이학박사 학위논문

# Boundary updating as a source of history effects on choice and decision uncertainty

## 판단 경계의 업데이트가 선택과 선택 불확실성의 역사 효과에 미치는 영향

2023년 8월

서울대학교 대학원

뇌인지과학과

이 희 승

# Boundary updating as a source of history effects on choice and decision uncertainty

지도 교수 이 상 훈

이 논문을 이학박사 학위논문으로 제출함
2023년 8월

서울대학교 대학원
뇌인지과학과
이 희 승

이희승의 이학박사 학위논문을 인준함
2023년 8월

위 원 장 Lee Sang Ah (인)

부위원장 이 상 훈 (인)

위 원 이 대 열 (인)

위 원 강 민 석 (인)

위 원 안 우 영 (인)

# Abstract

One crucial structure of sensory objects is the relative scale, which can be expressed throughd relative adjectives such as ‘large.’ Linguists and cognitive scientist have assumed that determining the appropriate relative adjective of objects requires a boundary value that separates the stimulus values of ‘small’ from those of ‘large.’

However, the existence of a boundary for relative adjectives has been a hypothetical factor, and its neural evidence remains uncertain. Thus, I delved into investigating the neural evidence of boundary for relative adjectives that describe the size of a ring, like ‘large’ or ‘small.’ To accomplish this, I leveraged a unique property of boundaries known as boundary updating. Linguists and cognitive scientists have assumed that boundaries assimilate to previous objects on each trial and thus bias current choices repelled from previous stimuli. I analyzed brain signals from human fMRI data to identify the neural signal that reflects boundary updating. The results demonstrated the presence of brain signals associated with boundary updating in the parietal and temporal cortices, providing neural evidence that supports the existence of boundary updating as a source of the repulsive bias.

On another note, while the impact of boundary updating on choices has been extensively studied, its effect on decision uncertainty has not been previously reported. Therefore, I expanded the exploration of boundary updating to decision uncertainty. By using the three correlates of decision uncertainty – response time, pupil size, and the neural signal of dorsal anterior cingulate cortex – the findings revealed that boundary updating systematically biases

the current level of decision uncertainty based on previous stimuli. This discovery highlights that boundary updating is a fundamental cognitive process that influences both history effects of choice and decision uncertainty.

The findings of this thesis may have several implications for our understanding of how the brain perceives the relative scale of objects. One notable implication is that the brain may process the relative scale of objects by rescaling the one-dimensional magnitude representation in the parietal cortex using a preceding neural signal of the boundary while also engaging the language-related circuits of the superior temporal cortex.

**주요어** : Boundary updating, History effect, Repulsive bias, Decision-making, Decision uncertainty
**학 번** : 2017-38129

# Table of Contents

Chapter 1 Introduction ..... 1

1.1 From the Structure of Thoughts to the Structure of Sensory Perception ..... 1

1.2 The Significance of Reference in the Structure of Objects ..... 2

1.3 Reference Updating as a Source of the Repulsive Bias ..... 3

1.4 Structure of the Thesis ..... 5

Chapter 2 Neural evidence for boundary updating as the source of the repulsive bias in classification ..... 7

2.1 Introduction ..... 7

2.2 Materials and Methods ..... 10

2.3 Results ..... 31

2.4 Discussion ..... 56

Chapter 3 Boundary updating as a source of history effect on decision uncertainty ..... 61

3.1 Introduction ..... 61

3.2 Materials and Methods ..... 64

3.3 Results ..... 69

3.4 Discussion ..... 75

Chapter 4 General Discussion ..... 79

4.1 A review of history effects of previous stimuli and previous choices ..... 79

4.2 A new perspective on the system-level neural processing of perceptual decision-making of the relative structure ..... 83

4.3 Limitations ..... 91

Bibliography ..... 96

Abstract in Korean ..... 107

# Figures and Tables

**Figure 1** Two contending hypotheses on the origin of the repulsive bias in binary classification..... 8

**Figure 2** Binary classification task on ring size..... 11

**Figure 3** Influences of previous and current stimuli on classification behavior and V1 activity in Experiment 1 ..... 33

**Figure 4** Origin of the covariation between the stimulus-encoding signal of V1 and the current choice..... 36

**Figure 5** Repulsive bias in Experiment 2 and a Bayesian model of boundary updating (BMBU) ..... 41

**Figure 6** Brain signals of the latent variables of BMBU..... 48

**Figure 7** The probable causal structures between the brain signals of the latent variables in BMBU ..... 50

**Figure 8** Origin of the covariation between the current choice and the brain signals of the latent variables in BMBU ..... 51

**Figure 9** The pre-congruence effect on decision uncertainty62

**Figure 10** The pre- and the current-congruence effects on RT, cortical activity, and pupil-size ..... 70

**Figure 11** The pre- and current-congruence effects on reaction time (RT) in other task conditions ..... 73

**Figure 12** The time courses of dACC, insula, and pupil-size and their linear regressions to decision uncertainty ..... 75

**Figure 13** The regression coefficient between consecutive decision uncertainties..... 76

**Figure 14** Controlling the false discovery rate (FDR) of the ***P*** values of the whole-brain analysis..... 77

**Figure 15** A schematic overview of the visual processing for decision-making. The visual processing is composed of five stages ..... 84

**Figure 16** The two decision-making scenarios ..... 94

**Table 1** The sets of regressions that BMBU requires the brain signals of its latent variables to satisfy ..... 28

**Table 2** Specification of the brain signals of the latent variables of BMBU ..... 47

**Table 3** The brain regions where there was a significant correlation between BOLD signals and simulated decision uncertainty ..... 74

# Chapter 1

# Introduction

## 1.1. From the Structure of Thoughts to the Structure of Sensory Perception

The understanding that our thoughts have a structured nature has a longstanding history in Western philosophy, tracing back to Aristotle (Chase 2019). This structure of thought, commonly known as logic, was further developed by Immanuel Kant, who argued that our thoughts are inherently tied to a specific logical framework (Smith 2011). Kant's ideas have had a significant influence on modern linguistics, which posits that although languages vary across cultures, a universally shared underlying structure exists due to the presence of a genetically pre-programmed universal grammar (Chomsky 2014).

