in which activity was informative of the latent variables in the model (s, c, and v).

Figure 7-figure supplement 4: Correspondence between the causal graph of the latent variables that is prescribed *a priori* by the model ($c \rightarrow v \leftarrow s$) and the *maximum likelihood* causal graph that is inferred from the brain signals representing those variables.

definition ($v \sim s - c$), must be regressed positively on itself ($y_c \sim +c$) and negatively on v ($y_c \sim -v$), but must not be regressed on v when v is orthogonalized against c ($y_c \nsim -v_{\perp c}$) (*Figure 7B*). In this way,additional regressions were further deduced from the causal relationship of c with the remaining latent (s,u), observable (d), and manipulated (Z) variables, resulting in 13 regressions for c. Likewise, we deduced 13 and 16 regressions for s and v, respectively (*Figure 7-figure supplement 1*). We then evaluated each candidate region by testing whether the decoded target variable satisfied those regressions. Six regions survived the test (*Figure 7C; Figure 7-figure supplement2, 3*). The signal of c appeared in three different regions at different time points relative to stimulus onset, as follows: a region in the inferior parietal lobe at 1.1 s (IPL_{c1}), followed by two regions in the superior temporal gyrus at 3.3 and 5.5 s (STG_{c3}, STG_{c5}). The signal of s appeared both in the dorsolateral prefrontal cortex at 3.3 s ($DLPFC_{s3}$) and in the cerebellum at 5.5 s ($Cereb_{s5}$). The signal of v appeared in the left superior temporal gyrus at 5.5 s (STG_{v5}).

The interplays between brain signals also satisfied the causal structure defined by the model. First, the interplays between c, s, and v singals were well captured by a multiple regression model, $STG_{v5} \sim \beta_{s3}DLPFC_{s3} + \beta_{s5}Cereb_{s5} + \beta_{c1}IPL_{c1} + \beta_{c3}STG_{c3} + \beta_{c5}STG_{c5}$, which was derived from the 'common-effect' structure of causal relationships between the variables ($c \rightarrow v \leftarrow s$) (*Figure 7A*). The regression coefficients matched the model predictions in sign for all of the five brain signals and were all significant except for $DLPFC_{s3}$ ($\beta_{s3} = 0.011, P = 0.55$; $\beta_{s5} = 0.047, P = 0.013$; $\beta_{c1} = -0.046, P = 0.0088$; $\beta_{c3} = -0.056, P = 0.0037$; $\beta_{c5} = -0.15, P = 1.9\times10^{-17}$; *Figure 7D*). 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 $(\mathrm{d} \leftarrow \mathrm{v} \rightarrow \mathrm{u})$ (*Figure 7E*). Importantly, the interactive regression on $\mathrm{d}\times\mathrm{STG_{v5}}$ was significant for $\mathrm{dACC_{u5}}$ $(\beta_{\mathrm{d\times v}} = -0.040, \mathrm{P} = 0.04)$, as well as for RT $(\beta_{\mathrm{d\times v}} = -0.045, \mathrm{P} = 0.013)$ and $\mathrm{u_{(t)}}$ $(\beta_{\mathrm{d\times v}} = -0.054, \mathrm{P} = 0.0012$; *Figure 7F*). Finally, to complement the model-driven regression analyses performed above, we used a data-driven approach to calculate 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; *Figure 7-figure supplement 4*). In line with the regression analyses, the causal structure of $\mathrm{c} \rightarrow \mathrm{v} \leftarrow \mathrm{s}$ emerged as one of the most probable causal structures (*Figure 7G; Figure 7-figure supplement 4*).

Discussion

The present work gives c firm computational and representational presences in classification by specifying the generative model, which defines the origins of s inference and c inference, dissects their respective contributions to the trial-to-trial variability of classification, and accounts for how their interplay underlies the relativity of decision making. Such presences were further secured against diverse variations in task conditions such as different participants, stimulus range, feedback type, and inter-trial interval. Our work also give c a neural presence by uncovering the brain network within which the brain signal of c interacts with the brain signals of s and v, as the model prescribes.

Bayes-optimal criterion inference

Although the true c is the population median of quantities, without access to quantities that will be encountered in the future, people have to rely on the quantities experienced so far. Further, due to memory limitations (*Gorgoraptis et al., 2011; Zokaei et al., 2015*), these past quantities cannot be retrievd entirely or reliably. Our model claims that, under these constraints, the best possible way to access the true c is to probabilistically integrate past observations by assigning them optimal weights based on their reliability.

In this sense, our model is Bayes-optimal, and our results imply that people performs our task with bounded rationality (*Simon, 1955*) by considering the limits of their own memory faculty. In this regard, the variability of the inferred c across trials should not be taken as suboptimal digressions from the true, fixed c, but rather as the best possible guesses for the true c, which vary trial to trial because of the variability of retrieved past quantities. This may explain the seemingly suboptimal behavioral strategies used by animals in an olfactory classification task (*Zariwala et al., 2013*).

Breaking the monopoly of stimulus inference

In classification models without c inference (*Gold and Shadlen, 2007; Green and Swets, 1966; Kepecs et al., 2008; Meyniel et al., 2015*), s inference monopolizes the variability of v via a single causal chain ($s \rightarrow v$; *Figure 8A*). The c-inference model creates another causal route to v by having the inferred c as another parent for v ($c \rightarrow v$) and a child of past stimuli ($Z \rightarrow \mathbb{r} \rightarrow c$) (*Figure 8B*). This alternative route breaks the monopoly of s inference and brings fresh perspectives to several aspects of classification processes, which have not been adequately addressed by the models without c inference.

First, c inference plays the role of a 'historian' by letting past episodes partake in a present decision. Our model's account for this role of c warns that one should not regress current decisions onto present stimuli without including past stimuli a co-regressor (*Fründ et al., 2014*).

Second, c claims a fair share of credit for the variability in classification, which has so far been monopolized by $s$in previous formalisms (*Gold and Shadlen, 2007*). In our work, the contributions of c to decision, RT and dACC activity were about a third of that of s (*Figure 4-figure supplement 3*).

Third, c, as a legitimate parent to v along with s, precisely defines the relativity and metamerism in classification with a rigor to predict when and to what extent decisions become indiscernible or discernible.

Finally, as a new species brings a native ecosystem a new configuration (*Richardson and Rejmánek, 2011*), c reconfigures the old causal architecture of classification (*Figure 8*). This, by redefining s, v, and u, leads to better quantitative predictions for the variability in decision and RT, and detection of the brain signal of u, which was undetectable by the constant-c model, in the dACC.

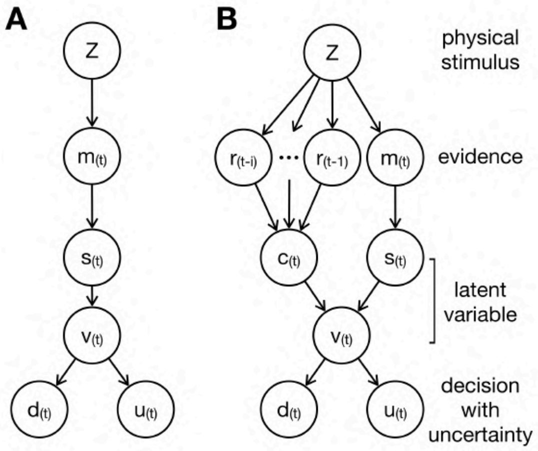


Figure 8. Causal structures of variables without and with criterion inference. (**A**) Causal structure without c inference. The information originating from a current value of the stimulus variable (Z) flows in v only via a single causal stream $(\mathrm{Z} \rightarrow \mathrm{m} \rightarrow s \rightarrow v)$. (**B**) Causal structure with c inference. The information originating from a current value of Z is joined by another originating from past values of Z via the conduit of c inference $(\mathrm{Z} \rightarrow \mathrm{r} \rightarrow c \rightarrow v)$, creating the confluence of information streams of past experiences and current sense of external stimulus quantities. This information confluence due to criterion inference redefines v and thus distinguishes v and s. See also *Figure 8-figure supplement 1, 2*.

Figure 8-figure supplement 1: Specifications of the latent variables of the criterion-inference model.

Figure 8-figure supplement 2: The 'classification' with a generative model for 'categorization' cannot give a coherent account for the history effects of previous stimuli and decisions on current decisions

Brain signals of inferred criterion

The brain signals of c appeared initially at the IPL and migrated towards the STG, where the v signal 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 turned into c to partake in classification of current stimuli. This conjecture appears to be consistent 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 priamte IPL (*Goard et al., 2016; Hanks et al., 2015; Roitman and Shadlen, 2002*) – selectively diminished the influence of previous stimuli (history effects) but left the task performance on current stimuli intact (*Akrami et al., 2018*), indicating a critical role of the IPL in encoding previous stimulus information. In addition, perturbating the PPC after current stimulus onset failed to diminish history effects, suggesting that the information of previous stimuli is likely to be maintained in some brain regions other than the PPC (*Hwang et al., 2017*). On the other hand, the c signal at the STG is 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. The functional flow from the IPL to the STG is further corroborated by their anatomical connection (*Seltzer and Pandya, 1994*). In line width these findings, our data showed that c decoded in IPL temporally preceded current stimulus onset whereas c in STG occurred at or following current stimulus onset.

Indirect neural signatures of c have been recently reported (*Crapse et al., 2018; Li et al., 2009; Luo and Maunsell, 2018*), but the neural correlate of c, per se, has been explored by one previous study, which reported the inferior temporal pole region as the locus of c (*White et al., 2012*). Unfortunately, we could not confirm this report, because this region was not fully imaged for all participants (see Methods). Nonetheless, their failure of finding neural correlates of c in IPL or STG could be due to the fact that the effect of past stimuli was not considered in defining c. This conjecture is grounded in the contrast between the constant-c model's failure to detect the u signal in the dACC and our c inference model's redemption of the u signal in dACC.

Brain signals of inferred stimulus and decision variable

The representational inconsistency between s and c is problematic for dissecting neural signals of s and v, because s, a single parent of v, is highly correlated with v in the constant-c model (*Figure 8A, Figure 8-figure supplement 1A,B*), which could have led to arbitrary or ambiguous interpretations of given neural activities either as s or as v, depending on research contexts or objectives (*Baumann et al., 2015; Curtis and Lee, 2010; Haller et al., 2018; Heekeren et al., 2006; Kim and Shadlen, 1999*). By contrast, our model, in which s is only regressed onto a current stimulus whereas v both onto a current and past stimuli (*Figure 8B, Figure 8-figure supplement 1B*), could identify distinct brain signals of s and v.

The presence of s signal in the DLPFC is compatible with its suggested roles in maintaining current sensory information during decision-making (*Curtis and Lee, 2010; Haller et al., 2018*). Our findings also suggest that the cerebellum encodes task-relevant stimulus features during classification, in line with previous studies indicating the involvement of the cerebellum in sensory perception (*Baumann et al., 2015; Gao et al., 1996*). Note that the latent variable s should not be confused with physical stimulus size Z, which is a variable manipulated by experimenters (*Figure 3; Figure 8*). Accordingly, the s signals found in the DLPFC and cerebellum did not represent external stimuli per se but the inferred estimates of external stimuli that partake in classification (*Figure 8-figure supplement 1C*). In contrast, the s signals in the retinotopic visual areas were much weaker than those in the DLPFC and cerebellum (*Figure 7-figure supplement 2B*). To further explore the prensence of s signals in the early visual cortex, we revisited one of our previous imaging studies (*Choe et al., 2014*), where high-resolution fMRI responses were acquired in V1 while observers performed the same task as in the current study, and examined the V1 population activity with a fine-retinotopy-based population decoding analysis. Unlike the DLPFC and cerebellum, where the brain signals were significantly regressed onto both external stimuli and decisions made by observers as expected from the causal chain through s ($Z \rightarrow m \rightarrow s \rightarrow v \rightarrow d$; *Figure 8-figure supplement 1D*), the V1 activity was regressed only onto external stimuli but not onto decisions (*Figure 8-figure supplement 1E*). This implies, as concluded previously (*Choe et al., 2014; Jasper et al., 2019; Lee et al., 2007; Nienborg and Cumming, 2009*), that the V1 population activity represents physical stimuli (Z) robustly but its trial-to-trial variability does not seem to be causally involved in determining the trial-to-trial variability in choice, which contradicts the property of s in our model.

Given the location of v at the confluence of the two upstream ($Z \rightarrow s \rightarrow v$; $Z \rightarrow c \rightarrow v$) and two downstream flows of information ($v \rightarrow d$; $v \rightarrow u$), the v signal in the STG should be considered importantly because its variability was consistent with the nuanced interplays between all the four streams. That said, if a neural correlate of v was to be identified, it should be interpreted with caution because it might simply represent an ancestor or offspring of v rather than v itself (*Pearl et al., 2016*). This might explain a recent report (*Katz et al., 2016*) of no behavioral consequences of lesioning of an area that has been presumed to be a locus of v for several decades.

Classification versus categorization

The difference between classification and categorization tasks has rarely been appreciated by previous studies, presumably because the two tasks appear similar on a surface level in that they both require translating continuous variables into discrete variables. As footnoted in Introduction, however, classification (e.g., “Is your cat big?”) should not be confused with categorization (“Is that a cat or a dog?”) since they fundamentally differ in what to be computed (*Jacob, 2004*). In terms of our apple-sorting scenario, categorization would be to answer the question of ‘which farm is this apple from?’. Specifically, while classification requires judging the inequality between an instance quantity (‘size of your cat’) and a criterion (‘typical size of cats’), categorization requires judging which category (‘cat’ or ‘dog’) an instance (‘that particular animal’) belongs to (*Green and Swets, 1966*). In this regard, inference of c is necessary for classification but not for categorization whereas a generative model of categories (i.e., how instances are generated from respective categories) is necessary for categorization but not for classification. Nevertheless, previous computational studies on categorization (*Norton et al., 2017; Qamar et al., 2013*) explained human categorization behavior based on generative models of categories, which is necessary for categorization, but also described the model behavior using the concept of criterion, which is unnecessary for categorization. This prompted us to consider the possibility that people could have performed our classification task using the model of categorization despite the suboptimality of the strategy (*Ma, 2012*). However, this possibility was negated because the categorization model could not provide a coherent account for the history effects in our classification task (*Figure 8-figure supplement 2*), which supports that c inference is crucial for performing our task.

History effects in classification via dynamic criterion update

Stimulus history effects in perception (‘history effects’ in short) refer to various phenomena in which past stimuli affect perception of current stimuli, and have recently received considerable attention. Our c-inference model offers a normative account for the history effect in a classification task: c, which is a Bayes-optimal probabilistic integration of working-memory representations of past stimuli, is attracted toward past stimuli, and this leads to decisions that are repulsed from recent stimuli. This history effect, and our account for it, is distinct from those in previous studies as follows.

First, the attractive history effects reported in estimation tasks, whereby perceptual estimates of current stimuli tend to be attracted toward past stimuli (*Bliss, Sun, & D'Esposito, 2017; Cicchini, Anobile, and Burr, 2014; Fritsche, Mostert, and de Lange, 2017; Fischer and Whitney, 2014; Liberman, Fischer, and Whitney, 2014*), may appear conflicting with the repulsive history effect found in the current paper. However, these seemingly conflicting results might be grounded in a common process, whereby a belief over stimulus values is dynamically updated toward recent stimuli but manifest its impact in the opposite directions due to task difference. In estimation tasks, prior expectations (e.g., 'what angle of orientation is expected to appear on an upcoming trial?') will tend to attract the posterior belief of a current stimulus toward past stimuli, resulting in attractive history effects. In classification tasks, c (e.g., 'what angle of orientation is likely to be an appropriate criterion for upcoming classification?'), though being also attracted to past stimuli, will lead to decisions that are repulsed from recent stimuli because decisions are made from relative comparison of a current stimulus against c. Checking this possibility empirically (e.g., demonstrating that stimulus size estimation and classification are attracted to and repulsed from, respectively, recent stimuli in a single experiment) will be an interesting extension of the current work and will bridge 'estimation' and 'classification' tasks via a common process of dynamic stimulus prior update.

Second, Glaze and his colleagues (*Glaze et al., 2015; Glaze et al., 2018*) have recently proposed a normative formalism for history effects in perceptual choice. Their formalism is relevant for categorization while ours for classification. As mentioned earlier, what needs to be computed differs between the two tasks, and this difference leads to the difference in what information needs to be gleaned from past trials. The observers in Glaze et al. need to update 'category priors' – a pair of prior probabilities of task-relevant categories (e.g., the probabilities of male and female in a gender categorization task; (*Glaze et al., 2015; Gold and Shadlen, 2007; Green and Swets, 1966*)) whereas those in our experiment need to update c, which is a median value of the single distribution of task-relevant stimulus features (e.g., a typical height of people in a height classification task; *Petzschner et al., 2015*).

Third, unlike many previous studies on history effects (*Burr and Cicchini, 2014; Dyjas et al., 2012; Fritsche et al., 2017; Glaze et al., 2015; Raviv et al., 2012; Treisman and Williams, 1984*), our account does not require any premises regarding environmental volatility at all. Our formalism for c inference is grounded in the premise that working memory precision decays as trials elapse (*Bays, 2015; Gorgoraptis et al., 2011; Ma et al., 2014; Zokaei et al., 2015*), which makes c dynamically attracted toward recent stimuli. For this reason, although there was no autocorrelation between stimuli (thus, no information regarding environmental volitility), history effects were still observed in our experiment via dynamic criterion update. For that matter, our normative formalism of probabilistic inference of c based on precision decay of working memory should also be distinguished from other accounts in which working memory decay was only formalized in descriptive or mechanistic manners (*Norton et al., 2017; Treisman and Williams, 1984; Wickelgren and Norman, 1966*).