This perspective, highlighting the constraints on both thought and language by structures, raises questions about perceptual decision-making (PDM). PDM refers to making categorical judgments based on presented stimuli (Gold and Shadlen 2007). Suppose a particular perceptual decision corresponds to a specific proposition, which can be expressed as a sentence in a language. Then, our interpretation of the structure of sensory objects is also constrained by the structures that shape our thoughts. Consequently, how the brain generates propositions about the structure of sensory objects is closely connected to how the brain represents the structure of abstract concepts, which are essentially thoughts.

In laboratory experiments, perceptual processing is more manageable compared to the complex processing involved in abstract concepts, as brain signals related to sensory stimuli directly relate to the physical features of the stimuli. On the other hand, it remains unclear what brain signals are equivalent to abstract concepts.

Therefore, exploring how the brain processes the structure of sensory objects serves as a promising initial area of research to investigate how the brain processes the structure of concepts.

## 1.2 The Significance of Reference in the Structure of Objects

The understanding that objects have a minimum set of principles governing their structure originated nearly a century ago with the Gestalt psychologists (Wertheimer 1912, Wagemans, Elder et al. 2012). Among the fundamental relationships between objects, the relative scale holds significant importance. Numerous studies have examined how we perceive and process relative scale, with linguists particularly intrigued by its unique properties (Huttenlocher, Higgins et al. 1971, Tribushinina 2011). Linguists refer to adjectives that describe relative relationships between objects as relative adjectives, such as 'big' and 'old' (Kennedy 2007).

A relative adjective like 'tall' can have various meanings. Consider the following examples:

(1) There are many tall mountains in the Himalayas.
(2) She is young, beautiful, tall, and intelligent.
(3) He wanted to lie down in this tall grass.

The scale of 'tall' differs in each example. In (1), tall mountains would be over a thousand meters high, while in (2), a tall girl might be around 170 centimeters tall. The tallness of grass in (3) would be more than five centimeters. However, individuals have no difficulty adjusting the interpretation of 'tall' based on the context (Kamp and Partee 1995, Partee 2007). Linguists have been investigating this intriguing property of relative adjectives for several decades.

Semanticists have argued that interpreting relative adjectives involves locating a property on a gradual scale relative to a context-specific reference point (Tribushinina 2011). This reference point has been referred to as a norm (Apresjan 1974, Lehrer and Lehrer

1982, Bierwisch 1989), standard (Rotstein and Winter 2004, Maat 2006), or relative standard of comparison (Kennedy 2007). In this text, it will be referred to as a reference.

The central hypothesis regarding a reference is that it represents the average value of objects described by relative adjectives (Katz 1972, Klein 1980, Bierwisch 1989, Tribushinina 2011). This assumption implies that a reference is inherently subjective. For example, a person from Western Europe would likely have different expectations regarding average winter temperatures compared to a resident of Siberia and thus assign a different value to the adjective 'warm' in a sentence like 'This winter is surprisingly warm' (Tribushinina 2011). Additionally, the hypothesis suggests that a reference is updated with each new experience of the object denoted by the relative adjectives. If a person from Western Europe moves to Siberia, their average winter temperature expectation will align with that of Siberia. As a result, they will assign a temperature to 'warm' based on their experience in Siberia.

In summary, linguists have significantly contributed to our understanding of the relative scale, a fundamental structure between objects, by investigating relative adjectives. Their primary conclusion is that the interpretation of the relative adjectives is determined by a reference, which represents a context-specific average value of objects described by relative adjectives. This conclusion suggests that a reference is updated through new experiences as they may introduce new object samples and modify the belief regarding the expected value of objects.

## 1.3 Reference Updating as a Source of the Repulsive Bias

As previously mentioned, linguists' conclusion that our internal reference for relative adjectives is updated with new instances has piqued the interest of cognitive scientists, prompting them to investigate reference updating through laboratory experiments.

Cognitive scientists have developed computational models to

understand the process of reference setting (Treisman and Williams 1984, Lages and Treisman 1998, Lages and Treisman 2010, Dyjas, Bausenhart et al. 2012, Raviv, Lieder et al. 2014, Norton, Fleming et al. 2017, Hachen, Reinartz et al. 2021). These computational models, similar to the hypothesis proposed by linguists, suggest that the reference boundary assimilates to previous stimulus quantities. However, unlike the linguists' hypothesis, these computational models make a quantifiable prediction about boundary updating.

This prediction is known as the repulsive bias, which indicates that current choices tend to be biased away from previous stimulus values. For instance, after seeing a short tree, we are more likely to classify a tree of intermediate height as 'tall.' The repulsive bias has been observed in laboratory experiments involving both humans and non-human animals (Lages and Treisman 1998, Lages and Treisman 2010, Nakashima and Sugita 2017, Bosch, Fritsche et al. 2020, Fritsche, Spaak et al. 2020, Hachen, Reinartz et al. 2021). Thus, empirical evidence of the repulsive bias appears to support the concept of boundary updating.

However, it is important to note that the repulsive bias can also be attributed to a bias in current stimulus perception rather than a bias of the boundary itself. Sensory adaptation, a phenomenon in which previous stimulus values influence the perception of the current stimulus, has been considered a potential cause of the repulsive bias (Gibson and Radner 1937, Stocker and Simoncelli 2006, Knapen, Rolfs et al. 2010, Fritsche, Mostert et al. 2017, Nakashima and Sugita 2017, Fritsche, Spaak et al. 2020, Collins 2021, Fritsche, Solomon et al. 2022). Therefore, the mere existence of the repulsive bias does not definitively establish the empirical basis for boundary updating.

Traditionally, the neural mechanism underlying sensory adaptation has been attributed to fatigued neural populations in the sensory cortex becoming less responsive to previous stimuli (Webster and Mollon 1997, Clifford, Webster et al. 2007, Kohn 2007). This reduced neural response to previous stimuli leads to the biased perception of the current stimulus being repelled away from previous

stimuli. More recent theories propose other neural mechanisms, such as normalization, which increase the sensitivity of sensory neurons (Stocker and Simoncelli 2006, Solomon and Kohn 2014, Weber, Krishnamurthy et al. 2019, Fritsche, Spaak et al. 2020). Despite these diverse theories, a common hypothesis is that the neural mechanism of sensory adaptation involves biased neural responses in the sensory cortex.

On the other hand, the cortical region involved in boundary updating may be the associative area of the brain, as the boundary is not determined by present stimuli but is maintained by the memory system. Therefore, some researchers suggest that distinct cortical involvements in boundary updating and sensory adaptation can be useful for uncovering the true source of the repulsive bias (Hachen, Reinartz et al. 2021).

This thesis aims to provide insights to the implementation of boundary updating in the human brain and its influence on current behavior and neural responses. To address this question, I collaborated with my colleagues and conducted a series of investigations.

## 1.4 Structure of the Thesis

In Chapter 2, I aimed to investigate the origin of the repulsive bias by analyzing whole-brain imaging data, specifically focusing on the sensory and associative areas. I investigated using functional magnetic resonance imaging (fMRI) to analyze the primary visual cortex (V1) signal, as well as the whole-brain fMRI signal of human subjects. The findings revealed that the adaptation signal in V1 did not have an impact on the repulsive bias in decision-making. However, the boundary updating signal in the parietal and temporal cortices influenced the repulsive bias in decision-making. This suggests that the boundary updating process serves as the source of the repulsive bias. Importantly, the presence of neural signals related to boundary updating demonstrates the neural basis of the boundary, which plays a fundamental role in perceiving the structure of objects.

In Chapter 3, building upon the confirmation of the role of boundary updating in inducing the repulsive bias, I further explored the implications of boundary updating on decision uncertainty. While previous research primarily focused on the effect of boundary updating on choice behavior, I investigated how it also affects the variability of decision uncertainty. To assess this, I examined well-established correlates of decision uncertainty, including response time (RT) (Urai, Braun et al. 2017, Braun, Urai et al. 2018), pupil size (Urai, Braun et al. 2017), and the brain signal of the dorsal anterior cingulate cortex (dACC) (Grinband, Hirsch et al. 2006, Shenhav, Straccia et al. 2014). The results indicated that all three correlates of decision uncertainty exhibited systematic biases induced by boundary updating. This highlights the crucial role of boundaries in processing the structure of objects, which is to influence not only decision-making but also decision uncertainty.

In Chapter 4, I presented how my findings can modify previous perspectives on how the brain generates the representation of the structure of visual objects, with a specific focus on the structure of the relative scale.

# Chapter 2

# Neural evidence for boundary updating as the source of the repulsive bias in classification

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

## 2.1 Introduction

We commit to a proposition about a specific world state when making a perceptual decision. One basic form of such commitment is binary classification. It is to decide whether an item's magnitude lies on the smaller or larger side of the magnitude distribution across items of interest (Fig. 1*A*). For example, when uttering "this tree is tall" while walking in a wood, we are implicitly judging the height of that tree to be taller than the typical height of the trees in the wood (Klein 1980, Bierwisch 2009), where 'typical height' works as the

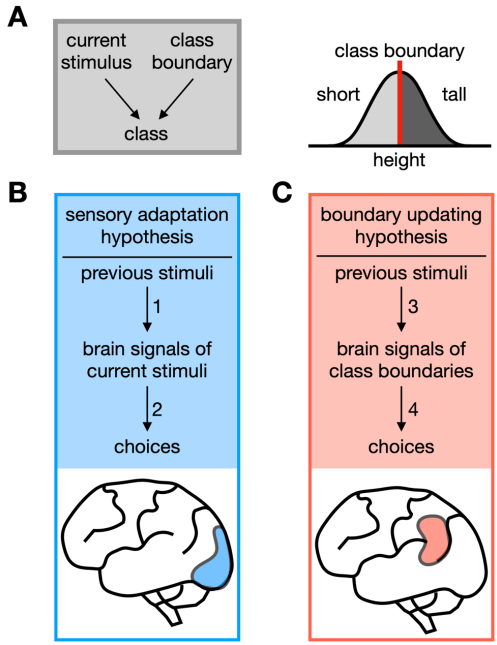


**Figure 1.** Two contending hypotheses on the origin of the repulsive bias in binary classification. *A*, Task structure (left) and statistical knowledge (right) for binary classification. For any given item, its class is determined by its position relative to the class boundary in the distribution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

boundary dividing the 'short' and 'tall' classes. Like this, binary classification is exercised in our daily language use—whenever modifying a subject with relative adjectives (Rips and Turnbull 1980, Tribushinina 2011, Solt 2015, Lassiter and Goodman 2017)—and has been adopted as an essential paradigm for studying perceptual decision-making (Lages and Treisman 1998, Grinband, Hirsch et al.

2006, Kepecs, Uchida et al. 2008, Nahum, Daikhin et al. 2010, Lak, Costa et al. 2014, Bosch, Fritsche et al. 2020, Hachen, Reinartz et al. 2021).

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

As one most straightforward scenario for repulsive bias, the previous stimuli may repel away our perception of the current stimulus from themselves because the sensory system adapts to earlier stimuli (Gibson and Radner 1937, Stocker and Simoncelli 2006, Clifford, Webster et al. 2007, Knapen, Rolfs et al. 2010, Pavan, Marotti et al. 2012, Morgan 2014, Nakashima and Sugita 2017) (Fig. 1*B*). According to this 'sensory-adaptation' hypothesis, the current tree is *biasedly* classified as 'tall' since the sensory system's adaptation to the previous short tree makes the current tree appear taller than its physical height. However, there is an alternative scenario, which considers the possibility that the internal class boundary adaptively shifts toward recent samples of property magnitude (Treisman and Williams 1984, Lages and Treisman 1998, Lages and Treisman 2010, Dyjas, Bausenhart et al. 2012, Raviv, Lieder et al. 2014, Norton, Fleming et al. 2017, Hachen, Reinartz et al. 2021) (Fig. 1*C*). According to this 'boundary-updating' hypothesis, the current tree is *biasedly* classified as 'tall' since the shift of the class boundary toward the previous short tree makes the current tree be positioned in the taller side of the boundary.

As discussed previously (Hachen, Reinartz et al. 2021), it is hard to assess which hypothesis is more viable based on behavioral data. This difficulty arises because binary classification is a matter of the relativity between the perceived stimulus and the class boundary: the identical bias in classification can be caused either by sensory-

adaptation or boundary-updating. However, the two hypotheses involve distinct neural routes through which repulsive bias transpires. The sensory-adaptation hypothesis predicts that the sensory brain signals subject to adaptation—such as those in the early sensory cortex with substantive adaptation to earlier stimuli—contribute to the choice variability. By contrast, the boundary-updating hypothesis predicts that the brain signals of the shifting boundary—such as those in the high-tier cortices involved in the working memory of previous stimuli—contribute to the choice variability.