Future extensions of the formalism for criterion inference

We think that internal criterion plays crucial roles not just in perceptual classification but also in other important cognitive phenomena such as 'perceptual disability (*Lau, 2007*)', 'metacognition of decision confidence (*Kepecs et al., 2008; Sanders et al., 2016*)', and 'attentional modulation (*Luo and Maunsell, 2018*)'. Previous stuides have so far been treating internal criterion as a constant in accountinng for such phenomena. To check whether the 'inferred criterion' can offer better accounts for behavior and neural activity in such phenomena will be important steps toward establishing the generalizability of our new formalism proposed here. Another important direction of extending our formalism of criterion inference will be to incorporate other crucial variables that are known to contribute to the behavioral or neural variabilities in perceptual decision-making such as feedback from the environment (*Dobres and Watanabe, 2012; Herzog and Fahle, 1999*). For instance, although we demonstrated the robust stimulus history effects on inferred criterion in the presence of various types of feedback, we conjecture that feedback itself is another important factor for criterion inference and are currently extending our formalism such that a Bayesian decision-maker probabilistically infers internal criterion by taking into account both past stimuli and feedback in an integrative manner (*Lee et al., 2019*).

Methods

The main data were acquired from 18 participants (9 females, aged 20–30 years), who performed the task inside an MRI scanner. The sample size was determined based on previous SVR studies on fMRI data (*Kahnt et al., 2011a; Kahnt et al., 2011b*). The auxiliary data used in some of supplementary figures (*Figure 4-figure supplement 4,5*; *Figure 5-figure supplement 1*; *Figure 6-figure supplement 2*) were borrowed from previously published or in-preparation papers in our lab (*Choe et al., 2014, 2016; Lee et al., 2016*). The main data were never published nor used in any previous work, and the origins of the data are specified in *Figure 4-figure supplement 4A* to avoid confusions. 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

Figure 2 illustrates the behavioral task. While fixating at the screen center, observers were instructed to view a brief (0.3 s) ring-shape stimulus and classify its size within 1.2 s after stimulus onset into either 'small' or 'large' by pressing the left-hand and right-hand keys, respectively. The response (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 criterion (radius of 9°) to help them understand the goal of the task (, which is 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 2.7 s. 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 13.2 s. 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 fMRI runs, observers inside the MRI scanner repeated what they did during the training sessions by performing 180 threshold-calibration trials and 54 practice trials with trial-to-trial feedback with short inter-trial interval (2.7s). 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 performance levels in the main session (proportion correct = 75±7%) were similar to those targeted by the staircase method (70.7%).

Imaging data acquisition and preprocessing

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

For each observer, we acquired three types of MRI images, as follows: (1) 3D, T1-weighted, whole-brain images (MPRAGE; resolution, 1×1×1 mm; field of view (FOV), 256 mm; repetition time (TR), 1.9 s; time for inversion, 700 ms; time to echo (TE), 2.36; and flip angle (FA), 9°), (2) 2D, T1-weighted, in-plane images (MPRAGE; resolution, 1.08×1.08×3.3 mm; TR, 1.5 s; T1, 700 ms; TE, 2.79 ms; and FA, 9°), and (3) 2D, T2*-weighted, functional images (gradient EPI; TR, 2.2 s; TE, 40 ms; FA, 73°; FOV, 208 mm; image matrix, 90×90; slice thickness, 3 mm with 10% space gap; slice, 32 oblique transfers slices; bandwidth, 790 Hz/ps; and effective voxel size, 3.25×3.25×3.3 mm). For univariate analysis, the images of individual observers were normalized to the MNI template using the following steps: motion correction, coregistration to whole-brain anatomical images via the in-plane images (*Nestares and Heeger, 2000*), spike elimination, slice timing correction, normalization using the SPM DARTEL Toolbox (*Ashburner, 2007*) to 3×3×3 mm voxel size, and smoothing with 8×8×8 mm full-width half-maximum Gaussian kernel. All the procedures were implemented with SPM8 and SPM12 (http://www.fil.ion.ucl.ac.uk.spm) (*Friston et al., 1996; Jenkinson et al., 2002*), except for spike elimination for which we used the AFNI toolbox (*Cox, 1996*). The first six frames of each functional scan (the first trial of each run) were discarded to allow hemodynamic responses to reach a steady state. Then, the normalized BOLD time series at each voxel, each run, and each subject were preprocessed using linear detrending and high-pass filtering (132 s cut-off frequency with a Butterworth filter), conversion into percent-change signals, and correction for non-neural nuisance signals by regressing out mean BOLD activity of cerebrospinal fluid (CSF). To define anatomical masks, probability tissue maps for individual participants were generated from T1-weighted images, normalized to the MNI space, and smoothed as were done 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.

Generative model

Figure 3A graphically illustrates the 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 i 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 three following probabilistic terms, as follows: 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 (*Figure 3B-D*), 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 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 $\kappa > 0$. In summary, the generative model is completely specified by the four parameters, $\Theta = \{\mu_0, \sigma_0, \sigma_m, \kappa\}$.

Bayesian estimates of stimuli

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 $m_{(t)}$ (*Equation 1*). The posterior $p\big(Z\big|m_{(t)}\big)$ 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 (*Figure 3c*):

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

$$\sigma_{s_{(t)}} = \sigma_s = (\sigma_0\sigma_m)/\sqrt{{\sigma_0}^2 + {\sigma_m}^2}.$$

Whenever Z takes one of the ring sizes $\{-1,0,1\}$ on each trial as $Z_{(t)}$ (*Figure 3E*), 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 (*Figure 5A*):

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

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

Bayesian estimates of criterion

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 $\mathbb{r}_{(t)} = \{r_{(t-1),} r_{(t-2),} \ldots, r_{(t-n)}\}$ (*Equation 2*), where the maximum number of measurements that can be retrieved, n, was set to 7. Here, $p(Z|\mathbb{r}_{(t)})$ is a conjugate normal distribution of the prior and likelihoods of Z given the evidence $\mathbb{r}_{(t)}$,

$$p(Z|\mathbb{r}_{(t)}) \propto p(\mathbb{r}_{(t)}|Z)p(Z)$$
$$= p(r_{(t-1)}|Z)p(r_{(t-2)}|Z) \ldots p(r_{(t-7)}|Z)p(Z)$$

, whose mean and standard deviation were calculated (*Bromiley, 2003*) based on the knowledge of how the retrieved stimulus becomes noisier as trials elapse (*Figure 3B*):

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

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

, 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 (*Figure 5A*):

$$\mu_{p(c_{(t)})} = \beta_0 \mu_0 + \sum_{i=1}^{7} \beta_i Z_{(t)};$$

$$\sigma_{p(c_{(t)})} = \sqrt{\sum_{i=1}^{7} \beta_i{}^2 \sigma_{r_{(t-i)}}{}^2}.$$

The constant-criterion model

The constant-criterion model has two parameters, bias 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 value, μ_0; thus $\sigma_{p(c_{(t)})} = \sigma_{c_{(t)}} = 0$.

Converting Bayesian estimates into choice fraction, decision uncertainty, and decision variable

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 $v_{(t)} = \Phi[\frac{s_{(t)} - c_{(t)}}{\sqrt{\sigma^2_s + \sigma^2_{c_{(t)}}}}]$, and the decision uncertainty, $u_{(t)}$, which represents the odds that the current decision will be incorrect, as follows: $u_{(t)} = \Phi[\frac{-|s_{(t)} - c_{(t)}|}{\sqrt{\sigma^2_s + \sigma^2_{c_{(t)}}}}]$ (*Figure 3E,F*).

Estimating generative model parameters

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

$$\widehat{\Theta}_j = \arg\max_{\Theta} \sum_{t=1}^{N} \log\, p(D_{j(t)}|\Theta)$$

, where $\mathrm{p}\left(\mathrm{D}_{\mathrm{j(t)}}\middle|\Theta\right) = \mathrm{pL}_{(\mathrm{t})}$ for $\mathrm{D}_{\mathrm{j(t)}}$ =‘large’ and $\mathrm{p}\left(\mathrm{D}_{\mathrm{j(t)}}\middle|\Theta\right) = 1 - \mathrm{pL}_{(\mathrm{t})}$ for $\mathrm{D}_{\mathrm{j(t)}}$ =‘small’. See ***Figure 3-figure supplement 1*** for the detailed procedure and results of parameter estimation.

Comparing Bayesian and human decision behavior

Estimating the generative model parameters separately for the 18 human observers 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 stimulus orders that were identical to those presented to their human counterparts, we acquired a sufficient number (10^6 repetitions) of simulated choices, $\mathrm{d}_{(\mathrm{t})}$, decision uncertainty values, $\mathrm{u}_{(\mathrm{t})}$, and stimulus estimate, $\mathrm{s}_{(\mathrm{t})}$, and criterion estimate, $\mathrm{c}_{(\mathrm{t})}$, based on pairs of random samples drawn from the stimulus and criterion estimate distributions (which are specified by $\mu_{\mathrm{p}(\mathrm{s}_{(\mathrm{t})})}$with $\sigma_{\mathrm{p}(\mathrm{s}_{(\mathrm{t})})}$, and $\mu_{\mathrm{p}(\mathrm{c}_{(\mathrm{t})})}$ with $\sigma_{\mathrm{p}(\mathrm{c}_{(\mathrm{t})})}$, respectively) on each trial for each observer. Then, those simulated outcomes were each averaged across the 10^6 simulations. When predicting $\mathrm{s}_{(\mathrm{t})}$, $\mathrm{c}_{(\mathrm{t})}$, $\mathrm{v}_{(\mathrm{t})}$, and $\mathrm{u}_{(\mathrm{t})}$ for the observed choice $\mathrm{D}_{(\mathrm{t})}$, we only included the simulation outcomes in which the relation between $\mathrm{s}_{(\mathrm{t})}$ and $\mathrm{c}_{(\mathrm{t})}$ (i.e., $\mathrm{d}_{(\mathrm{t})}$) matched the observed choice $\mathrm{D}_{(\mathrm{t})}$.

To check the correspondence between the Bayesian and human observers in choice fraction, we first sorted trials based on current stimuli, which resulted in three different stimulus size conditions. Then, for each condition, we further sorted trials into normalized bins of $\mu_{\mathrm{p}(\mathrm{c}_{(\mathrm{t})})}$ (***Figure 5D***) or $\mu_{\mathrm{p}(\mathrm{s}_{(\mathrm{t})})} - \mu_{\mathrm{p}(\mathrm{c}_{(\mathrm{t})})}$ (***Figure 6D***), and pitted Bayesian $\mathrm{pL}_{(\mathrm{t})}$ against human $\mathrm{pL}_{(\mathrm{t})}$ over the different stimulus conditions. To check the Bayesian-human correspondence in the impact of previous stimuli and previous choices on current choices, we logistically regressed the human choice sequence onto stimuli and choices, both on current and previous trials concurrently, using the following model to obtain regression coefficients $\vec{\mathrm{p}} = [\mathrm{p}_1, \cdots, \mathrm{p}_{11}]$ for each observer (***Figure 4C***): $\vec{\mathrm{D}}_{(\mathrm{t})} \sim \mathrm{e}^{\vec{\mathrm{K}}_{(\mathrm{t})}}/(1 + \mathrm{e}^{\vec{\mathrm{K}}_{(\mathrm{t})}})$, where $\vec{\mathrm{K}}_{(\mathrm{t})} = \mathrm{p}_0 + \mathrm{p}_1\vec{\mathrm{Z}}_{(\mathrm{t})} + \sum_{\mathrm{i}=1}^{5}(\mathrm{p}_{1+\mathrm{i}}\vec{\mathrm{Z}}_{(\mathrm{t-i})} + \mathrm{p}_{6+\mathrm{i}}\vec{\mathrm{D}}_{(\mathrm{t-i})})$, the independent variables were each standardized into z-scores for each observer and $\vec{\mathrm{D}}_{(\mathrm{t})} = \left[\mathrm{D}_{(1)}, \cdots \mathrm{D}_{(\mathrm{N})}\right]$, $\vec{\mathrm{K}}_{(\mathrm{t})} = \left[\mathrm{K}_{(1)}, \cdots \mathrm{K}_{(\mathrm{N})}\right]$, $\vec{\mathrm{Z}}_{(\mathrm{t})} = \left[\mathrm{Z}_{(1)}, \cdots \mathrm{Z}_{(\mathrm{N})}\right]$. The Bayesian observers’ choices were also regressed with the logistic regression model by substituting $\vec{\mathrm{d}}_{(\mathrm{t})}$ for $\vec{\mathrm{D}}_{(\mathrm{t})}$ (***Figure 4A***) where $\vec{\mathrm{d}}_{(\mathrm{t})} = \left[\mathrm{d}_{(1)}, \cdots \mathrm{d}_{(\mathrm{N})}\right]$. The regression was repeatedly carried out for each simulation, and the estimated coefficients $\vec{\hat{\mathrm{p}}}$ were averaged across the simulations. Due to the time consumption of regression, the number of simulations for regression (10^5 repetitions) was compromised, and was smaller than that for $\mathrm{pL}_{(\mathrm{t})}$ prediction (10^6). However, we confirmed that the simulation number was sufficiently large to produce stable simulation outcomes.

To check the correspondence between Bayesian decision uncertainty $\mathrm{u}_{(\mathrm{t})}$ and human RT, we took both stimuli and choices into account, because $\mathrm{u}_{(\mathrm{t})}$ under the same stimulus condition is contingent on the choice made in a given trial (***Figure 5C***). Thus, we sorted trials into six different conditions (three stimulus sizes × two choices) and then further sorted the trials in each condition into normalized bins of $\mu_{\mathrm{p}(\mathrm{c}_{(\mathrm{t})})}$ (***Figure 5E***), where the normalized values of $\mathrm{u}_{(\mathrm{t})}$ and human RT were compared. To convert $\mathrm{u}_{(\mathrm{t})}$ values into the values comparable to RT, we regressed RT onto $\mathrm{u}_{(\mathrm{t})}$ using a generalized linear mixed model (GLMM), $\overrightarrow{\mathrm{RT}}_{(\mathrm{t})} \sim \beta_0 + \beta_1\vec{\mathrm{u}}_{(\mathrm{t})}$, with a random effect of individual observers, where $\overrightarrow{\mathrm{RT}}_{(\mathrm{t})} = \left[\mathrm{RT}_{(1)}, \cdots \mathrm{RT}_{(\mathrm{N})}\right]$ and $\vec{\mathrm{u}}_{(\mathrm{t})} = \left[\mathrm{u}_{(1)}, \cdots \mathrm{u}_{(\mathrm{N})}\right]$. Then, ‘$\beta_0 + \beta_1\vec{\mathrm{u}}_{(\mathrm{t})}$’ were pitted against human RT data. To check the Bayesian-human correspondence in the impact of previous stimuli on decision uncertainty, we regressed the Bayesian decision uncertainty estimates $\vec{\mathrm{u}}_{(\mathrm{t})}$ or human RTs $\overrightarrow{\mathrm{RT}}_{(\mathrm{t})}$ on current trials onto the congruency of current and past stimuli with a current choice using the following model to obtain regression coefficients by GLMM with a random effect of individual observers (***Figure 4D***):

$$\overrightarrow{\mathrm{RT}}_{(\mathrm{t})} \sim \mathrm{q}_0 + \textstyle\sum_{\mathrm{i}=0}^{5} \mathrm{q}_{\mathrm{i}+1}\mathrm{g}_{\mathrm{Z(t-i)}} + \sum_{\mathrm{i}=1}^{5} \mathrm{q}_{\mathrm{i}+6}\mathrm{g}_{\mathrm{D(t-i)}}, \text{ where}$$
$$\mathrm{g}_{\mathrm{Z(t-i)}} = 1 \text{ if } \mathrm{Z}_{(\mathrm{t-i})} = \mathrm{D}_{(\mathrm{t})} \text{ and } \mathrm{Z}_{(\mathrm{t-i})} \neq \mathrm{'M'};$$
$$\mathrm{g}_{\mathrm{Z(t-i)}} = -1 \text{ if } \mathrm{Z}_{(\mathrm{t-i})} \neq \mathrm{D}_{(\mathrm{t})} \text{ and } \mathrm{Z}_{(\mathrm{t-i})} \neq \mathrm{'M'};$$
$$\mathrm{g}_{\mathrm{Z(t-i)}} = 0 \text{ if } \mathrm{Z}_{(\mathrm{t-i})} = \mathrm{'M'};$$
$$\mathrm{g}_{\mathrm{D(t-i)}} = 1 \text{ if } \mathrm{D}_{(\mathrm{t-i})} = \mathrm{D}_{(\mathrm{t})};$$
$$\mathrm{g}_{\mathrm{D(t-i)}} = -1 \text{ if } \mathrm{D}_{(\mathrm{t-i})} \neq \mathrm{D}_{(\mathrm{t})}.$$

For the regression analysis, the independent and dependent variables were both standardized into z-scores for each observer. For comparison, the regression coefficients for Bayesian decision uncertainty were calculated as $\vec{\mathrm{u}}_{(\mathrm{t})} \sim \hat{\mathrm{q}}_0 + \sum_{\mathrm{i}=0}^{5} \hat{\mathrm{q}}_{\mathrm{i}+1}\mathrm{g}_{\mathrm{Z(t-i)}} + \sum_{\mathrm{i}=1}^{5} \hat{\mathrm{q}}_{\mathrm{i}+6}\mathrm{g}_{\mathrm{D(t-i)}}$ by GLMM with random effect of individual observers (*Figure 4B*), and linearly re-scaled to those for human RT using the following GLM: $\sum_{i=0}^{11} q_i \sim \gamma(\sum_{i=0}^{11} \hat{q}_i)$.

Comparing Bayesian decision uncertainty and human BOLD activity

At each of the six time points, i, within a single trial, t, BOLD responses, $B_{(v,t,i)}$, of each cortical voxel, v, were regressed onto $u_{(t)}$ to identify brain regions whose activity was correlated with $u_{(t)}$ 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 $\vec{B}_{(v,t,i)} = [B_{(v,1,i)}, \cdots, B_{(v,N,i)}]$. For the regression analysis, the independent and dependent variables were both standardized into z-scores for each observer. We first localized 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, we identified voxel clusters covering 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} as regions of interest (ROIs). For ROI analysis, BOLD signals were averaged over individual voxels, and their correspondences with Bayesian decision uncertainty were calculated using the same procedure as that used for RT data analysis.