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

## 2.2 Materials and Methods

The data of Experiment 1 (Exp1) and Experiment 2 (Exp2) were acquired from 19 (9 females, aged 20-30 years) and 18 (9 females, aged 20-30 years) participants, respectively. Among the participants, 17 of them participated in both experiments. The Research Ethics Committee of Seoul National University approved the experimental procedures. All participants gave informed consent and were naïve to the purpose of the experiments. High-spatial-resolution images were acquired only from the early visual cortex in Exp1 while the images in Exp2 were acquired from the entire brain with a conventional spatial resolution. The 17 people who provided the data for both experiments participated in three to six behavior-only

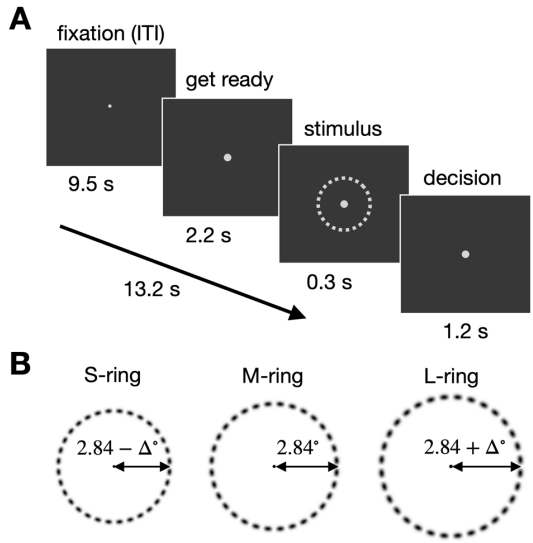


**Figure 2.** Binary classification task on ring size. ***A***, Within-trial procedure. With the eyes fixed, human participants were pre-warned (2.2s), with the increase of the fixation dot, to get ready for the upcoming trial after a long inter-trial interval (9.5s), briefly viewed the ring stimulus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

sessions for training and stimulus calibration, one fMRI session for retinotopy, and two experimental fMRI sessions (one for each experiment). The remaining people also completed the behavioral and retinotopy fMRI sessions with the same protocols but participated in only one of the two experiments.

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

## Experimental setup

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

## Behavioral data acquisition

Figure 2 illustrates the experimental procedures. On each trial, the observer initially viewed a small fixation dot (diameter in visual angle, 0.12°; luminance, 321cd/$m^2$) appearing at the center of a dark (luminance, 38cd/$m^2$) screen. A slight increase in the size of the fixation dot (from 0.12° to 0.18° in diameter), which was readily detected with foveal vision, forewarned the observer of an upcoming presentation of a test stimulus. The test stimulus was a brief (0.3s) presentation of a thin (full-width at half-maximum of a Gaussian envelope, 0.17°), white (321cd/$m^2$), dashed (radial frequency, 32cycles/360°) ring that counter-phase-flickered at 10Hz. After each presentation, participants classified the ring size into *small* or *large* using a left-hand or right-hand key, respectively, within 1.5s from stimulus onset. They were instructed to maintain strict fixation on the fixation dot throughout experimental runs. This behavioral task was performed in three different environments: i) the training sessions, ii) the practice runs of trials inside the MR scanner, and iii) the main scan runs inside the MR scanner, in the following order.

In the training sessions, participants practiced the task intensively over several (3 to 6) sessions (about 1,000 trials per session) in a dim room outside the scanner until they reached an asymptotic level of accuracy. Note that we opted to train observers with the stimuli that were much larger than those for the main experiments (mean radius of 9°) to avoid any unwanted perceptual learning effects at low sensory levels and to train participants to

learn the task structure of classification.

In the practice runs of trials inside the MR scanner, participants performed 54 practice trials and then 180 threshold-calibration trials while lying in the magnet bore. On each of the threshold-estimation trials in which consecutive trials were apart from one another by 2.7s., one of 20 different-sized rings was presented according to a multiple random staircase procedure (four randomly interleaved 1-up-2-down staircases, two starting from the easiest stimulus and the other two starting from the hardest one) with trial-to-trial feedback based on the class boundary with the radius of 2.84° . A Weibull function was fit to the psychometric curves obtained from the threshold-calibration trials using a maximum-likelihood procedure. From the fitted Weibull function, the threshold difference in size (Δ in Fig. 2*B*) associated with a 70.7% correct proportion of responses was estimated. By finding this threshold for each participant, three threshold-level ring sizes were individually tailored as 2.84−Δ° (S-ring), 2.84° (M-ring), 2.84+Δ° (L-ring).

In the main scan runs, one of these rings with threshold-level differences was presented in the order defined by an m-sequence (base = 3, power = 3; nine S and L-rings and eight M-rings were presented; all scan runs started with two M-rings) (Buracas and Boynton 2002) to null the autocorrelation between stimuli. Participants were not informed of the existence of medium-ring. Importantly, participants did not receive trial-to-trial feedback. Instead, only their run-averaged percent correct based on the trials of S-ring and L-ring was shown during a break after each run, to prevent trial-to-trial feedback from evoking any unwanted brain responses associated with rewards (Marco-Pallarés, Müller et al. 2007, Carlson, Foti et al. 2011) or errors (Carter, Braver et al. 1998, Holroyd, Nieuwenhuis et al. 2004, Cavanagh and Frank 2014). Consecutive trials were apart from one another by 13.2s. In the main scan runs of Exp1 and Exp2, observers performed 156 (6 runs X 26 trials) and 208 (8 runs X 26 trials) trials in total, respectively.

## MRI equipment and acquisition

We acquired three types of MRI images. (1) 3D, T1-weighted, whole-brain images were acquired at the beginning of each functional session: MPRAGE; resolution, 1×1×1mm; field of view (FOV), 256mm; repetition time (TR), 1.9s; time for inversion, 700ms; time to echo (TE), 2.36ms; and flip angle (FA), 9°.

(2) 2D, T1-weighted, in-plane images were acquired at the beginning of each functional session. The parameters for the retinotopy-mapping, the V1 mapping, and the whole brain mapping differed slightly as follows (retinotopy, followed by the V1 mapping, and then by the whole brain mapping): MPRAGE; resolution, 1.078x1.078x2.0 mm, 1.083x1.083x2.3 mm 1.08×1.08×3.3 mm; TR, 1.5s; T1, 700ms; TE, 2.79ms; and FA, 9°).

(3) 2D, T2*-weighted, functional images were acquired during each functional session: gradient EPI; TR, 2.7s, 2.2s, 2.2s; TE, 40ms; FA, 77°, 73°, 73°; FOV, 208mm, 207mm, 208mm; image matrix, 104x104, 90x90, 90×90; slice thickness, 1.8mm with 11% gap, 2mm with 15% slice gap, 3mm with 10% space gap; slice, 30, 22, 32 oblique transfers slices; bandwidth, 858Hz/px, 750Hz/px, 790Hz/px; and effective voxel size, 2.0x2.0x1.998mm, 2.3x2.3x2.3mm, 3.25×3.25×3.3mm).

## Retinotopy-mapping protocol

Standard traveling wave methods (Engel, Rumelhart et al. 1994, Sereno, Dale et al. 1995) were used to define V1, to estimate each participant's hemodynamic impulse response function (HIRF) of V1, and to estimate V1 voxels' receptive field center and width. High-contrast and flickering (1.33Hz) dartboard patterns were presented either as 0.89°-thick expanding or contracting rings in two scan runs, as 40°-width clockwise or counterclockwise rotating wedges in four runs or in one run as four stationary, 15°-wide wedges forming two bowties centered on the vertical and horizontal meridians. Each scanning run consisted of 9 repetitions of 27s period of stimulation. The fixation behavior during the scans was assured by monitoring participants' performance on a fixation task, in which

they had to detect any reversal in direction of a small dot rotating around the fixation.

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

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

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

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

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

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

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

Unfortunately, in a few of the sessions, functional images did not cover the entire brain. Especially, the lost part was much larger

in one participant's session than the others including the orbitofrontal cortice and posterior cerebellum. Thus, not to lose too many of voxels for analysis due to this single session, we relaxed the criterion of voxel selection a bit by including the voxels that were shared by more than 16 brains in the normalized MNI space. As a result, some voxels in the temporal pole, ventral orbitofrontal, and posterior cerebellum were excluded from data analysis.

### Estimation of the eccentricities in retinotopic space for V1 voxels

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

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

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

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

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

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

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

Fifth, the eccentricity $e$ and the other model parameters ($\sigma$, $\beta$, $b$)

were found by fitting $fMRI_p(t)$ to the predicted time-series of fMRI responses to the actual stimulation, $fMRI_o(t)$, by minimizing the residual sum of squared errors between $fMRI_p(t)$and $fMRI_o(t)$ over all time frames, $RSS$:

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

Extraction of the size-encoding signal from V1 voxels

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

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

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

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

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

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

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

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

where $\sigma_L$ and $\sigma_S$ are the tuning widths with $r_L$ and $r_S$, and the baseline offset, $\delta$, is as follows:

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

### A Bayesian model of boundary-updating (BMBU)

*The generative model.* The generative model is the observers' causal account for noisy sensory measurements, where the true ring size, $S$, causes a noisy sensory measurement on a current trial, $m_{(t)}$, which becomes noisier as $i$ trials elapse, thus turning into a noisy retrieved measurement of the value of $S$ on trial $t - i$, $r_{(t-i)}$ (Fig. 5*D*). Hence, the generative model can be specified with the following three probabilistic terms: a prior of $S$, $p(S)$, a likelihood of $S$ given $m_{(t)}$, $p(m_{(t)}|S)$, and a likelihood of $S$ given $r_{(t-i)}$, $p(r_{(t-i)}|S)$. These three terms were all modeled as normal distribution functions, the shape of which is specified with mean and standard deviation parameters, $\mu$ and $\sigma$: $\mu_0$ and $\sigma_0$ for the prior, $\mu_{m_{(t)}}$ and $\sigma_{m_{(t)}}$ for the likelihood for $m_{(t)}$, and $\mu_{r_{(t-i)}}$ and $\sigma_{r_{(t-i)}}$ for the likelihood for $r_{(t-i)}$. The mean parameters of the two likelihoods, $\mu_{m_{(t)}}$ and $\mu_{r_{(t-i)}}$, are identical to $m_{(t)}$ and $r_{(t-i)}$; therefore, the parameters that must be learned are reduced to $\mu_0$, $\sigma_0$, $\sigma_{m_{(t)}}$, and $\sigma_{r_{(t-i)}}$.

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

The primary purpose of BMBU is to build a generative Bayesian model which allows us to estimate the trial-to-trial latent states of the class boundary variable that are likely to be used by

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

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

$$p(s_{(t)}) = p(S|m_{(t)}) \propto p(m_{(t)}|S)p(S)$$

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

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

*Class boundary inference (b).* The Bayesian observer infers the value of class boundary on a current trial, $b_{(t)}$, by inferring the posterior function of a given set of retrieved sensory measurements $\vec{r}_{(t)} =$

$\{r_{(t-1)}, r_{(t-2)}, \ldots r_{(t-n)}\}$:

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

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

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

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

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

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

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

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

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

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

### Fitting the parameters of BMBU

For each human participant, the parameters of the generative model,

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

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

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

### A constant-boundary model

The constant-boundary model has two parameters, bias of class boundary $\mu_0$ and measurement noise $\sigma_m$. Stimulus estimates, $s_{(t)}$, were assumed to be sampled from a normal distribution, $\mathcal{N}(S_{(t)}, \sigma_m)$. Each stimulus sample has uncertainty $\sigma_{s_{(t)}} = \sigma_m$. Class boundary $b_{(t)}$ was assumed to be a constant, $\mu_0$; so $\sigma_{p(b_{(t)})} = \sigma_{b_{(t)}} = 0$.