Searching for multivoxel patterns of activity signaling latent model variables

To decode the model's latent variables from the BOLD signal, 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 BOLD signals of a voxel cluster centered at a gray-matter voxel, e, where $k \in \{1,2,\dots\}$ 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 in the dACC. Although the searchlight cluster had a radius of 9 mm 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 model's 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). We note that one might expect the fine-retinotopy-based population decoding method developed in our previous imaging work (*Choe et al., 2014*) to be applied to decode $s_{(t)}$ in the early visual areas. We chose not to do so for several reasons. First, while the population decoding method in Choe et al. (2014) was developed to decode external stimuli by assigning decoding weights to invidual voxels based on the retinotopy (eccentricity) map acquired a priori, the latent variable $s_{(t)}$ does not represent external physical stimuli per se but inferred stimuli that partake with $c_{(t)}$ in causing $v_{(t)}$. For such a latent variable, whose values stochastically vary on a trial-to-trial basis, the time-resolved SVR method is more appropriate because voxel weights can flexibly learned via training. Second, we, of course, did not want to preclude the possibility that $s_{(t)}$ is represented in the early visual cortex or to limit our search of brain signals of $s_{(t)}$ within the early visual cortex either. In this regard, the SVR method in conjunction with searchlight is more appropriate than the retinotopy-based population decoding method, because the former can be applied impartially to any local regions throughout the entire brain whereas the latter can be applied only to those with fine retinotopy maps. Lastly, the spatial resolution of fMRI signals in the current study (voxel size, 3.25x3.25x3.3mm) is not appropriate for the fine-grained-retinotopy-based population decoding method (e.g., voxel size of 2.0x2.0x1.998mm (retinotopy scans) and 2.3x2.3x2.3mm (experimental scans) were used in Choe et al. (2014)). We performed the SVR using the LIBSVM (http://www.csie.ntu.edu.tx/~sjlin/libsvm) with a linear kernel and constant regularization parameter as 1. $B_{(e,k,t,i)}$ and the latent variables were z-scored on each voxel and for each subject before decoding. To calculate the significance level of the decoded information, each 3D map that represented the 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 FWHM Gaussian kernel. We regressed the smoothed $v_{B(e,t,i)}$, $c_{B(e,t,i)}$, and $s_{B(e,t,i)}$ to $v_{(t)}$, $c_{(t)}$, and $s_{(t)}$, respectively, using 16, 13, and 13 regression models, respectively. We concluded that a given cluster carries the neural signals of $v_{(t)}$, $c_{(t)}$, or $s_{(t)}$, only when all those regression models are satisfied over 12 contiguous searchlights. Those regression models were predicted by the causal relationships between the model latent variables (*Figure 7-figure supplement 1*). For the ROI analysis (*Figure 7C-G*), the decoded values of a given latent variable were averaged over the all searchlights within each ROI. For the data-driven Bayesian network analysis (*Figure 7G; Figure 7-figure supplement 4*), we derived an exhaustive set of causal structures between the brain signals of latent model variables and calculated BIC for each structure (*Scutari, 2009*). The brain imaging results were visualized using xjView toolbox for the cross sectional images and Connectome Workbench (*Marcus et al., 2011*) for the inflated images.

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, 18 individuals were used except the whole brain analysis, in which statistics at some of ventral area were calculated with 17 individuals. Statistical significances were calculated using two-tailed tests, except for the regression models for decoding the model's latent variables form neural information (*Figure 7-figure supplement 1*).

Data and code availability

The codes for reproducing main results are available at https://github.com/Heeseung-Lee/LeeLeeChoeLee2019/tree/master/code. The behavior data, raw MRI data, BOLD activity of dACC, and the decoded latent variables with SVR are available at https://github.com/Heeseung-Lee/LeeLeeChoeLee2019/tree/master/data. The statistical parametric maps of the results of whole brain analysis are available at https://github.com/Heeseung-Lee/LeeLeeChoeLee2019/tree/master/SPM.

Acknowledgement

We thank J.S. Lim and J. Ahn for their advice on data analysis, D.S. Lim for his advice on modeling, H.W. Gu for sharing his tips on Bayesian-network analysis, and I. Rhim for sharing his data. This research was supported by the Brain Research Programs and Basic Research Laboratory Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT of South Korea (No. NRF-2015M3C7A1031969, NRF-2017M3C7A1047860 for the Brain Research Programs; NRF-2018R1A4A1025891 for the Basic Research Laboratory Program).

Additional information

Author contributions

K.W.C. and S.H.L. designed the experiments. K.W.C. ran experiments. H.L. performed the analyses. H.L. and S.H.L. implemented the models and wrote the manuscript. H.J.L. and K.W.C. preprocessed the auxiliary data. K.W.C. provided critical revisions.

REFERENCES

Akrami, A., Kopec, C.D., Diamond, M.E., and Brody, C.D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature 554, 368.

Anderson, D., and Burnham, K. (2004). Model selection and multi-model inference. Second NY: Springer-Verlag, 63.

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

Ashby, F. (2001). Categorization and similarity models: Neuroscience applications.

Ashby, F.G., and Ell, S.W. (2001). The neurobiology of human category learning. Trends in cognitive sciences 5, 204-210.

Baumann, O., Borra, R.J., Bower, J.M., Cullen, K.E., Habas, C., Ivry, R.B., Leggio, M., Mattingley, J.B., Molinari, M., and Moulton, E.A. (2015). Consensus paper: the role of the cerebellum in perceptual processes. The Cerebellum 14, 197-220.

Bays, P.M. (2015). Spikes not slots: noise in neural populations limits working memory. Trends in cognitive sciences 19, 431-438.

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of information in an uncertain world. Nature neuroscience 10, 1214.

Benjamin, A.S., Diaz, M., and Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. Psychological review 116, 84.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological), 289-300.

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

Burr, D., and Cicchini, G.M. (2014). Vision: efficient adaptive coding. Current Biology 24, R1096-R1098.

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

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., and Cohen, J.D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280, 747-749.

Cavanagh, J.F., and Frank, M.J. (2014). Frontal theta as a mechanism for cognitive control. Trends in cognitive sciences 18, 414-421.

Choe, K.W., Blake, R., and Lee, S.-H. (2014). Dissociation between neural signatures of stimulus and choice in population activity of human V1 during perceptual decision-making. Journal of Neuroscience 34, 2725-2743.

Choe, K.W., Blake, R., and Lee, S.-H. (2016). Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. Vision research 118, 48-59.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical research 29, 162-173.

Crapse, T.B., Lau, H., and Basso, M.A. (2018). A role for the superior colliculus in decision criteria. Neuron 97, 181-194. e186.

Curtis, C.E., and Lee, D. (2010). Beyond working memory: the role of persistent activity in decision making. Trends in cognitive sciences 14, 216-222.

Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific and statistical inference (Macmillan International Higher Education).

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in psychology 5, 781.

Dobres, J., and Watanabe, T. (2012). Response feedback triggers long-term consolidation of perceptual learning independently of performance gains. Journal of vision 12, 9-9.

Dyjas, O., Bausenhart, K.M., and Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. Attention, Perception, & Psychophysics 74, 1819-1841.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., and Turner, R. (1996). Movement-related effects in fMRI time-series. Magnetic resonance in medicine 35, 346-355.

Fritsche, M., Mostert, P., and de Lange, F.P. (2017). Opposite effects of recent history on perception and decision. Current Biology 27, 590-595.

Fründ, I., Wichmann, F.A., and Macke, J.H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. Journal of vision 14, 9-9.

Gao, J.-H., Parsons, L.M., Bower, J.M., Xiong, J., Li, J., and Fox, P.T. (1996). Cerebellum implicated in sensory acquisition and discrimination rather than motor control. Science 272, 545-547.

Ghiselin, M.T. (1981). Categories, life, and thinking. Behavioral and Brain Sciences 4, 269-283.

Glaze, C.M., Kable, J.W., and Gold, J.I. (2015). Normative evidence accumulation in unpredictable environments. Elife 4, e08825.

Goard, M.J., Pho, G.N., Woodson, J., and Sur, M. (2016). Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. Elife 5, e13764.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annual review of neuroscience 30.

Good, I.J. (1979). Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in World War II. Biometrika, 393-396.

Gorgoraptis, N., Catalao, R.F., Bays, P.M., and Husain, M. (2011). Dynamic updating of working memory resources for visual objects. Journal of Neuroscience 31, 8502-8511.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics (Oxford, England: John Wiley).

Griffiths, T.L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012).

Haller, M., Case, J., Crone, N.E., Chang, E.F., King-Stephens, D., Laxer, K.D., Weber, P.B., Parvizi, J., Knight, R.T., and Shestyuk, A.Y. (2018). Persistent neuronal activity in human prefrontal cortex links perception and action. Nature Human Behaviour 2, 80.

Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature 520, 220.

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

Heekeren, H.R., Marrett, S., Ruff, D.A., Bandettini, P., and Ungerleider, L.G. (2006). Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. Proceedings of the National Academy of Sciences 103, 10023-10028.

Hempel, C.G. (1965). Aspects of scientific explanation.

Herzog, M.H., and Fahle, M. (1999). Effects of biased feedback on learning and deciding in a vernier discrimination task. Vision research 39, 4232-4243.

Holroyd, C.B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R.B., Coles, M.G., and Cohen, J.D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. Nature neuroscience 7, 497.

Hwang, E.J., Dahlen, J.E., Mukundan, M., and Komiyama, T. (2017). History-based action selection bias in posterior parietal cortex. Nature communications 8, 1242.

Jacob, E.K. (2004). Classification and categorization: a difference that makes a difference.

Jasper, A.I., Tanabe, S., and Kohn, A. (2019). Predicting perceptual decisions using visual cortical population responses and choice history. Journal of Neuroscience, 0035-0019.

Jeffreys, H. (1961). Theory of probability, Clarendon. (Oxford).

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

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

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

Karnath, H.-O. (2001). New insights into the functions of the superior temporal cortex. Nature Reviews Neuroscience 2, 568.

Karnath, H.-O., Ferber, S., and Himmelbach, M. (2001). Spatial awareness is a function of the temporal not the posterior parietal lobe. Nature 411, 950.

Karnath, H.-O., Milner, A.D., and Vallar, G. (2002). The cognitive and neural bases of spatial neglect (Oxford University Press Oxford:).

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. Journal of the american statistical association 90, 773-795.

Katz, L.N., Yates, J.L., Pillow, J.W., and Huk, A.C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature 535, 285.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature 455, 227.

Kim, J.-N., and Shadlen, M.N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. Nature neuroscience 2, 176.

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. Nature 427, 244.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences 103, 3863-3868.

Lau, H.C. (2007). A higher order Bayesian decision theory of consciousness. Progress in brain research 168, 35-48.

Lee, H.-J., Lee, H., Rhim, I., Lim, C.Y., and Lee, S.-H. (2019). Learning criteria from feedback in perceptual classification.

Lee, H.-J., Rhim, I., and Lee, S.-H. (2016). Optimal and suboptimal integration of sensory and value information in perceptual decision-making. CoSyNe.

Lee, S.-H., Blake, R., and Heeger, D.J. (2007). Hierarchy of cortical responses underlying binocular rivalry. Nature neuroscience 10, 1048.

Li, S., Mayhew, S.D., and Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. Neuron 62, 441-452.

Li, W., Piëch, V., and Gilbert, C.D. (2004). Perceptual learning and top-down influences in primary visual cortex. Nature neuroscience 7, 651.

Luo, T.Z., and Maunsell, J.H. (2018). Attentional changes in either criterion or sensitivity are associated with robust modulations in lateral prefrontal cortex. Neuron 97, 1382-1393. e1387.

Ma, W.J. (2012). Organizing probabilistic models of perception. Trends in cognitive sciences 16, 511-518.

Ma, W.J., Husain, M., and Bays, P.M. (2014). Changing concepts of working memory. Nature neuroscience 17, 347.

Marco-Pallarés, J., Müller, S.V., and Münte, T.F. (2007). Learning by doing: an fMRI study of feedback-related brain activations. Neuroreport 18, 1423-1426.

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

Marr, D., and Poggio, T. (1976). Cooperative computation of stereo disparity. Science 194, 283-287.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. Weather and Forecasting 19, 1106-1114.

Meyniel, F., Sigman, M., and Mainen, Z.F. (2015). Confidence as Bayesian probability: From neural origins to behavior. Neuron 88, 78-92.

Nakashima, Y., and Sugita, Y. (2017). The reference frame of the tilt aftereffect measured by differential Pavlovian conditioning. Scientific reports 7, 40525.

Nestares, O., and Heeger, D.J. (2000). Robust multiresolution alignment of MRI brain volumes. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 43, 705-715.

Nienborg, H., and Cumming, B.G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459, 89.

Norton, E.H., Fleming, S.M., Daw, N.D., and Landy, M.S. (2017). Suboptimal criterion learning in static and dynamic environments. PLoS computational biology 13, e1005304.

Palmer, J., Huk, A.C., and Shadlen, M.N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. Journal of vision 5, 1-1.

Pavan, A., Marotti, R.B., and Campana, G. (2012). The temporal course of recovery from brief (sub-second) adaptations to spatial contrast. Vision research 62, 116-124.

Pearl, J., Glymour, M., and Jewell, N.P. (2016). Causal inference in statistics: a primer (John Wiley & Sons).

Petzschner, F.H., Glasauer, S., and Stephan, K.E. (2015). A Bayesian perspective on magnitude estimation. Trends in cognitive sciences 19, 285-293.

Pouget, A., Beck, J.M., Ma, W.J., and Latham, P.E. (2013). Probabilistic brains: knowns and unknowns. Nature neuroscience 16, 1170.

Pourtois, G., Rauss, K.S., Vuilleumier, P., and Schwartz, S. (2008). Effects of perceptual learning on primary visual cortex activity in humans. Vision research 48, 55-62.

Qamar, A.T., Cotton, R.J., George, R.G., Beck, J.M., Prezhdo, E., Laudano, A., Tolias, A.S., and Ma, W.J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. Proceedings of the National Academy of Sciences 110, 20332-20337.

Rahnev, D., and Denison, R.N. (2018). Suboptimality in perceptual decision making. Behavioral and Brain Sciences 41.

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. Neural computation 20, 873-922.

Raviv, O., Ahissar, M., and Loewenstein, Y. (2012). How recent history affects perception: the normative approach and its heuristic approximation. PLoS computational biology 8, e1002731.

Renart, A., and Machens, C.K. (2014). Variability in neural activity and behavior. Current opinion in neurobiology 25, 211-220.

Richardson, D.M., and Rejmánek, M. (2011). Trees and shrubs as invasive alien species–a global review. Diversity and distributions 17, 788-809.

Rips, L.J., and Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. Cognition 8, 145-174.

Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of neuroscience 22, 9475-9489.

Sanders, J.I., Hangya, B., and Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. Neuron 90, 499-506.

Schwartz, S., Maquet, P., and Frith, C. (2002). Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. Proceedings of the National Academy of Sciences 99, 17137-17142.

Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:09083817.

Seltzer, B., and Pandya, D.N. (1994). Parietal, temporal, and occipita projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. Journal of Comparative Neurology 343, 445-463.

Shenhav, A., Straccia, M.A., Cohen, J.D., and Botvinick, M.M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. Nature neuroscience 17, 1249.

Sheth, S.A., Mian, M.K., Patel, S.R., Asaad, W.F., Williams, Z.M., Dougherty, D.D., Bush, G., and Eskandar, E.N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. Nature 488, 218.

Simon, H.A. (1955). A behavioral model of rational choice. The quarterly journal of economics 69, 99-118.

Treisman, M., and Williams, T.C. (1984). A theory of criterion setting with an application to sequential dependencies. Psychological Review 91, 68.

Urai, A.E., Braun, A., and Donner, T.H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nature communications 8, 14637.

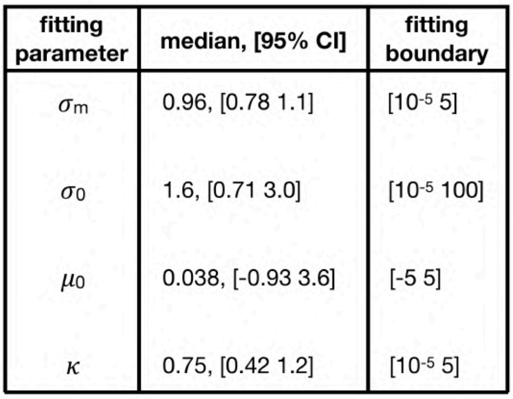
White, C.N., Mumford, J.A., and Poldrack, R.A. (2012). Perceptual criteria in the human brain. Journal of Neuroscience 32, 16716-16724.

Wickelgren, W.A., and Norman, D.A. (1966). Strength models and serial position in short-term recognition memory. Journal of Mathematical Psychology 3, 316-347.

Wod, I. (1985). Weight of evidence: A brief survey. Bayesian statistics, 249-270.

Zariwala, H.A., Kepecs, A., Uchida, N., Hirokawa, J., and Mainen, Z.F. (2013). The limits of deliberation in a perceptual decision task. Neuron 78, 339-351.

Zokaei, N., Burnett Heyes, S., Gorgoraptis, N., Budhdeo, S., and Husain, M. (2015). Working memory recall precision is a more sensitive index than span. Journal of Neuropsychology 9, 319-329.



fitting parameter	median, [95% CI]	fitting boundary
σ_m	0.96, [0.78 1.1]	[10^{-5} 5]
σ_0	1.6, [0.71 3.0]	[10^{-5} 100]
μ_0	0.038, [-0.93 3.6]	[-5 5]
κ	0.75, [0.42 1.2]	[10^{-5} 5]

Figure3-figure supplement 1. Estimation of model parameters and confidence intervals with uniform prior bounds. Across-observer median value, its 95% bootstrap confidence interval, and prior bounds are shown for each estimated model parameter. 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 estimates for the model parameters. The first trial of each run and the trials on which RTs were too short (less than 0.3 s) or during which no responses were given were used neither for parameter estimation nor for any further analyses. We discarded the first trial of each run for two reasons; first, the criterion-inference model cannot estimate criterion on the first trial since there is not yet an existing measurement to be retrieved; second, the first trial is susceptible to non-specific fMRI signals that are irrelevant to the task. RT was shorter than 0.3s (0.0059s) in one trial, and this was too short to be considered as a task-relevant response. Note that some aspects of the procedure are arbitrary (e.g., the fitting boundaries and the number of initial parameter randomization). So, the outcomes of the fitted model parameters are subect to slight changes depending if those aspects are modified. However, we confirmed that such changes in outcomes are small (results were not shown here) and do not affect the main claims of the current stusy. The fitting codes and data are available in GitHub (see Methods: Data and code availability).