### Estimation of the latent states of the variables of BMBU

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

### Recovery of the true states of the model variables

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

First, we created 256 different sets of parameter values by taking the possible combinations of the four different values of each of the four model parameters, where the four different values corresponded to the 20, 40, 60, and 80 percentiles of the parameter values fitted to the observers' choices. Second, we acquired the synthetic choices and the *true* model variables $b$, $s$, $v$, and $u$ by plugging one parameter set into BMBU and simulating it on the actual stimulus sequence presented to the observers. Third, we fitted the parameters of BMBU to the synthetic choices in the same procedure

conducted for fitting BMBU to the observed choices. Fourth, we simulated a set of the *recovered* states of the model variables using the fitted model parameters. Fifth, we calculated the R-squared between the true and the recovered variables to assess how reliably our model fitting procedure can recover the true states of the model variables. Finally, we repeated the above procedure for all the remained parameter sets and used the R-squared averaged across the 256 parameter sets as the performance measure of the recovery test.

## The multiple logistic regression model for capturing the repulsive bias

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

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

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

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

## The average marginal effect analysis

Average marginal effect (AME) was calculated by using the R-

package 'margins' (Leeper, Arnold et al. 2018). AME quantifies the average marginal effect between an ordinal dependent variable (i.e., binary choice) and an independent variable of a multiple logistic (or probit) regression model (Williams and Jorgensen 2023). To calculate the AMEs of any given variable on the current choice ($D_{(t)}$) without controlling the previous ($S_{(t-1)}$) and current stimuli ($S_{(t)}$) (i.e., the baseline AME), we implemented a logistic regression model with two regressors -the variable of interest $X$ (i.e., V1, $b$, $s$, or $v$) and the previous choice ($D_{(t-1)}$):

$$D_{(t)} \sim logit\left(\beta_0 + \beta_X X + \beta_{D_1} D_{(t-1)}\right).$$

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

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

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

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

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

We assumed that i) activity patterns of neural population for representing the latent variables are different between participants, but ii) locations and iii) timings of the activity patterns overlap across participants. Therefore, to identify the brain signals of the latent variables of BMBU in fMRI responses, the support vector regression (SVR) decoding was carried out for each human participant within

specific spatial and temporal windows.

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

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

Before training SVR, the BOLD responses in a searchlight and a target latent variable were z-scored across trials. Then, the z-scored variable was decoded for each searchlight using the cross-validation method of leave-one-run-out (8-fold cross-validation). As a result, for each searchlight and at each time point, we acquired a set of decoded latent variables in all trials. In other words, on each time point, we acquired the 4-dimensional map of the decoded variable (i.e., 3 spatial dimensions and 1 trial dimension). The 3D spatial dimensions of the decoded variables were smoothed with a 5mm FWHM Gaussian kernel on each trial.

After this subject-wise decoding analysis, we conducted the across-subject analysis to test whether the decoded variables are significantly informative. To do so, for each searchlight locus and each time point, we regressed the smoothed decoded variable onto the regression conditions of the target variable by using a generalized linear mixed effect regression model (GLMM) with a random effect of subjects. The number of regression conditions was 14, 14, and 17 for $b_{(t)}$, $s_{(t)}$, and $v_{(t)}$, respectively (Table 1). Those

regression models were deduced from the causal structure between the variables of BMBU (see the next section). We accepted a given cluster as the brain signals of $b_{(t)}$, $s_{(t)}$, or $v_{(t)}$ only when they satisfied those regression models over more than 12 contiguous searchlights. For the ROI analysis, the decoded variables were averaged over all searchlights within each ROI.

SVR was conducted using LIBSVM (http://www.csie.ntu.edu.tx/~sjlin/libsvm) with a linear kernel and constant regularization parameter of 1 (Soon, Brass et al. 2008, Kahnt, Heinzle et al. 2011). The brain imaging results were visualized using Connectome Workbench (Marcus, Harwell et al. 2011) and xjview.

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

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

The 14 regressions for the brain signal of $s_{(t)}$ (Table 1): (#1-4), $y_s$, $s$ decoded from brain signals, must be regressed positively onto $s$–the variable it represents–even when the false discovery rate is controlled (Benjamini and Hochberg 1995), and $s$ orthogonalized to $v$ or $d$ because it should reflect the variance irreducible to the offspring variables of $s$; (#5), $y_s$ must not be regressed onto $b$ because $s$ and $b$ are independent of each other ($b \nleftrightarrow s$ Fig. 5*G*); (#6,7), $y_s$ must be positively regressed onto $v$ ($s \rightarrow v$ Fig. 5*G*) but not when $v$ is orthogonalized to $s$ because the influence of $s$ on $v$ is removed; (#8,9) $y_s$ must be positively regressed onto $d$ ($s \rightarrow v \rightarrow d$ Fig. 5*G*) but not onto $u$ because $u$ cannot be linearly correlated with $s$ ($s \rightarrow v \rightarrow u$ is blocked by the interaction between $u$ and $v$ Fig. 5*G*);

| test index | $b_{(t)}$ | | | $s_{(t)}$ | | | $v_{(t)}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | regressor | threshold p-value | tail | regressor | threshold p-value | tail | regressor | threshold p-value | tail |
| 1 | $b$ | 0.001 ↓ | right | $s$ | 0.001 ↓ | right | $v$ | 0.001 ↓ | right |
| 2 | $b$ | 0.05 ↓ (fdr) | right | $s$ | 0.05 ↓ (fdr) | right | $v$ | 0.05 ↓ (fdr) | right |
| 3 | $b_{\perp v}$ | 0.05 ↓ | right | $s_{\perp v}$ | 0.05 ↓ | right | $v_{\perp b}$ | 0.05 ↓ | right |
| 4 | $b_{\perp d}$ | 0.05 ↓ | right | $s_{\perp d}$ | 0.05 ↓ | right | $v_{\perp s}$ | 0.05 ↓ | right |
| 5 | $s$ | 0.05 ↑ | both | $b$ | 0.05 ↑ | both | $v_{\perp d}$ | 0.05 ↓ | right |
| 6 | $v$ | 0.05 ↓ | left | $v$ | 0.05 ↓ | right | $b$ | 0.05 ↓ | left |
| 7 | $v_{\perp b}$ | 0.05 ↑ | left | $v_{\perp s}$ | 0.05 ↑ | right | $b_{\perp v}$ | 0.05 ↑ | left |
| 8 | $d$ | 0.05 ↓ | left | $d$ | 0.05 ↓ | right | $s$ | 0.05 ↓ | right |
| 9 | $u$ | 0.05 ↑ | both | $u$ | 0.05 ↑ | both | $s_{\perp v}$ | 0.05 ↑ | right |
| 10 | $S_{(t)}$ | 0.05 ↑ | both | $S_{(t)}$ | 0.05 ↓ | right | $d$ | 0.05 ↓ | right |
| 11 | $S_{(t-1)}$ | 0.05 ↓ | right | $S_{(t-1)}$ | 0.05 ↑ | both | $u$ | 0.05 ↑ | both |
| 12 | $S_{(t-2)}$ | 0.05 ↑ | left | $S_{(t-2)}$ | 0.05 ↑ | both | $S_{(t)}$ | 0.05 ↓ | right |
| 13 | $D_{(t-1)}$ | 0.05 ↑ | both | $D_{(t-1)}$ | 0.05 ↑ | both | $S_{(t-1)}$ | 0.05 ↓ | left |
| 14 | $D_{(t-2)}$ | 0.05 ↑ | both | $D_{(t-2)}$ | 0.05 ↑ | both | $S_{(t-2)}$ | 0.05 ↑ | right |
| 15 | | | | | | | $D_{(t)}$ | 0.05 ↓ | right |
| 16 | | | | | | | $D_{(t-1)}$ | 0.05 ↑ | both |
| 17 | | | | | | | $D_{(t-2)}$ | 0.05 ↑ | both |

**Table 1.** The sets of regressions that BMBU requires the brain signals of its latent variables to satisfy. The regressions required for the brain signal of the inferred class boundary ($b_{(t)}$; left sector), the inferred stimulus ($s_{(t)}$; middle sector), and the decision variable ($v_{(t)}$; right sector). The top sector (#1~#9 for $b_{(t)}$; ; #1~#9 for $s_{(t)}$; #1~#11 for $v_{(t)}$) specifies the individual, simple regression models in which the brain signal of interest is regressed on a single regressor (second column). Any regressor subscripted with another variable with the perpendicular symbol (e.g., $b_{\perp v}$) means that the residuals of the left-side variable (e.g., $b$) from the regression of the right-side variable with the perpendicular symbol (e.g., $v$) were used as the regressor. This regression with the residual regressor was created to check whether the brain variable of interest has a unique covariation with the original

regressor by withholding the influence of the perpendiculared variable (e.g., $pSTG_{b5}$ must be positively correlated with $b$ even when the part of $b$'s variability associated with $v$ is withheld). The bottom sector of each table (#10~#14 for $b_{(t)}$; #10~#14 for $s_{(t)}$; #12~#17 for $v_{(t)}$) specifies the multiple-regression model in which the brain signal of interest is regressed concurrently on the current and previous stimuli and the past or current choices. The third and fourth column of each table specify the statistical criteria used for significance test, where fdr indicates a multiple comparison test controlling the false discovery rate.

(#10-12), $y_s$ must be positively regressed onto the current stimuli and not the past stimuli because $s$ is inferred solely from the current stimulus measurement; (#13,14), $y_s$ must not be regressed onto previous decisions because $s$ is inferred solely from the current stimulus measurement. #10-14 were investigated by a multiple regression with regressors $[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t-1)}, D_{(t-2)}]$. We did not include $D_{(t)}$ as a regressor because $D_{(t)}$ may induce a spurious correlation between $b$ and $s$ by controlling the collider $v$ (Elwert and Winship 2014) ($b \rightarrow v \leftarrow s$ and $v \rightarrow d$ Fig. 5*G*).

The 17 regressions for the brain signal of $v_{(t)}$ (Table 1). (#1-5), $y_v$, $v$ decoded from brain signals, must be positively regressed onto $v$–the variable it represents–even when the false discovery rate is controlled (Benjamini and Hochberg 1995), and $v$ orthogonalized to $b, s,$ or $d$, because it should reflect the variance irreducible to the offspring variables of $v$; (#6,7), $y_v$ must be negatively regressed onto one of its parents $b$ ($b \rightarrow v$ Fig. 5*G*), but not when $b$ is orthogonalized to $v$, because the influence of $b$ on $v$ is removed; (#8,9), $y_v$ must be positively regressed onto one of another parent $s$ ($s \rightarrow v$ Fig. 5*G*), but not when $s$ is orthogonalized to $v$, because the influence of $s$ on $v$ is removed; (#10,11), $y_v$ must be regressed onto $d$ but not onto $u$ because $u$'s correlation with its parent $v$ cannot be revealed without holding the variability of $d$ (the interaction between $u$ and $v$); (#12-14), $y_v$ must be positively regressed onto the current stimulus because the influence of the current stimulus on $v$ is propagated via $s$ ($S_{(t)} \rightarrow s \rightarrow v$), and negatively regressed onto the past stimuli because the influence of the past stimuli on $v$ is propagated via $b$ ($S_{(t-1)} \rightarrow b \rightarrow v$) –strongly

onto the 1-back stimulus and more weakly onto the 2-back stimulus (thus, non-significant regression with one-tailed regression in the opposite sign is modeled moderately); (#15-17), $y_v$ must be regressed onto the current decision and not the past decisions because the current decision is a dichotomous translation of $v$ ($v \rightarrow d$ Fig. 5*G*), whereas past decisions have nothing to do with the current state of $v$. #12-17 were investigated by a multiple regression with regressors $[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t)}, D_{(t-1)}, D_{(t-2)}]$. $D_{(t)}$ was included as a regressor because $v$ does not suffer from a spurious correlation that arises by controlling a collider variable which is absent in this case.

## Bayesian network analysis

To investigate whether the relationship between decoded $b$, $s$, and $v$ is consistent with the causal structure postulated by BMBU, we calculated the BIC values for all the three-node networks consisting of the time series of three brain signals $\{y_b, y_s, y_v\}$ (Scutari 2009) and determined the causal graph whose likelihood is maximal. The three-node network has 162 possible structures, as follows. A total of 27 edge structures can be created out of three nodes since three types of edges are possible for any given pair of nodes (i.e., $x \rightarrow y$, $x \leftarrow y$ or $x \nleftrightarrow y$) and there are three pairs (i.e., $\{b, v\}$, $\{v, s\}$, $\{s, b\}$; $3^3$). Also, a total of 6 combinations of three nodes exist for $\{y_c, y_s, y_v\}$ since we have three ($IPL_{b1}$, $pSTG_{b3}$, $pSTG_{b5}$), two ($DLPFC_{s3}$, $Cereb_{s5}$), and single ($aSTG_{v5}$) brain signals of $b$, $s$, and $v$, respectively ($3 \times 2 \times 1$). Thus, because each of the 6 possible node combinations can have 27 edge structures, there are 162 possible three-node causal networks.