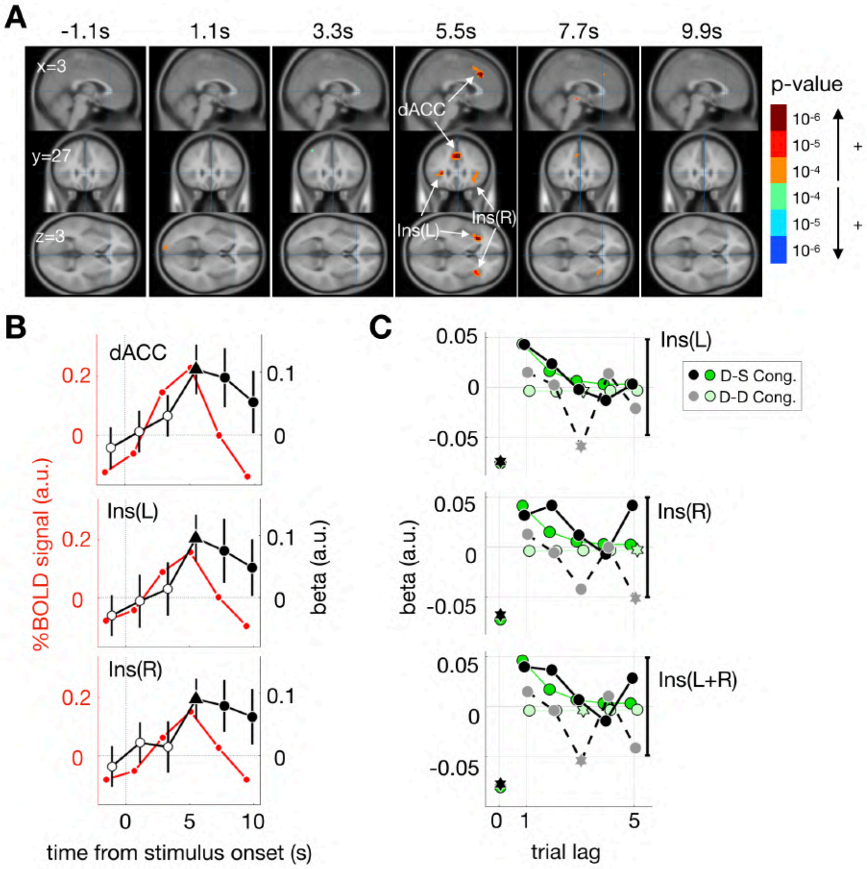
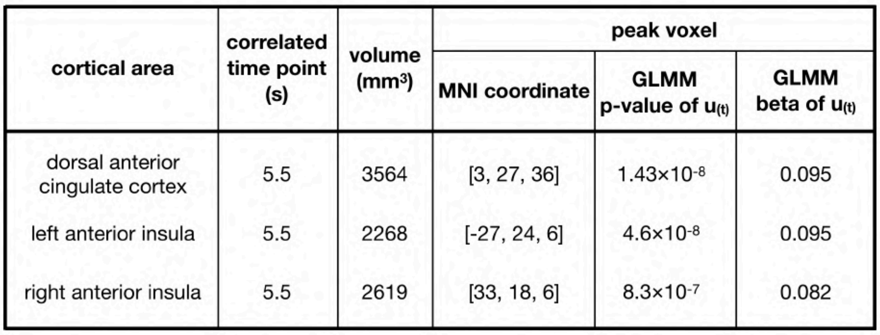


Figure4-figure supplement 1. Univariate BOLD signals correlated with decision uncertainty. **(A)** Regressions of BOLD signals at individual cortical sites (voxels) onto $u_{(t)}$ for each of the 6 time points comprising single trials. Each column shows regression results at each time point, as marked in time elapsed from stimulus onset at the top, by overlaying the thresholded ($P < 10^{-4}$, uncorrected) p-values of regression coefficients on the sagittal, coronal, and axial slices of a template brain. Statistically significant ($P_{FDR} < 0.05$, FDR-corrected), all positive in sign, regressions on $u_{(t)}$ were found in the clustered regions in the dACC and the bilateral insula at 5.5 s after stimulus onset. **(B)** For the three ROIs in the dACC and the insula (see Table S2 for specifications of these ROIs), the time courses of the within-ROI averages of BOLD signals (red) are juxtaposed with their regression coefficients on $u_{(t)}$ (black). Filled circles and filled triangles, coefficients with $P < 0.05$ and $P < 10^{-5}$, respectively. Error bars, standard errors of the mean across observers. **(C)** Results of the multiple linear regressions of BOLD signals onto the congruence of current decisions with stimuli and past decisions are shown for the two ROIs in the insula. The observed regression coefficients (black symbols) are juxtaposed with the simulated regression coefficients (green symbols). Significant coefficients are indicated by hexagons, whereby black hexagons indicate that the 95% CIs of the observed GLMM coefficients did not include zero and green hexagons indicate that the simulated regression coefficients are outside the 95% CIs of the observed GLMM coefficients. Vertical black error bars indicate the average of 95% bootstrap CI for the mean of observed coefficients.



cortical area	correlated time point (s)	volume (mm^3)	peak voxel		
			MNI coordinate	GLMM p-value of $u_{(t)}$	GLMM beta 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

Figure4-figure supplement 2. Specifications of the ROIs in which activity was correlated with decision uncertainty. When a set of constraints (uncorrected $P < 10^{-4}$, FDR-corrected $P < 0.05$, cluster size > 324 mm^3 (or contiguous voxels > 12)) was applied, three regions of interest (ROIs) were significantly correlated with $u_{(t)}$ at the fourth fMRI time point (5.5s after stimulus onset)– the dACC, left insula, and right insula. GLMM was two-tailed test.

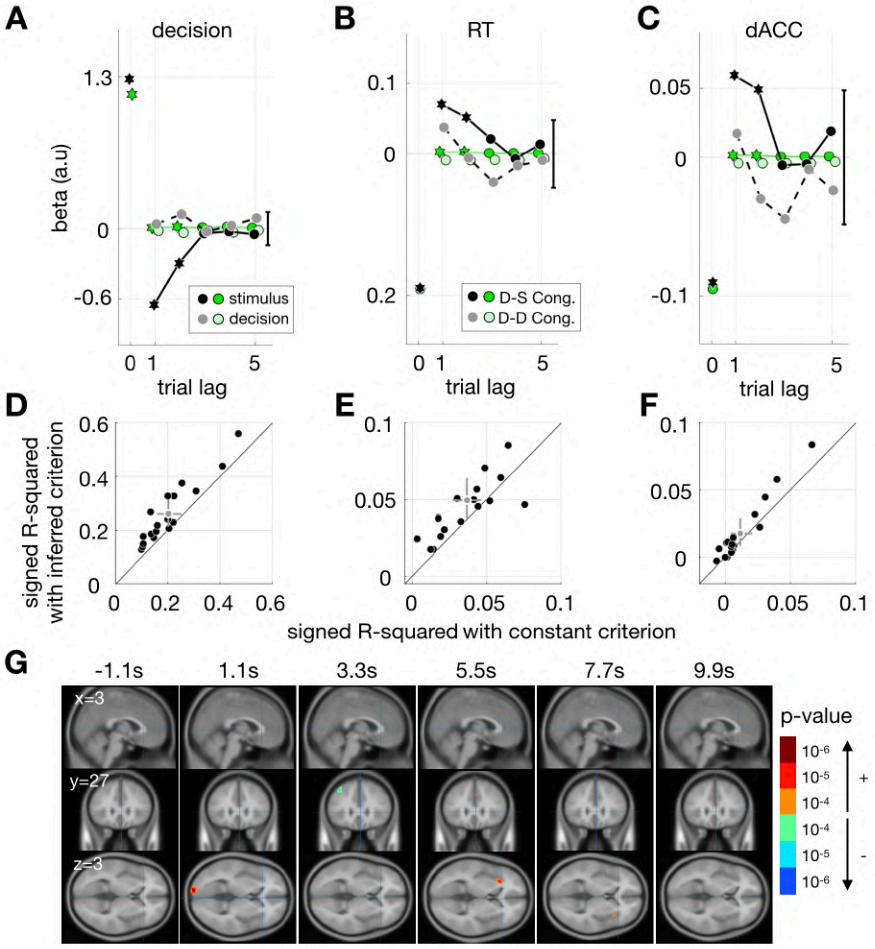
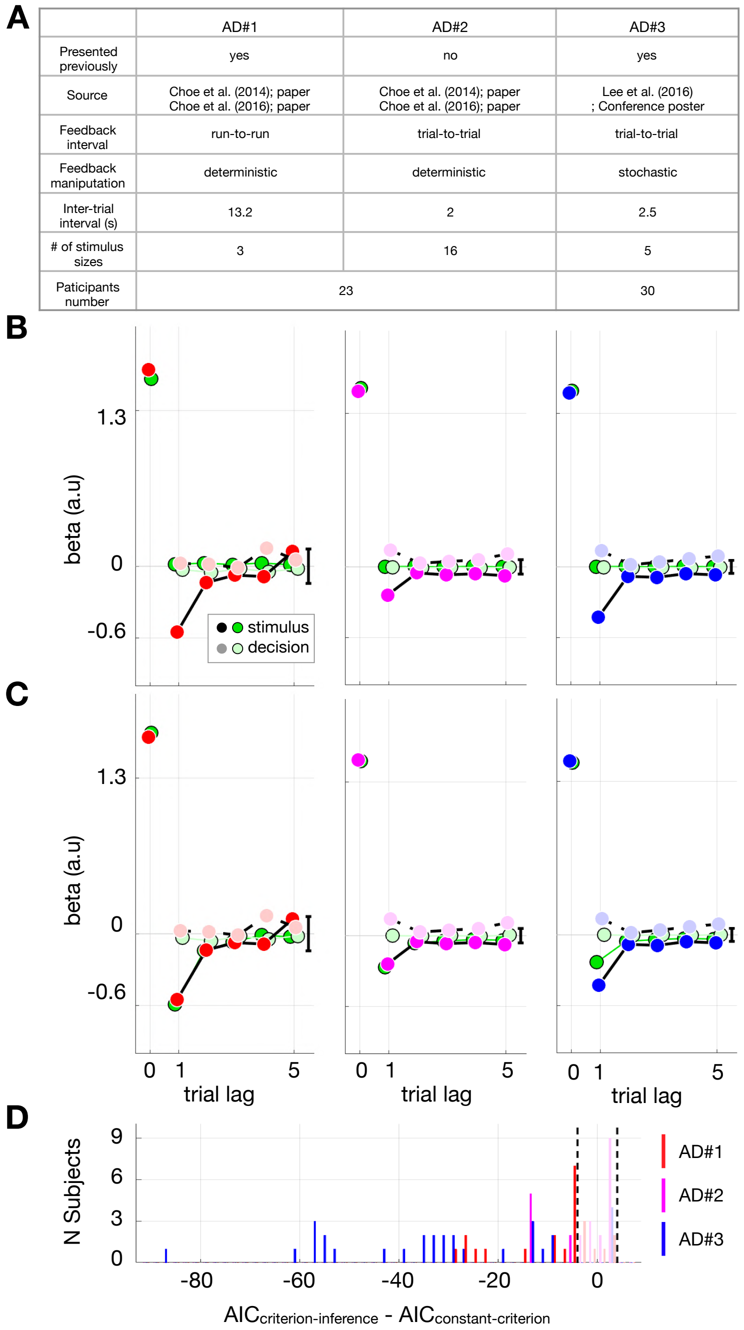


Figure 4-figure supplement 3. Limitations of the constant-criterion model.The constant-criterion model offers no account for stimulus history effects on current decision-making (**A-C**), leaves substantial trial-to-trial variability in decision-making unexplained (**D-F**), and does a poor job of detecting the brain signals associated with decision uncertainty (**G**). (**A**) Multiple logistic regressions of current decisions onto stimuli and past decisions. Black hexagons indicate that the logistic regression coefficients of the observed decision significantly deviated from zero (pairwise t-test P < 0.05), and green hexagons indicate that the model's predictions significantly deviated from the coefficients of observed decision (pairwise t-test P < 0.05). (**B, C**) Multiple regressions of RT and dACC activity onto the congruence of current decisions with stimuli and past decisions. Black hexagons mark the observed GLMM coefficients whose 95% CIs did not include zero, which indicates significant effects of stimuli and stimulus-decision congruences on human observers' decision-making. Green hexagons mark the simulated regression coefficients that are outside the 95% CIs of the observed coefficients, which indicates significant deviations of the simulation from the observation. (A-C) Black and green symbols represent the coefficients for observed and stimulated data, respectively. Vertical black error bars indicate the average of 95% CI for the mean of observed coefficients (A: bootstrap CI; B, C: GLMM CI). Note that the constant-criterion model manages to simulate the effects of current stimuli on current decision-making but, not surprisingly, is incapable of providing any account for the effects of past stimuli on current decision-making, as indicated by the near-zero green regression coefficients. (**D-F**) The proportions of variance explained by the criterion-inference model are pitted against those by the constant-criterion model across individual observers (indicated by black dots), for decision (**D**), RT (**E**), and dACC activity (**F**). Gray dots with cross hairs represent across-observer means with 95% bootstrap CI of the means. As indicated by the placements of black and gray dots above the diagonal identity line, the substantive proportions of variance unexplained by the constant-criterion model could be explained by the criterion-inference model. These increases of the proportions of explained variance, $\frac{\mathrm{Var}_{\mathrm{inferred}}-\mathrm{Var}_{\mathrm{constant}}}{\mathrm{Var}_{\mathrm{constant}}}$, were 29%, 34%, and 49% for decision, RT, and dACC activity, respectively. (**G**) Regressions of BOLD signals at individual cortical sites (voxels) onto $u_{(t)}$ for each of the 6 time points comprising single trials. The formats are identical to those in ***Figure 4-figure supplement 1***, except for the fact that the regressor $u_{(t)}$ was estimated by the constant-criterion model. When compared to the results shown in ***Figure 4-figure supplement 1A*** the cortical sites with significant regressions substantially decreased both in quantity (as indicated by the reduced cluster size) and in quality (as indicated by the high range of p-values), failing to appear in the dACC. This implies that the constant-criterion model is limited in accounting for dACC activity, which has been previously shown to reflect decision uncertainty or task difficulty.



	AD#1	AD#2	AD#3
Presented previously	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	trial-to-trial	trial-to-trial
Feedback maniputation	deterministic	deterministic	stochastic
Inter-trial interval (s)	13.2	2	2.5
# of stimulus sizes	3	16	5
Paticipants number	23		30

Figure 4-figure supplement 4. Verifying the model predictions of the impacts of past stimuli on current decisions with different sets of stimuli, feedback types, and inter-trial-interval lengths. Four auxiliary sets of data were used for this verification.

(A) Specifications of the data sets for their origins and experimental procedures. Auxiliary data set #1, #2, and #3 (AD#1~#3) were borrowed from the works conducted with different purposes in our lab. The specifics of these data sets were as follows: AD#1 was previously published in one of our previous studies (Choe et al., 2014, 2016), which was contributed by 23 observers, each of whom performed 162 trials on the ring stitmuli of 3 different sizes (2.80° ~ 2.88°), with run-to-run and deterministic (see below for definition) feedback and inter-trial interval of 13.2 s. AD#2 is the training sessions' data of AD#1 and has not been published before. It was contributed by the 23 observers participated in AD#1, each of whom performed 315 trials on the ring stimuli of 16 different ring sizes (2.72° ~ 2.95°) presented via a staircase method (1-up-2-down), with trial-to-trial and deterministic feedback and inter-trial interval of 2 s. (see (Choe et al., 2014, 2016) for other methodological details); AD#3 was contributed by 30 observers, each of whom performed 1,700 trials on the ring stimuli of 5 different ring sizes (3.84°, 3.92°, 4.00°, 4.08 °, 4.16°), with trial-to-trial and stochastic (see below for definition) feedback and inter-trial interval of 2.5 s. AD#3 corresponds to the data of our preliminary poster presentation (*Lee et al., 2016*). Deterministic feedback refers to the use of a single fixed objective criterion (AD#1 and AD#2) whereas stochastic feedback refers to the use of objective criteria with a small degree of trial-to-trial variability (AD#3). Stimulus duration was 0.3s for all the data sets. (**B**-**C**) Mutiple logistic regressions of current decisions onto past/current stimuli and past decisions. Green symbols represent the regression coefficients for the simulated data by the constant-criterion model (**B**) or by the criterion-inference model (**C**). Non-green symbols represent the regression coefficients for the auxiliary sets of observed data and are re-plotted in **B** and **C** for model comparisons. Dark colors, coefficients for stimulus regressors; Light colors, coefficients for decision regressors. Importantly, as predicted by the criterion-inference model (dark green symbols in **C**), a previous stimulus with trial lag of 1 showed a strong negative correlation with a current decision in each and every data set (non-green color symbols in **B** and **C**; see *Figure 4* and related parts in the main text for this prediction), which was failed to be captured by the constant-criterion model (dark green symbols in **B**).

(**D**) Comparisons of the criterion-inference model and the constant-criterion model in the accountability for the decisions in the auxiliary data sets. The data accountability was assessed by AIC values. The AIC value differences greater than 4 between nested models, which are demarcated by dashed vertical lines, are conventionally considered to be significant (*Anderson and Burnham, 2004*). When fit to the auxiliary data sets, the criterion-inference model was significantly superior to the constant-criterion model for substantive fractions of observers for all of the auxiliary data sets (70%, 30% and 87% for AD#1, 2 and 3 respectively, bold bars of colors).

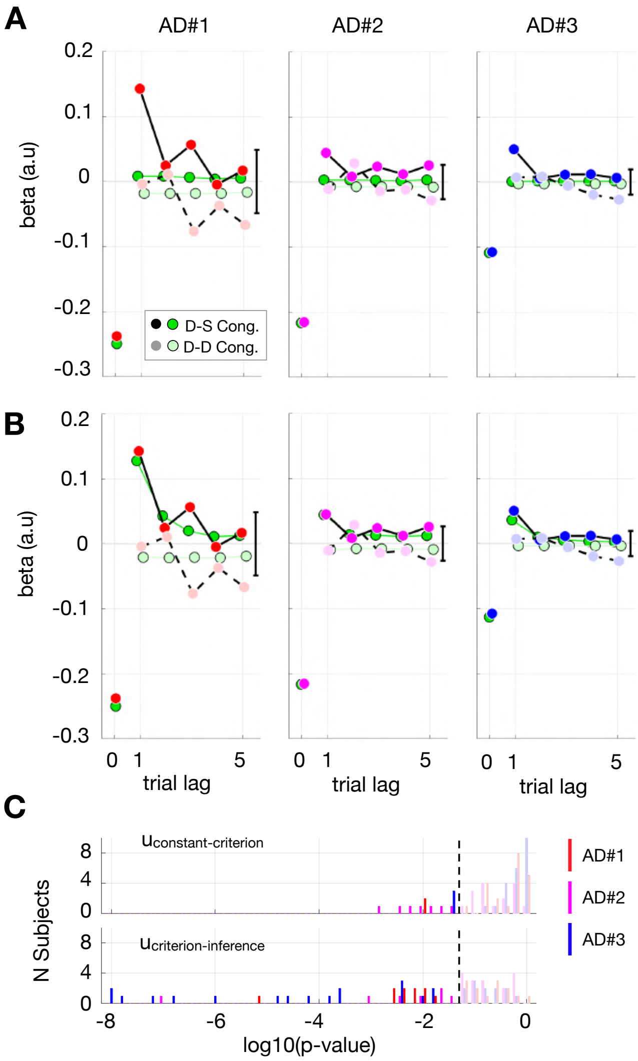


Figure 4-figure supplement 5. Verifying the model predictions of the impacts of past stimuli on current RTs with different sets of stimuli, feedback types, and inter-trial-interval lengths. This verification was based on the same auxiliary sets of data that were described in ***Figure 4-figure supplement 4A*.**

(**A-B**) Mutiple linear regressions of RTs (observed data) or decision uncertainty (simulated model behavior) onto the congruency between current decisions and current/past stimuli (D-S Cong.; dark symbols) and the congruency between current decisions and current/past decisions (D-D Cong.; light symbols). Green symbols represent the regression coefficients for the simulated data by the constant-criterion model (**A**) or by the criterion-inference model (**B**). Non-green symbols represent the regression coefficients for the auxiliary sets of observed data and are re-plotted in **B** and **C** for model comparisons. Importantly, as predicted by the criterion-inference model (dark green symbols in **B**), current decision RTs were positively regressed onto the congruency between current decisions and past stimuli in each and every data set (non-green color symbols in **A** and **B**; see *Figure 4* and related parts in the main text for this prediction), which was failed to be captured by the constant-criterion model (dark green symbols in **A**).

(**C**) Comparisons of the criterion-inference model and the constant-criterion model in the accountability for the decision RTs in the auxiliary data sets. The data accountability was assessed by the multiple regression of the RTs observed from individual observers onto the decision uncertainty values simulated by the both models using a model of regression, $RT \sim \beta_0 + \beta_C u_{Const.Cri.} + \beta_I u_{Inf.Cri.}$, where $u_{Const.Cri.}$ $u_{Inf.Cri.}$ refer to the trial-to-trial expected values of decision uncertainty simulated by the constant-criterion and criterion-inference model, respectively. The statistically significant regression coefficients at the individual level are demarcated by bold bars. The criterion-inference model explained the RT data better than the constant-criterion model: the significant regression coefficients between observed RTs and decision uncertainty were more prevalent for the criterion-inference model (46.1%) than for the constant-criterion model (15.8%);

the regression coefficients between observed RTs and decision uncertainty were higher for the criterion-inference model than for the constent-criterion model in 73.7% of observers.

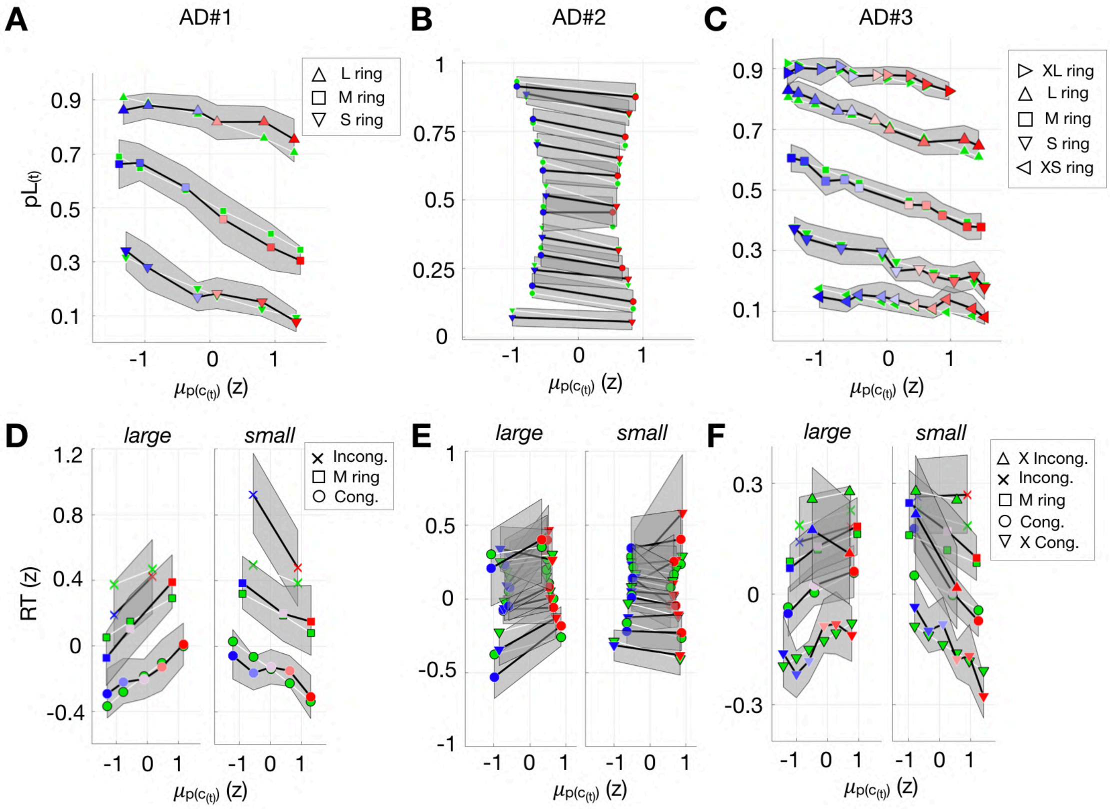


Figure 5-figure supplement 1. Verifying the model's accountability for respective contributions of stimulus and criterion with different sets of stimuli, feedback types, and inter-trial-interval lengths.

(**A**-**F**) Juxtaposition of the observed data (markers in a blue-to-red spectrum) and the criterion-inference model predictions based on simulation (green markers). As for the markers in blue-to-red spectrum, color shifts from blue to red as the inferred criterion ($\mu_p(c_{(t)})$) increases. Gray patches represent 95% bootstrap confidence interval of the across-participant means.

(**A**-**C**) Proportions of 'large' responses (pL) plotted as a function of inferred criterion, separately for different ring sizes. The criterion-inference model (green markers) well predicted the contributions of stimuli (as indicated by different symbols) and inferred criteria (as indicated by different colored symbols) to pL, which were confirmed statistically by the significant positive and negative (multiple) logistic regressions of decisions onto ring size ($Z_{(t)}$) and inferred criterion (onto $\mu_{p(c_{(t)})}$), respectively, for each and every auxiliary data set: $\beta_{Z_{(t)}} = 1.36$ ($P < 10^{-86}$), $\beta_{\mu_{p(c_{(t)})}} = -0.54$ ($P < 10^{-28}$) for AD#1 (**A**); $\beta_{Z_{(t)}} = 1.47$ ($P < 10^{-323}$), $\beta_{\mu_{p(c_{(t)})}} = -0.25$ ($P < 10^{-17}$) for AD#2 (**B**); $\beta_{Z_{(t)}} = 1.44$ ($P < 10^{-72}$), $\beta_{\mu_{p(c_{(t)})}} = -0.35$ ($P < 10^{-44}$) for AD#3 (**C**). As for AD#2, we did not plot the pL data for the largest two and the smallest two stimuli, because only small numbers of trials were avaiable for several observers due to the staircase procedure applied to determine the stimulus sequence for AD#2.

(**D**-**F**) RT plotted as a function of inferred criterion, separately for three different congruence conditions (as specified by different symbols). 'X Cong./Incong.' in (**F**) refers to the congruence between current decision and XL (extra large, 4.16°) or XS (extra small, 3.84°) sized rings for the data set AD#3 (see *Figure 4-figure supplement 4A* for detailed specifications). As explained and stated in the main text (*Figure 5*), the criterion-inference model predicts the negative and positive regressions of decision RT onto ring size ($Z_{(t)}$) and inferred criterion ($\mu_{p(c_{(t)})}$), respectively, for the 'large'-decision trials, whereas it predicts the positive and negative regressions of decision RT onto ring size ($Z_{(t)}$) and inferred criterion ($\mu_{p(c_{(t)})}$), respectively, for the 'small'-decision trials. When multiple linear regressions were carried out, these model predictions were statistically supported by each and every auxiliary data set: $\beta_{Z_{(t)}} = -0.21$ ($P < 10^{-15}$) & $\beta_{\mu_{p(c_{(t)})}} = 0.16$ ($P < 10^{-9}$)for 'large' decisions and $\beta_{Z_{(t)}} = 0.28$ ($P < 10^{-27}$) & $\beta_{\mu_{p(c_{(t)})}} = -0.13$($P < 10^{-5}$) for 'small' decisions for for AD#1 (**D**). $\beta_{Z_{(t)}} = -0.23$ ($P < 10^{-49}$) & $\beta_{\mu_{p(c_{(t)})}} = 0.085$ ($P < 10^{-7}$) for 'large' decisions and $\beta_{Z_{(t)}} = 0.21\ (P < 10^{-41})$ & $\beta_{\mu_{p(c_{(t)})}} = -0.038\ (P = 0.013)$ for 'small' decisions for AD#2 (**E**); $\beta_{Z_{(t)}} = -0.097\ (P < 10^{-6})$ & $\beta_{\mu_{p(c_{(t)})}} = 0.051\ (P < 10^{-6})$ for 'large' decisions and $\beta_{Z_{(t)}} = 0.12\ (P < 10^{-11})$ & $\beta_{\mu_{p(c_{(t)})}} = -0.058\ (P < 10^{-5})$ for 'small' decisions for for AD#3 (**F**). Again, as for AD#2, we did not plot the RT data for the largest two and the smallest three congruence conditions for the same reason in the above.

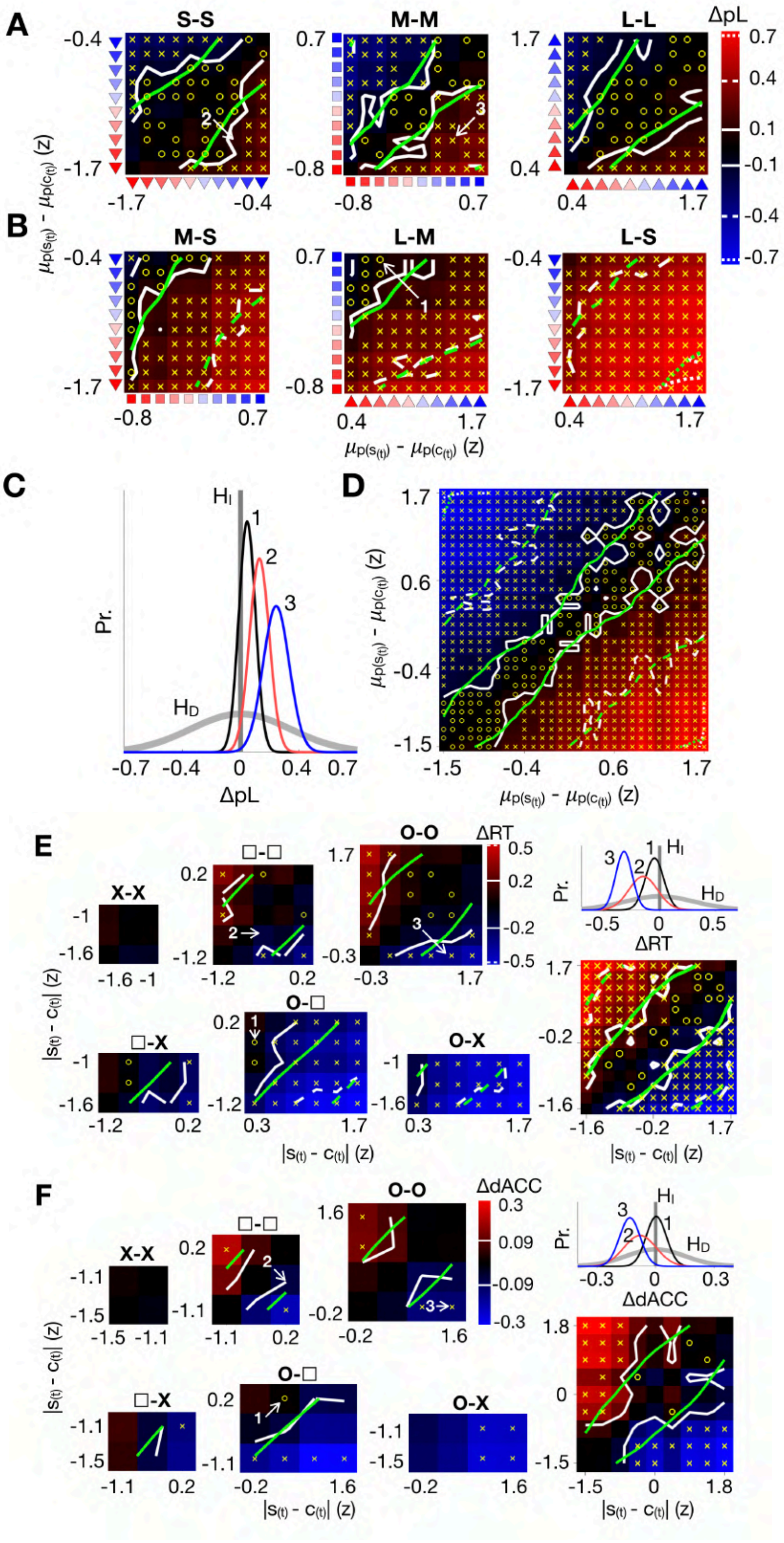


Figure 6-figure supplement 1. Definitions and predictions of decision-probability (pL) metamers and anti-metamers by the criterion-inference model. According to the criterion-inference model, decisions on two different trials, i and j, become 'metamers'—decisions on physically different stimuli that are indiscernible both in decision probability ($\mathrm{pL}_{(i)} \approx \mathrm{pL}_{(j)} | \mathrm{Z}_{(i)} \neq \mathrm{Z}_{(j)}$) and decision uncertainty ($\mathrm{u}_{(i)} \approx \mathrm{u}_{(j)} | \mathrm{Z}_{(i)} \neq \mathrm{Z}_{(j)}$), provided that their difference in inferred stimulus is sufficiently counteracted by the same amount of difference in inferred criterion ($[\mathrm{s}_{(i)} - \mathrm{s}_{(j)}] - [\mathrm{c}_{(i)} - \mathrm{c}_{(j)}] \approx 0$). By the same token, as long as the difference of inferred criterion between two decisions is large enough to go sufficiently beyond their difference of inferred stimulus ($[s_{(i)} - s_{(j)}] - [c_{(i)} - c_{(j)}] \not\approx 0$), two decisions become 'anti-metamers'—decisions on physically identical stimuli that are discernible in decision probability ($pL_{(i)} \not\approx pL_{(j)} | Z_{(i)} = Z_{(j)}$) or decision uncertainty ($u_{(i)} \not\approx u_{(j)} | Z_{(i)} = Z_{(j)}$). These definitions of decision metamers and anti-metamers are falsifiable because our model specifies each and every trial by a set of expected values of $\{s_{(t)}, c_{(t)}, pL_{(t)}, u_{(t)}\}$. Specifically, the set of expected values allows us to deterministically predict the metameric currency for each pair of trials, $[s_{(i)} - s_{(j)}] - [c_{(i)} - c_{(j)}] = [s_{(i)} - c_{(i)}] - [s_{(j)} - c_{(j)}]$, by which we derive when the trials in each pair differ or similar in decision probability ($\Delta pL_{(i,j)} = pL_{(i)} - pL_{(j)}$) and decision uncertainty ($\Delta u_{(i,j)} = u_{(i)} - u_{(j)}$). To statistically verify this 'metamer-to-anti-metamer' mapping for the case of $\Delta pL_{(i,j)}$, we proceeded as follows. First, as shown in ***Figure 6D***, we grouped individual trials according to physical ring sizes (left, middle, and right panels for S, M, and L rings, respectively), and sorted the trials within each stimulus group by $s_{(t)} - c_{(t)}$ into 10 equal-sized bins (as indicated by 10 markers in each panel). The symbols and colors represent the physical ring size and the mean of inferred criteria, respectively. Second, we made all possible (100) pairs of bins for each stimulus group, as shown in the 2D matrix format in (**A**), resulting in a total of 300 within-stimulus bin pairs. Third, we also made all possible (100) pairs of bins for each of the three pairs of stimulus groups resulting in a total of 600 between-stimulus bin pairs, as shown separately in three panels of (**B**) (left, S subtracted from M; middle, L subtracted from M; right, S subtracted from L; the opposite subtractions were not shown because the only difference is sign). Fourth, to judge whether the observed $\Delta pL_{(i,j)}$ at each of the bin pairs (a total of 900 pairs) is significantly "equal to zero (H_I: $\Delta pL_{(i,j)} = 0$)" or "not equal to zero (H_D: $\Delta pL_{(i,j)} \neq 0$), we computed the Bayes factor B (***Dienes, 2008, 2014; Good, 1979; Jeffreys, 1961; Kass and Raftery, 1995; Wod, 1985***) by taking the ratio of the marginal likelihoods of two competing hypotheses, H_I ("Indiscernible") and H_D ("Discernible"): $B_{(i,j)} \equiv \frac{p(\Delta pL_{(i,j)}|H_D)}{p(\Delta pL_{(i,j)}|H_I)}$, which is equivalent with the ratio of the posterior probabilities of two hypotheses $\frac{p(H_D|\Delta pL_{(i,j)})}{p(H_I|\Delta pL_{(i,j)})}$, because we did not have a bias to one of the hypotheses ($p(H_D) = p(H_I)$). Unlike the conventional significance test, the Bayes factor takes into account the posterior probabilities of competing hypotheses and tell us whether the observed data is likely to support the null hypothesis (H_I for Bayes factor < 1/3), the alternative hypothesis (H_D for Bayes factor > 3), or neither (for 1/3 < Bayes factor < 3) (***Jeffreys, 1961***). The Bayes factor was calculated using a free online Bayes factor calculator (***Dienes, 2008***) (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf) by assuming the two prior distributions of H_I ($p(\Delta pL|H_I)$) and H_D ($p(\Delta pL|H_D)$), respectively. $p(\Delta pL|H_I)$ supports only a single value at zero ($\Delta pL = 0$), which means H_I is Dirac delta function. On the other hand, $p(\Delta pL|H_D)$ was assumed to have normal distribution with a mean of zero and an SD identical to the SD of the distribution of non-zero $\Delta pL_{(i,j)}$ (870 pairs), as indicated by the gray curve in **c**. Next, we defined the likelihood $p(\Delta pL_{(i,j)}|\Delta pL)$ given the observed data $\{\Delta pL^1_{(i,j)}, \Delta pL^2_{(i,j)}, \dots, \Delta pL^N_{(i,j)}\}$, where N = 18 (number of observers), for each of the entire (900) cells by taking the mean of the observed data and its standard error. Example likelihood distributions (the colored curves in **C**) for the three bin pairs are overlaid on the prior distribution, whereby the number labels indicate for which cells in A and B the likelihoods were defined. These likelihood distributions were combined with the prior to define the marginal likelihoods, as follows: $p(\Delta pL_{(i,j)}|H_D) = \int_{-1}^{1} p(\Delta pL_{(i,j)}|\Delta pL)p(\Delta pL|H_D)d\Delta pL$ and $p(\Delta pL_{(i,j)}|H_I) = \int_{-1}^{1} p(\Delta pL_{(i,j)}|\Delta pL)p(\Delta pL|H_I)d\Delta pL = p(\Delta pL_{(i,j)}|0)$, which is the intercept of $p(\Delta pL_{(i,j)}|\Delta pL)$. The integration of the second part of the above equation was solved using the definition of Dirac delta function. Next, the Bayes factors were defined by taking the ratio of the marginal likelihoods. For the case "1" (black), although different stimuli ('S' ring and 'M' ring) were presented between these paired trials, the Bayes factor (0.2) was smaller than 1/3 and thus favors H_I ("Indiscernible"). Hence, this case demonstrates the presence of a pair of "decision metamers" (indiscernible decisions on different stimuli). By contrast, the case "3" (blue) has the Bayes factor of 8.4, which is greater than 3 and thus favors H_D ("Discernible"), although an identical ring size (M) was presented between these paired trials. Hence, this case demonstrates the presence of a pair of "decision anti-metamers" (discernible decisions on identical stimuli). The case "2" (red), whereby the Bayes factor had an intermediate value (1.1), illustrates a case in which neither hypotheses is supported. The entire Bayesian judgments of which hypothesis, H_I or H_D, wins in local bin pairs are marked by 'O' or 'X', respectively, in (A) and (B). We omitted the test for the bins with $\Delta pL_{(i,j)} = 0$, because it entails the indiscernibility. Note that decision metamers are indicated by 'O' in the between-stimulus pairs (B) and decision anti-metamers indicated by 'X' in the within-stimulus pairs (A). When the all bin pairs, whether they were within-stimulus or between-stimulus pairs, were put together in a single, unified matrix (**D**), regardless of the stimulus size, we can clearly see that the relative value of stimulus estimate to criterion estimate ($\mu_{p(s_{(t)})} - \mu_{p(s_{(t)})}$) modulates the degree of difference in pL between trials. Here, the color intensity in each cell of the matrix corresponds to the difference in decision probability for each bin pair ($\Delta\mathrm{pL}_{(i,j)}$), whereby the color is black when $\Delta\mathrm{pL}_{(i,j)} = 0$ and become increasingly saturated into blue and red as $\Delta\mathrm{pL}_{(i,j)}$ takes increasingly more negative and positive values, respectively. To help appreciate the correspondences between the Bayes factor outcomes and the model prediction, the observed and the Bayesian iso-$\Delta\mathrm{pL}_{(i,j)}$ contours are overlaid with white and green, respectively, in A, B, and D. Solid, dashed, and dotted contours demarcate $\Delta\mathrm{pL}$ of ± 0.1, ± 0.4, and ± 0.7, respectively. In sum, the results of the Bayes factor analysis show that our criterion-inference model offers accurate definitions of decision metamers and anti-metamers by mapping the estimates of $s_{(t)}$ and $c_{(t)}$ to $\mathrm{pL}_{(t)}$ on a trial-to-trial basis. The almost identical analyses were applied to the RT and dACC data to demonstrate that the model also successfully specifies the metamers and anti-metamers defined in terms of decision uncertainty ($u_{(t)}$). For $u_{(t)}$, the metameric currency is defined in an absolute term, $|s_{(t)} - c_{(t)}|$. (**E, F**) The format is identical to that of A-D. As for decision uncertainty (u), two decisions become metamers in RT and dACC when they have sufficiently similar values of $|s_{(t)} - c_{(t)}|$ (instead of $\mu_{p(s_{(t)})} - \mu_{p(s_{(t)})}$) despite their difference in stimulus-decision congruency ($u_{(i)} = u_{(j)} | \mathrm{Cong.}_{(i)} \neq \mathrm{Cong.}_{(j)}$; marked by ‘O’ in the panels labeled with ‘□-X’, ‘O-□’ and ‘O-X’ in E for RT data and F for dACC data) or anti-metamers in RT and dACC when they have sufficiently different values of $|s_{(t)} - c_{(t)}|$, despite their similarity in stimulus-decision congruency ($u_{(i)} \neq u_{(j)} | \mathrm{Cong.}_{(i)} = \mathrm{Cong.}_{(j)}$; marked by ‘X’ in the panels labeled with ‘X-X’, ‘□-□’ and ‘O-O’ in E and F). X, trials on which stimulus and decision are incongruent; □, on which M ring is presented; O, trials on which stimulus and decision are congruent. As in (C), the priors (gray curve and line) and likelihood (colored curves) distributions of differences in RT and dACC are shown for three example cases of RT/dACC metamers (“1”) and anti-metamers (“3”).

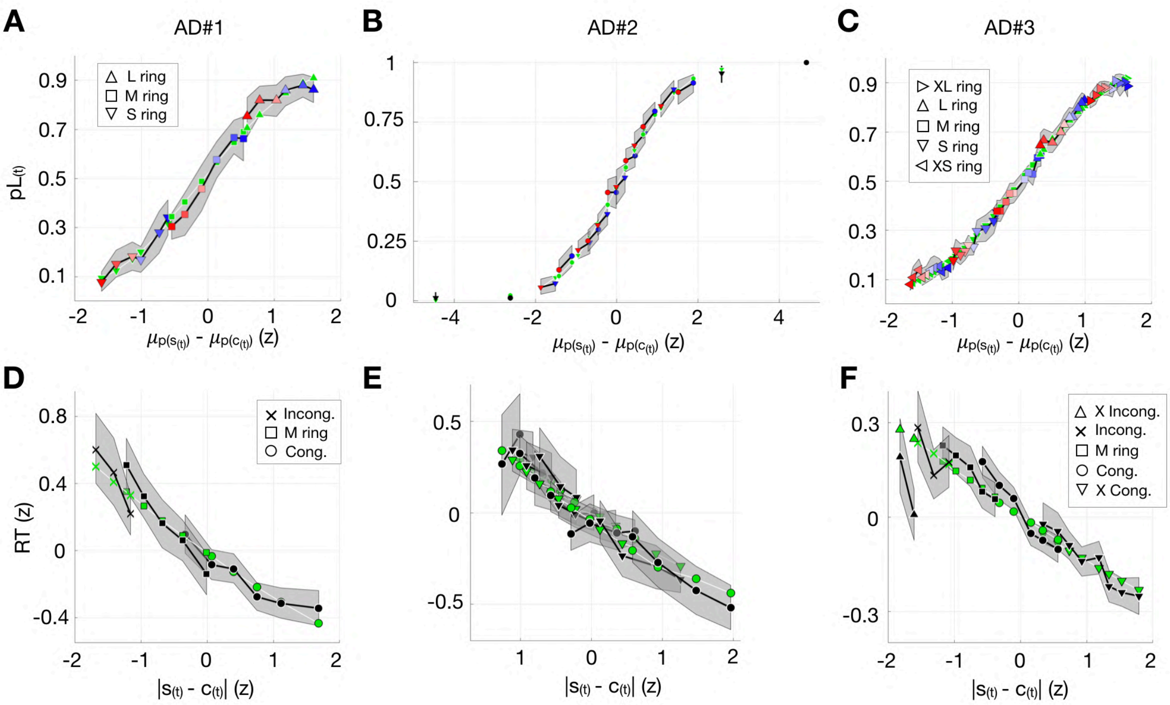
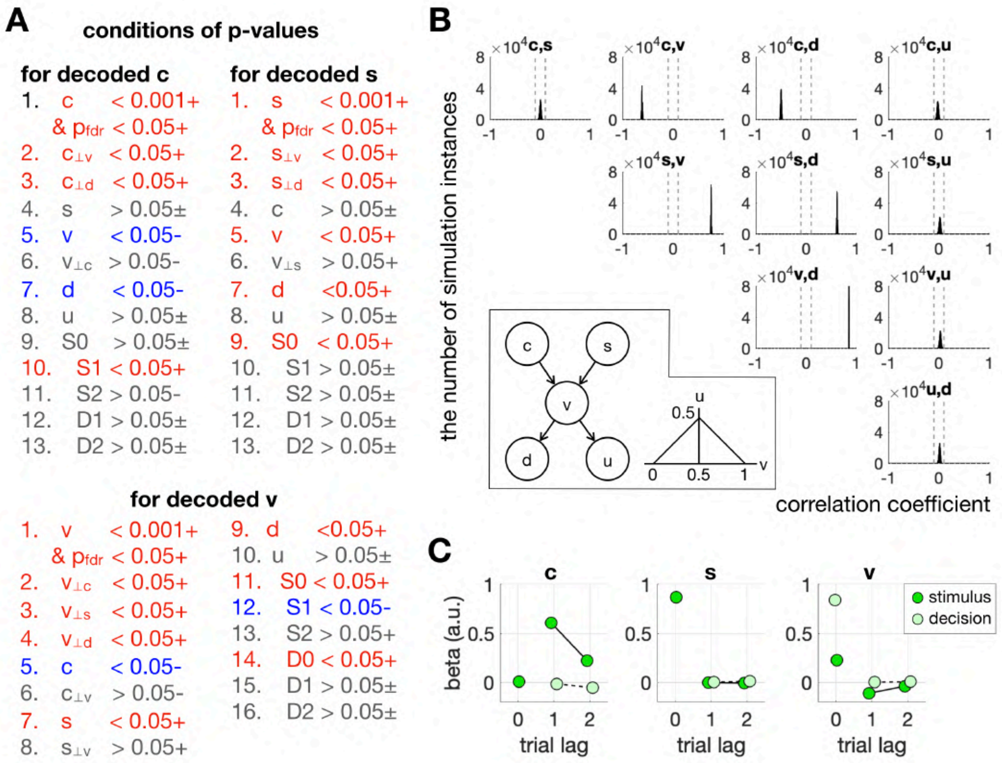


Figure 6-figure supplement 2: Verifying the model's ability to account for the relativity of classification with different sets of stimuli, feedback types, and inter-trial-interval lengths.

(**A**-**C**) Juxtaposition of the observed proportions of 'large' responses (pL; markers in a blue-to-red spectrum) and the criterion-inference model predictions of pLs based on simulation (green markers). pLs are plotted against the binned deviations of ring-size estimates from inferred criterion estimates for each stimulus condition. As for the markers in blue-to-red spectrum, color shifts from blue to red as the inferred criterion ($\mu_p(\mathrm{c}_{(t)})$) increases. Each symbol indicates different stimulus.

(**D**-**F**) Juxtaposition of the observed RT (black symbols) and the criterion-inference model predictions of decision-uncertainty based on simulation (green markers). RTs are plotted against the binned absolute difference between the stimulus and criterion estimates, separately for different stimulus-decision congruence conditions (as indicated by different symbols).

(**A**-**F**) Gray patches represent 95% bootstrap confidence interval. Just as in the main data (*Figure 6*), the predictions of the criterion-inference model fell well within the confidence intervals of the observed data at the majority of data bins for each and every auxiliary data set. As for AD#2, we did not plot the pL data for the largest two and the smallest two stimuli and did not plot the RT data for the largest two and the smallest three congruence conditions, because only small numbers of trials were avaiable for several observers due to the staircase procedure applied to determine the stimulus sequence for AD#2.



A conditions of p-values

for decoded c
1. c $< 0.001+$ & $p_{fdr} < 0.05+$
2. $c_{\perp v}$ $< 0.05+$
3. $c_{\perp d}$ $< 0.05+$
4. s $> 0.05\pm$
5. v $< 0.05-$
6. $v_{\perp c}$ $> 0.05-$
7. d $< 0.05-$
8. u $> 0.05\pm$
9. S0 $> 0.05\pm$
10. S1 $< 0.05+$
11. S2 $> 0.05-$
12. D1 $> 0.05\pm$
13. D2 $> 0.05\pm$

for decoded s
1. s $< 0.001+$ & $p_{fdr} < 0.05+$
2. $s_{\perp v}$ $< 0.05+$
3. $s_{\perp d}$ $< 0.05+$
4. c $> 0.05\pm$
5. v $< 0.05+$
6. $v_{\perp s}$ $> 0.05+$
7. d $<0.05+$
8. u $> 0.05\pm$
9. S0 $< 0.05+$
10. S1 $> 0.05\pm$
11. S2 $> 0.05\pm$
12. D1 $> 0.05\pm$
13. D2 $> 0.05\pm$

for decoded v
1. v $< 0.001+$ & $p_{fdr} < 0.05+$
2. $v_{\perp c}$ $< 0.05+$
3. $v_{\perp s}$ $< 0.05+$
4. $v_{\perp d}$ $< 0.05+$
5. c $< 0.05-$
6. $c_{\perp v}$ $> 0.05-$
7. s $< 0.05+$
8. $s_{\perp v}$ $> 0.05+$
9. d $<0.05+$
10. u $> 0.05\pm$
11. S0 $< 0.05+$
12. S1 $< 0.05-$
13. S2 $> 0.05+$
14. D0 $< 0.05+$
15. D1 $> 0.05\pm$
16. D2 $> 0.05\pm$

Figure 7-figure supplement 1. Deduction of the regressions for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ from the causal structure of manipulated (stimuli), observed (decisions), and latent model variables (c, s, v, u, and d). (**A**) Lists of regressions to be satisfied for $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$. To identify brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$, we defined three a priori lists of regressions onto the manipulated (past and current stimuli), observed (past and current decisions), and latent model variables (inferred criterion, inferred stimulus, decision variable, decision uncertainty, and decision identity) that must be satisfied, respectively, by the brain signals. We stress that each of these lists (1) is a "collectively exhaustive" set, because the constituent regressions encompass all the variables at work in the model as regressors, and (2) consist of "the necessary conditions to be satisfied, and the conditions that must not be satisfied as well", because both significant ($\beta > 0$ or $\beta < 0$) and non-significant ($\beta = 0$) regressions are deduced from the causal structure of variables that are defined by the criterion-inference model (see *Figure 7B*). Thus, these lists subject the candidate brain signals of the latent variables to strong tests, and, if a given brain signal satisfies the entire list of regressions, it must be considered, as one that may be not a mere "neural correlate or signature" of a latent variable, but rather a "neural representation or embodiment" of a latent variable. The meanings of symbols, numbers, and colors used in expressing the regressions are as follows: '+', '–', and '±', signs of regression indicating that the tail of significance test is right, left, or two-tailed, respectively; numbers with decimal points, threshold p-values for GLMM regression of a brain signal onto the variable of interest (P_{FDR}, FDR-corrected p-values) ; '<' and '>', significant and non-significant regression; '$A_{\perp B}$', residual from linear regression of A onto B (which will be referred to as 'A orthogonalized to B' from now on) ; 'D#' and 'S#', decision and stimulus on a #-back trial; red, blue, and gray, positive, negative, and non-significant regressions, respectively. The brain signals of a targeted variable must and must not satisfy the following regressions of the targeted variable:

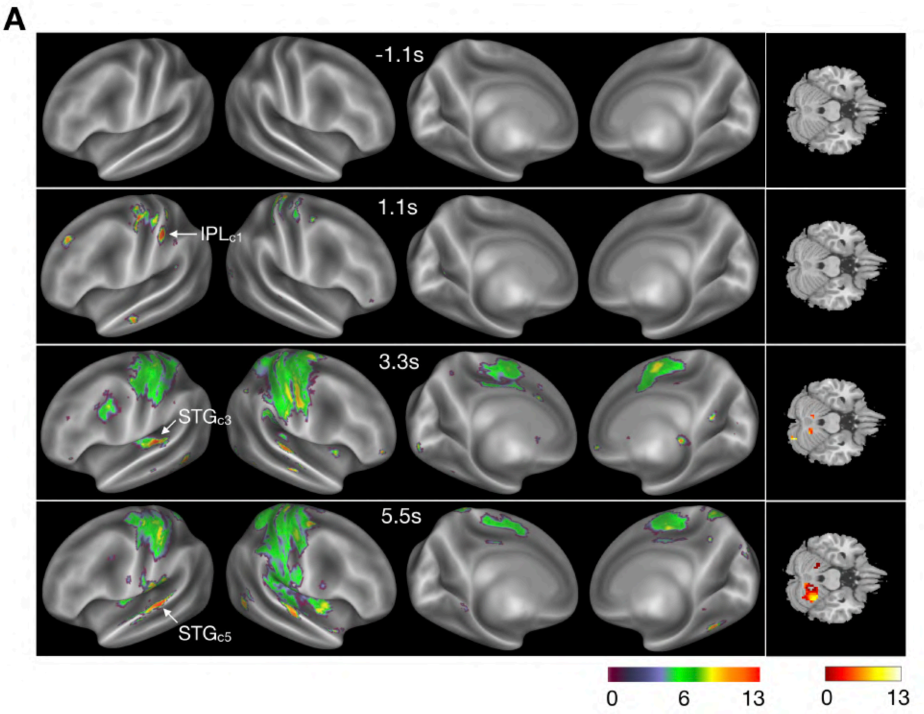
The brain signal of c (y_c; top left column of A). (c1)~(c3), y_c must be regressed onto c—the variable it represents—even when c is orthogonalized to v or d, because it should reflect the variance irreducible to the offspring variables of c; (c4), y_c must not be regressed onto s because c and s are independent; (c5),(c6), y_c must be regressed onto v but not when v is orthogonalized to c because the influence of c on v is removed; (c7),(c8) y_c must be regressed onto d but not onto u because u's relationship with its parents v and c is nonlinear (see *Figure 7E*); (c9)-(c11), 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),(c13), y_c must not be regressed onto previous decisions at all because c is inferred solely from retrieved stimulus measurements.

The brain signal of s (y_s; top right column of A). (s1)~(s3), y_s 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),(s6), 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),(s8) y_s must be regressed onto d but not onto u because u's relationship with its parents v and s is nonlinear (see *Figure 7E*); (s9)-(s11), 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),(s13), y_s must not be regressed onto previous decisions because s is inferred solely from the current stimulus measurement.

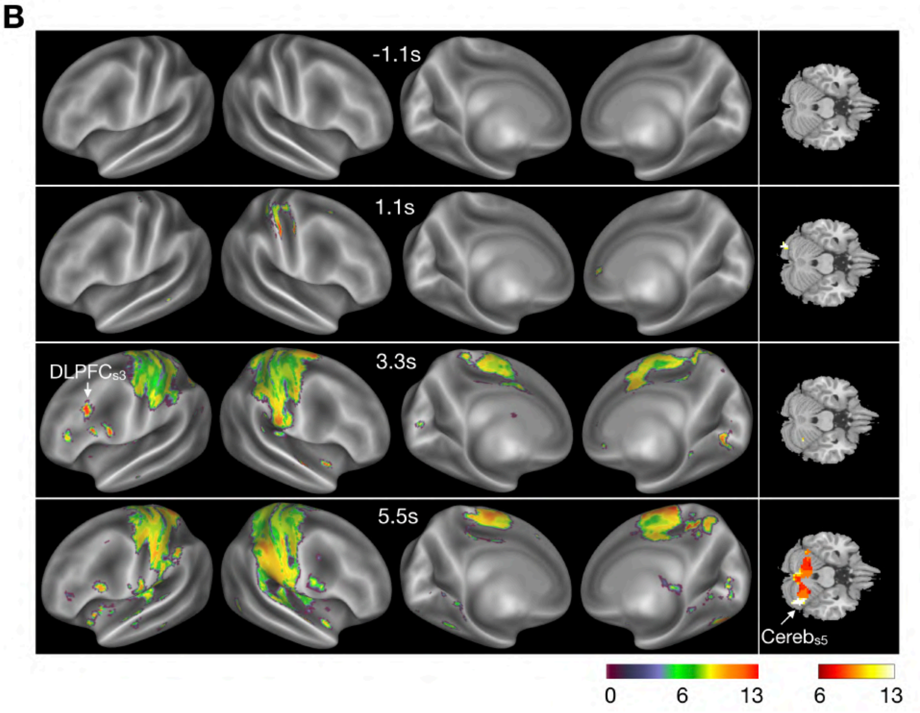
The brain signal of v (y_v; bottom columns of A). (v1)~(v4), y_v must be regressed onto v—the variable it represents—even when v is orthogonalized to c, s, or d, because it should reflect the variance irreducible to the offspring variables of s; (v5),(v6), y_v must be regressed onto one of its parents c, but not when c is orthogonalized to v, because the influence of c on v is removed; (v7),(v8), 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),(v10) y_v must be regressed onto d but not onto u because u's relationship with its parent v is nonlinear (see *Figure 7E*); (v11)-(v13), 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)-(v16), 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.

(**B**) Pairwise correlations between the simulated latent variables. To confirm the above regressions onto all latent variables except for themselves ((c4)~(c8) for y_c; (s4)~(s8) for y_s; (v5)~(v10) for y_v), we calculated the pairwise correlations between those variables from the simulated data as follows. First, we fitted the model to the corresponding human observers' decision behaviors and created the 18 Bayesian observers, each inheriting the fitted model parameters of their human partner. Second, we asked these 18 Bayesian observers to repeatedly (10^6 times) carry out the task on the same sequences of stimuli that were presented to their human partners, respectively, so that we could acquire a large, simulated time series of trial-to-trial values of the latent variables, $c_{(t)}$, $s_{(t)}$, $v_{(t)}$, $d_{(t)}$, and $u_{(t)}$. Finally, we calculated the Pearson's correlations between those variables for individual simulations, resulting in 10^6 coefficients for each of all possible (10) pairs of variables. The results are shown in a histogram, with one panel for each pair. Bilateral black dashed lines indicate ± 0.1 Pearson correlation coefficients around zero. The distributions of simulated variable values were consistent with the regressions deduced from the causal structure in the model, as follows: '$r(c, s) = 0.0023$ (median)' is consistent with the regressions of (c4) and (s4); '$r(c, v) = -0.62$' with the regressions of (c5), (c6), (v5), and (v6); '$r(c, d) = 0.50$' with the regressions of (c7); '$r(c, u) = 0.022$' with the regressions of (c8); '$r(s, v) = 0.75$' with the regressions of (s5), (s6), (v7), and (v8); '$r(s, d) = 0.62$' with the regressions of (s7); '$r(s, u) = 0.019$' with the regressions of (s8); '$r(v, d) = 0.85$' with the regressions of (v9); '$r(v, u) = 0.027$' with the regressions of (v10). To demonstrate how the regressions are deduced and related to the correlations, the causal graph and the nonlinear causal function from v to u are shown in the bottom left-hand corner.

(**C**) Multiple logistic regressions of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ onto stimuli and decisions. To confirm the regressions onto the manipulated (stimuli) and observed (decisions) variables ((c9)~(c13) for y_c; (s9)~(s13) for y_s; (v11)~(v16) for y_v), we carried out multiple logistic regressions of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$ onto the stimuli presented to and (simulated) decisions made by the Bayesian observers. The patterns of regression coefficients were consistent with the regressions deduced from the causal structure in the model, as follows: left panel was consistent with (c9)~(c13); middle panel was consistent with (s9)~(s13); right panel was consistent with (v11)~(c16).



1514
1515
1516



1517
1518

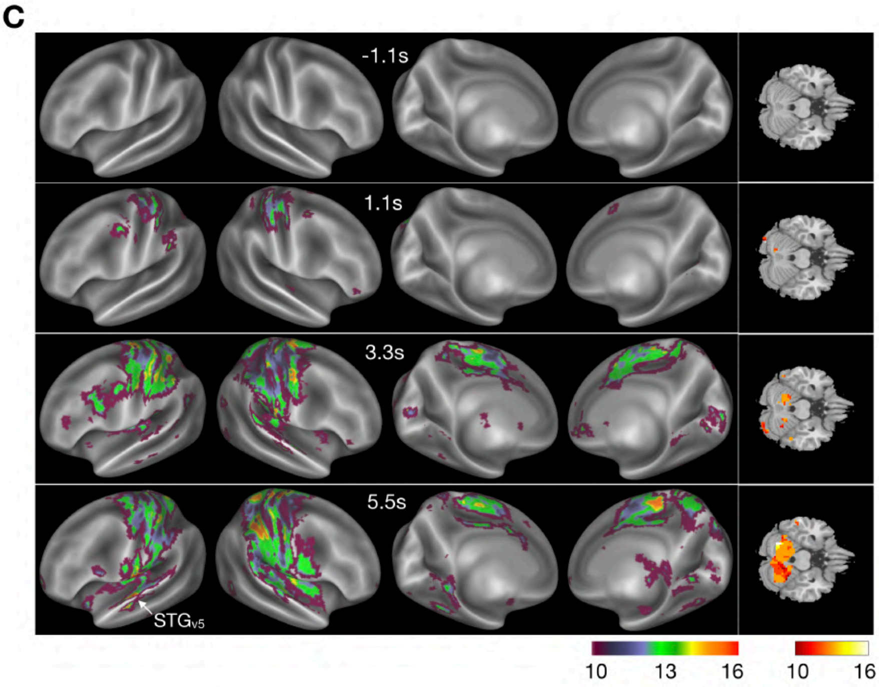
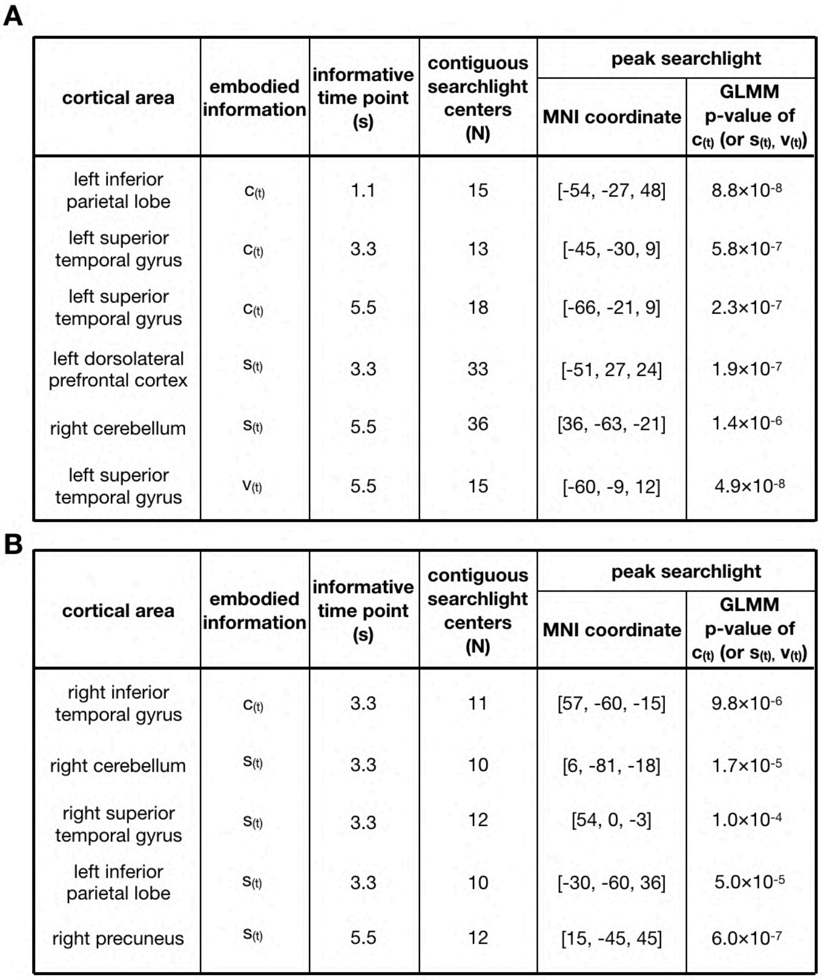


Figure 7-figure supplement 2. Maps of the numbers of regression tests satisfied for the latent variables in the model (s, c, and v). (**A**-**C**), Each row represents the decoded information of $c_{(t)}$ (**A**)**,** $s_{(t)}$(**B**), and $v_{(t)}$(**C**) at a specific time point relative to stimulus onset. The color hue represents how many of the regression tests (see *Figure 7-figure supplement 1* for complete definitions) were satisfied by the candidate brain signals of the latent model variables. The score maps in NIfTI format are available in Github (see Methods: Data and code availability).



A

cortical area	embodied information	informative time point (s)	contiguous searchlight centers (N)	peak searchlight	
				MNI coordinate	GLMM p-value of $c_{(t)}$ (or $s_{(t)}$, $v_{(t)}$)
left inferior parietal lobe	$c_{(t)}$	1.1	15	[-54, -27, 48]	8.8×10^{-8}
left superior temporal gyrus	$c_{(t)}$	3.3	13	[-45, -30, 9]	5.8×10^{-7}
left superior temporal gyrus	$c_{(t)}$	5.5	18	[-66, -21, 9]	2.3×10^{-7}
left dorsolateral prefrontal cortex	$s_{(t)}$	3.3	33	[-51, 27, 24]	1.9×10^{-7}
right cerebellum	$s_{(t)}$	5.5	36	[36, -63, -21]	1.4×10^{-6}
left superior temporal gyrus	$v_{(t)}$	5.5	15	[-60, -9, 12]	4.9×10^{-8}

B

cortical area	embodied information	informative time point (s)	contiguous searchlight centers (N)	peak searchlight	
				MNI coordinate	GLMM p-value of $c_{(t)}$ (or $s_{(t)}$, $v_{(t)}$)
right inferior temporal gyrus	$c_{(t)}$	3.3	11	[57, -60, -15]	9.8×10^{-6}
right cerebellum	$s_{(t)}$	3.3	10	[6, -81, -18]	1.7×10^{-5}
right superior temporal gyrus	$s_{(t)}$	3.3	12	[54, 0, -3]	1.0×10^{-4}
left inferior parietal lobe	$s_{(t)}$	3.3	10	[-30, -60, 36]	5.0×10^{-5}
right precuneus	$s_{(t)}$	5.5	12	[15, -45, 45]	6.0×10^{-7}

Figure 7-figure supplement 3. Specifications of the ROIs in which activity was informative of the latent variables in the model (s, c, and v). By subjecting the candidate correlates of $c_{(t)}$, $s_{(t)}$ and $v_{(t)}$, which were detected by the whole-brain search using the MVPA technique (see Methods: Searching for multivoxel patterns of activity signaling latent model variables), to the exhaustive sets of necessary regression tests for the model's latent variables (see *Figure 7-figure supplement 1* for complete definitions), we identified three, two, and one clustered regions in which the brain signals of $c_{(t)}$, $s_{(t)}$, and $v_{(t)}$, respectively, were embodied. The GLMM significances were calculated by one-tailed tests. (**A**) ROIs defined with the procedure in which the threshold for the number of contiguous searchlights satisfying all the regression tests was set to 13. (**B**) ROIs defined with the procedure in which a more relaxed threshold (the number of contiguous searchlights is more than 9) was used.

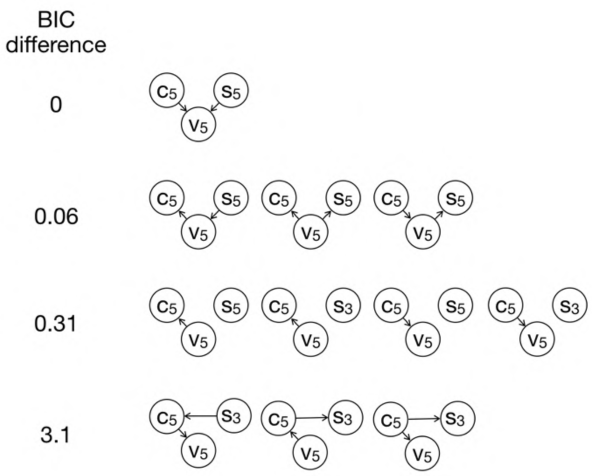


Figure 7-figure supplement 4. Correspondence between the causal graph of the latent variables that is prescribed *a priori* by the model ($c \rightarrow v \leftarrow s$) and the *maximum likelihood* causal graph that is inferred from the brain signals representing those variables. To see whether the patterns of c, s and v signals that were decoded in the brain regions ($\mathrm{IPL_{c1}}, \mathrm{STG_{c3}}, \mathrm{STG_{c5}}, \mathrm{DLPFC_{s3}}, \mathrm{Cereb_{s5}}, \mathrm{STG_{v5}}$; *Figure 7A-F*; *Figure 7-figure supplement 1, 2, 3*) are consistent with the causal graph structure that is prescribed *a priori* by the model, 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 ($\{\overleftarrow{\mathrm{X}}_\mathrm{c}, \overleftarrow{\mathrm{X}}_\mathrm{s}, \overleftarrow{\mathrm{X}}_\mathrm{v}\}$): $argmax\ p(G|\{c, s, v\} = \{\overleftarrow{\mathrm{X}}_\mathrm{c}, \overleftarrow{\mathrm{X}}_\mathrm{s}, \overleftarrow{\mathrm{X}}_\mathrm{v}\})$. A total of 27 causal graphs can be created out of three variables (3 x 3 x 3); a total of 6 triplets of brain signals can be used for $\{\overline{\mathrm{X}}_\mathrm{c}, \overline{\mathrm{X}}_\mathrm{s}, \overline{\mathrm{X}}_\mathrm{v}\}$ since we have three ($\mathrm{IPL_{c1}}, \mathrm{STG_{c3}}, \mathrm{STG_{c5}}$), two ($\mathrm{DLPFC_{s3}}, \mathrm{Cereb_{s5}}$), and single ($\mathrm{STG_{v5}}$) candidate brain signals for c, s, and v, respectively (3 x 2 x 1). Thus, we could evaluate which of the 27 states of G(raph) is most likely for each of the 6 triplets of brain signals. Since the 27 graph states differ in complexity, i.e., the number of parameters, we used Bayesian Information Criterion (BIC) values for evaluation (the smaller a BIC value is, the more likely a graph is). The outcomes of BIC evaluation were consistent with our model of c inference in the following two aspects.

First, out of the 162 possible pairs of possible causal graph structures and possible triplets of brain signals (162 = 27 graph structures x 6 triplets of brain signals), the smallest BIC value was found for the pair of the graph structure '$c \rightarrow v \leftarrow s$' and the triplet of brain signals $\{\overleftarrow{\mathrm{X}}_\mathrm{c} = \mathrm{STG_{c5}}, \overleftarrow{\mathrm{X}}_\mathrm{s} = \mathrm{Cereb_{s5}}, \overline{\mathrm{X}}_\mathrm{v} = \mathrm{STG_{v5}}\}$. This causal graph of 'common effect' is exactly the one prescribed by the model. Note that, given any correlation data, the graph of 'common effect' is distinguishable from the graphs of 'chain' and 'common cause', which are indistinguishable from one another by contrast, because the 'common effect' graph is Markov-equivalent neither with the 'chain' graph nor with the 'common cause' graph whereas the 'chain' graph and the 'common cause' graph are Markov equivalent. Note also that the winner graph, '$c \rightarrow v \leftarrow s$', is not Markov-equivalent with neither of the other possible 'common effect' graph structures, '$v \rightarrow c \leftarrow s$' and '$c \rightarrow s \leftarrow v$'. For that matter, '$c \rightarrow v \leftarrow s$' is unique in terms of Markov equivalence, being distinguishable from the entire rest of the possible (26) graphs. This is why '$c \rightarrow v \leftarrow s$' can stand out alone as the most likely causal graph given the evidence of brain signals.

Second, note that there exists one critical condition that can falsify our model: the 'significant' existence of any graph that includes causal arrows between c and s. This is so because our model is built upon the assumption that c and s are the random variables that are independent from one another. To examine the statistical significance of those graphs, we applied 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 passed the criterion. The closest Markov-equivalent group of graphs (shown at the bottom) was well apart from the winner graph, more by the BIC value of 3. The runners-up were the two Markov-equivalent groups of graphs (the second and third rows). In sum, our model prediction survived the two strong tests: (1) '$c \rightarrow v \leftarrow s$' wins the entire rest of the graph structures in the likelihood battle of causality; (2) the model's fundamental assumption, the independence between c and s, was not falsified. These outcomes must not be taken lightly but rather seriously as a strong support for the existence of brain embodiment of the latent variables and their interplay that our model of c inference predicts.

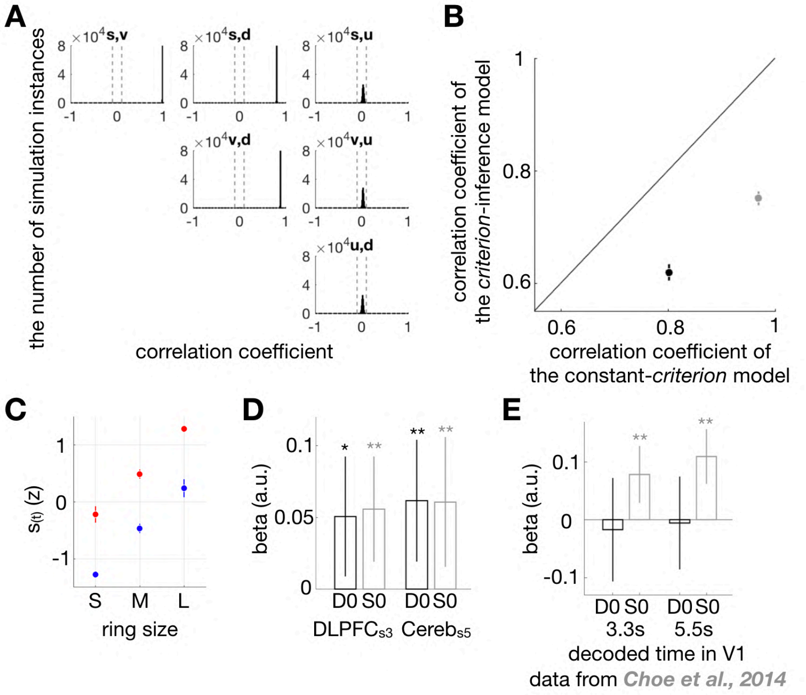


Figure 8-figure supplement 1. Specifications of the latent variables of the criterion-inference model. (**A, B**) To assess how the absence and presence of criterion inference influence the relationships between the remaining latent variables ($s_{(t)}$, $v_{(t)}$, $d_{(t)}$, and $u_{(t)}$), we compared the patterns of the pairwise correlations among those variables between the constant-criterion model and the criterion-inference model. To do so, as we did for the criterion-inference model (*Figure 7-figure supplement 1*), we created another group of 18 Bayesian observers by fitting the constant-criterion model separately to the individual human observers, and then acquired the pair-wise correlations based on the simulated data (**A**). We then pitted the median coefficients of the 10^6 simulated correlations and their range (2.5 and 97.5 percentiles) for the constant-criterion model against those for the criterion-inference model (**B**). Black and gray colors indicate correlations $r(s, d)$ and $r(s, v)$, respectively. By comparing the pairwise correlations, the repercussion of the computational absence of inferred criterion ($c_{(t)}$) was clearly deomontrated; all correlations involving *s* were stronger in the constant-criterion model. This implies that, due to the absence of *c*, *s* monopolizes the ecology of latent variables constituting the PDM process, making it difficult to determine the contribution of individual constituents and the interplays between them. (**C, D**) $c_{(t)}$ and $s_{(t)}$ interplay in PDM on equal computational footings; as $c_{(t)}$ is an inferred value of the median of external stimuli on trial t (*Equation 2; Figure 3D*), $s_{(t)}$ is an inferred value of the stimulus on trial t (*Equation 1; Figure 3C*). As defined in *Figure 3*, this implies that trial-to-trial measurement noise in the sensory system are substantive and propagate to $s_{(t)}$, then to $v_{(t)}$, and consequently affect decisions. In line with this, the simulated estimates of $s_{(t)}$ by the Bayesian observers (see *Figure 7-figure supplement 1* for how these simulations were carried out) differ substantively between the 'small'-decision (blue dots) and the 'large'-decision (red dots) trials, as shown in (**C**). Correspondingly, the brain signals of *s* found in the DLPFC and the cerebellum (DLPFC_{s3}, Cereb_{s5}) also varied not only as a function of external stimuli, but also as a function of decisions made by the observers, as shown in (**D**). Specifically, the multiple linear regressions 'DLPFC_{s3} (or Cereb_{s5}) $\sim\beta_0 + \beta_{S0}S0 + \beta_{D0}D0$' revealed that β_{S0} and β_{D0} are both significant (DLPFC_{s3}: $\beta_{D0} = 0.051, P_{D0} = 0.017$; $\beta_{S0} = 0.056, P_{S0} = 0.0028$; Cereb_{s5}: $\beta_{D0} = 0.062, P_{D0} = 0.0044$; $\beta_{S0} = 0.061, P_{S0} = 0.0084$). P-values, *< 0.05; **< 0.01. Error bars represent 95% GLMM confidence intervals. (**E**) To explore the possibility that *s* signals might reside in the early retinotopic visual cortex, we re-analyzed the fMRI data presented in Choe et al. (2014), the experimental procedure and stimuli of which were almost identical to those of the current study except that only the visual cortex was imaged with high spatial resolution. Specifically, we decoded stimulus sizes from the V1 population activity using the fine-retinotopy-based population decoding method used in Choe et al. (2014) and then conducted the same regression analyses that were applied to the brain signals of *s* in the DLPFC and the cerebellum (DLPFC_{s3}, Cereb_{s5}). Unlike DLPFC_{s3} and Cereb_{s5}, the decoded V1 signal of stimulus size was significantly regressed onto physical stimulus size ($\beta_{S0} = 0.078, P_{S0} = 0.0018$ at 3.3 s from stimulus onset; $\beta_{S0} = 0.11, P_{S0} < 10^{-5}$ at 5.5 s from stimulus onset) but not onto decisions made by observers ($\beta_{\mathrm{D0}} = -0.017, \mathrm{P}_{\mathrm{D0}} = 0.71$ at 3.3 s from stimulus onset; $\beta_{\mathrm{D0}} = -0.0054, \mathrm{P}_{\mathrm{D0}} = 0.90$ at 5.3 s from stimulus onset). This implies that the V1 population activity represents physical stimuli robustly but its trial-to-trial variability does not seem to be causally involved in determining the trial-to-trial variability in choice, which contradicts the property of s in our model.

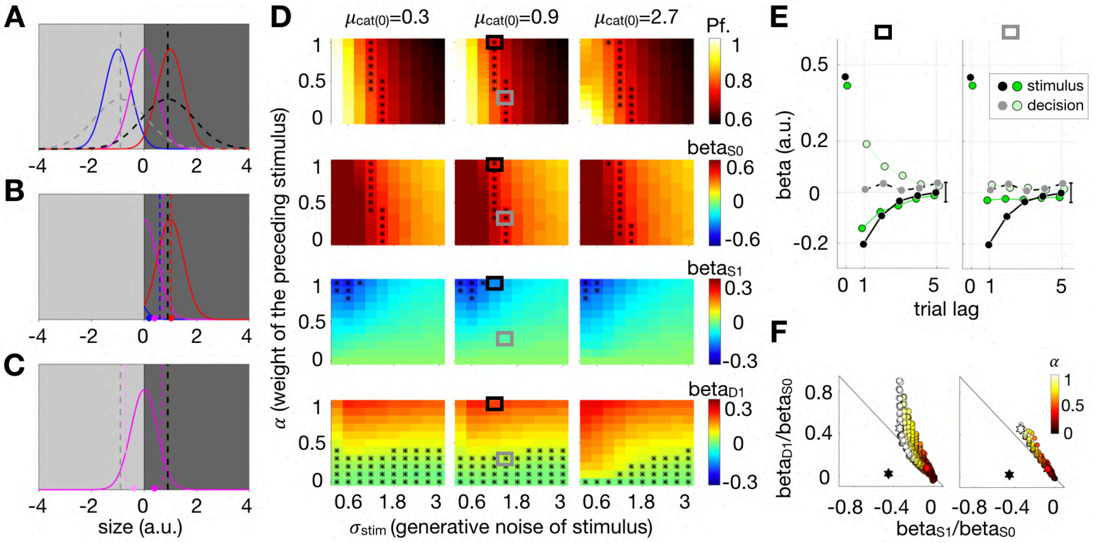


Figure 8-figure supplement 2. The ‘classification’ with a generative model for ‘categorization’ cannot give a coherent account for the history effects of previous stimuli and decisions on current decisions. As stated in the main text, the fundamental feature of the generative model for categorization is that observers assume that (i) there are two category states (‘small’ and ‘large’) and (ii) each category state engenders noisy samples with a respective mean. Norton and his colleagues contrived multiple algorithms by which one can make decisions on which category given samples come from and update the category means in the generative model on a trial-to-trial basis (Norton et al., 2017). They found the best algorithm to be the one whereby, on each trial t, each of the category means ($\mu_{'large'(t)}$ and $\mu_{'small'(t)}$) is updated by averaging past sensory evidences that were categorized to the corresponding category. For example, a decision on trial t, if a current sample, $z_{(t)}$, is greater than the average of the current category means, $(\mu_{'large'(t)} + \mu_{'small'(t)})/2$, the ‘large’ decision will be made (D(t) = ‘large’), and only the mean of the ‘large’ category will be updated such that $\mu_{'large'(t+1)} = \alpha z_{(t)} + (1-\alpha)\mu_{'large'(t)}$, where α is the relative weight given to the current sample, while the mean of the ‘small’ category will remain unchanged ($\mu_{'small'(t+1)} = \mu_{'small'(t)}$). Critically, this way of updating the parameters of the generative model, namely, the two category means over trials, make a ‘falsifiable’ prediction about the history effects of previous stimuli and decisions on current decisions, which is qualitatively different from the prediction made by our model. Our model predicted the ‘repulsive’ effects of previous stimuli on current decisions (as indicated by the negative regression weights in *Figure 4A*) but no effects of previous decisions on current decisions (as indicated by the zero regression weights in *Figure 4A*). By contrast, the categorization model predicts the ‘repulsive’ effects of previous stimuli on current decisions, which is same to ours, but the ‘attractive’ effects of previous decisions on current decisions, which is different from ours.

(**A**) Hypothetical sampling distributions of noisy stimulus measurements (indicated by colored curves) and a generative model of categorization (represented by gray curves). The blue, purple, and red curves correspond to the measurements generated by the ‘S’, ‘M’, and ‘L’ ring-size stimuli. The generative model for categorization, which is hypothetically assumed to be adopted by observers (instead of the generative model for classification), includes two distributions, one belonging to the ‘large’ category (dashed dark gray curve) and the other to the small category (dashed light gray curve). Vertical dashed lines demarcate the means of these two distributions of the generative model at the moment of trial t: $\mu_{'small'(t)}$ and $\mu_{'large'(t)}$. The center of these two means, $(\mu_{'large'(t)} + \mu_{'small(t)})/2$, is demarcated by the boundary between the light and dark gray background patches. This boundary can be considered as a decision rule. For example, if a noisy measurement, $z_{(t)}$, is located on the dark side, a decision of ‘large’ category will be rendered.

(**B**) An illustration of ‘repulsive’ history effects of previous stimuli on current decisions in the case of ‘large’ decision. Colored dots represent the respective means of noisy measurements sampled from the three ring-size stimuli (labeled by the corresponding colors) which resulted in ‘large’ decisions. According to the categorization model’s algorithm for generative model updating, these noisy measurements will attract the mean of generative-model distributions, as indicated by the vertical dashed lines of the corresponding colors, which represent the updated means $\mu_{'large'(t+1)}$. Note that $\mu_{'large'(t+1)}$ increases as the previous stimuli were larger, which results in the 'repulsive' decision bias because category boundary, $(\mu_{'large'(t)} + \mu_{'small(t)})/2$, also increases along with $\mu_{'large'(t+1)}$, which will lead to more decisions of 'small' category.

(**C**) An illustration of 'attractive' history effects of previous decisions on current decisions in the case of 'medium' stimulus. The purple dots of intense and weak colors indicate the means of noisy measurements sampled from medium stimulus that resulted in the decisions of 'large' and 'small' categories, respectively. Accordingly, the purple vertical dashed lines of intense and weak colors demarcate the updated means of the generative-model distributions, $\mu_{'large'(t+1)}$ and $\mu_{'small'(t+1)}$, respectively. Consequentially, if the decision on trial t was 'large', then $\mu_{'large'(t+1)}$ becomes smaller than $\mu_{'large'(t)}$, which will lead to more decisions of 'large' category, and vice versa. Therefore, unlike the stimulus history effects, previous decisions generally tend to attract current decisions.

(**D**) To confirm the above intuitive predictions by the categorization model, we ran simulation experiments while varying the model parameters, $[\alpha, \sigma_{stim}, \mu_{cat(0)}]$, within reasonable ranges, where α is the relative weight given to a stimulus sample, σ_{stim} is the generative noise of $z_{(t)}$, and $\mu_{cat(0)}$ is the initial mean of the 'large' category and the additive reciprocal of the 'small' category. $z_{(t)}$ was sampled from a normal distribution with mean $\in [-1, 0, 1]$, each represent 'S', 'M', and 'L' rings, respectively, and standard deviation σ_{stim}. The stimulus sequences identical to those used in our experiment were used. For each triplet of model parameters, we ran 10,000 stimulations and measured the overall performance (proportion of correct responses for 'S' and 'L' stimuli; top row) and regressed current decisions onto current stimuli (regression coefficients of S_0 shown in the second row), past stimuli (regression coefficients of S_1 in the third row), and past decisions (regression coefficients of D_1 in the bottom row). The asterisks mark a constricted set of model parameters with which simulated performances (top row) or regression coefficients (second to bottom rows) fell within the 95% bootstrap confidence intervals of the corresponding human data that were observed in our experiment. Importantly, two aspects of the simulation outcomes indicate the incapability of the categorization model to account for the observed history effects of previous stimuli and decisions on current decisions: (1) the coefficients of past stimuli and past decisions have opposite signs, which is in conflict with the observed ones; (2) there is no regime of the parameter space where performances or regression coefficients are concurrently in line with the observed human data.

(**E**) Temporally extended multiple linear regression results for the two triplets of model parameters that were marked by the black (left panel) and gray (right panel) boxes in (D). For comparison, the regression coefficients from the simulations (light and dark green circles) are juxtaposed with those from the actual experiments (light and dark gray circles, with their average 95% confidence interval shown in the right). In line with the outcome shown in (D), the coefficients of past stimuli and decisions deviated substantively from the observed ones. One might find the parameter regime at which the simulated coefficients for past stimuli fall near to the observed ones. However, in that regime, the simulated coefficients for past decisions fall far from the observed ones (left panel). Likewise, one might find the parameter regime at which the simulated coefficients for past decisions fall near to the observed ones. Now, in that regime, the simulated coefficients for past stimuli fall far from the observed ones (right panel). This testifies the fundamental incapability of the categorization model to give a coherent account for the observed history effects of past stimuli and past DECISIONS.

(**F**) The simulated regression coefficients for previous stimuli and previous decisions are plotted against one another, after being normalized by those for current stimuli for the comparisons across different parameter conditions. The left panel shows the results for the entire set of parameter triplet shown in (D), whereas the right panel shows those for the selective set of parameter triplet marked by the asterisks in the top panels in (D), where the simulated performances matched the human performances. In both of the panels, the warm of color indicates the value of parameter α. As α increased, both of the past decision and stimulus effects increased. The results for the parameters used in the left and right panels in (E) were marked by white and red hexagons respectively. Note that, regardless of α levels, the simulated pairs of past stimuli were stuck around or above the diagonal line and never approached the observed coefficients that are shown as the black hexagons at the left bottom corner. Again, this echoes the conclusion from (E): the categorization model cannot explain the observed history effects of past stimuli and past decisions, which were readily accounted for by our model of c inference.