We opted to apply this Bayesian network analysis to the three-node networks instead of the six-node network consisting of all the six brain signals identified by the searchlight analysis because the number of possible six-node networks ($N = 3^{6C2} = 3^{15} = 14{,}348{,}907$) was unrealistically large so that the statistical results are likely to suffer from type I errors. In addition, guided by BMBU, we were interested in identifying the causal structure of the three brain signals, each corresponding to one of the three model variables ($b$, $s$,

and $v$). In other words, we were not interested in the causal relationship between the brain signals representing the same model variable (e.g., between $pSTG_{b3}$, $pSTG_{b5}$).

### Statistics

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

## 2.3 Results

### Experimental paradigm

Over consecutive trials, participants sorted ring sizes into two classes, *small* and *large*, under moderate time pressure (Fig. 2*A*). To ensure decision-makings with uncertainty, we presented three rings (small, medium, and large) differing by a threshold size ($\Delta$), which was tailored for individuals (Fig. 2*B*; see Materials and Methods). The ring sizes were presented in m-sequence to rule out any

correlation between consecutive stimulus sizes (Buracas and Boynton 2002). We provided participants with feedback after each scan run by summarizing their performance with the proportion of correct trials.

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

### Repulsive bias in Experiment 1

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

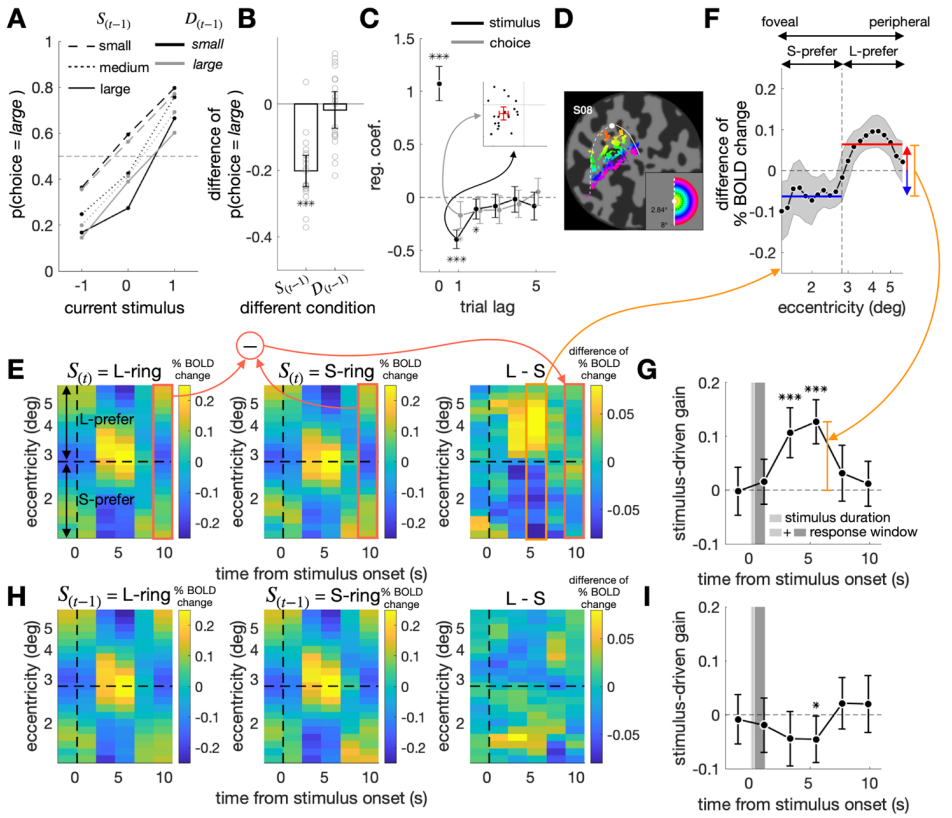


**Figure 3.** Influences of previous and current stimuli on classification behavior and V1 activity in Experiment 1. ***A*–*C***, Repulsive bias in psychometric curves (***A***,***B***) and regression analysis (***C***). The psychometric curves, where the fractions of *large* choices are plotted against the current stimulus, are shown separately for the six possible combinations defined by the previous stimulus and choice (***A***). As the summary of the effects of the previous stimulus on the current choice, the differences in the fractions of *large* choices between the previous stimuli were L-ring and S-ring ($p(D_{(t)} = large|S_{(t-1)} = 1) - p(D_{(t)} = large|S_{(t-1)} = -1)$) are computed separately for the six combinations of the current stimulus and previous choice and then averaged (***B***, left). As the summary of the effects of the previous choice on the current choice, the differences in the fractions of *large* choices between the previous choices were *large* and *small* ($p(D_{(t)} = large|D_{(t-1)} = large) - p(D_{(t)} = large|D_{(t-1)} = small)$) are computed separately for the nine combinations of the current and previous stimuli and then averaged (***B***, right). The small gray circles represent the individual observers. The multiple logistic regression coefficients of the current choice are plotted against trial lags (**C**). In the inset, the regression coefficients for the previous-stimulus ($S_{(t-1)}$) regressor are plotted against those for the previous-choice ($D_{(t-1)}$) regressor for individual observers, where the red error bars demarcate the 95% CIs of the means. ***D***, Eccentricity map of V1 on the flattened left occipital cortex of a representative brain, S08. The dot, curves, and colors correspond to

those in the inset depicting the visual field. The image is borrowed from our previous work (Choe, Blake et al. 2014). *E*,*H*, Spatiotemporal BOLD V1 responses to L-ring (left) and S-ring (middle), and their differentials (right), presented on the current (*E*) and previous (*H*) trials. The color bars indicate BOLD changes in the unit of % signal, averaged across all participants. The vertical dashed line marks the time point for stimulus onset. The horizontal dashed line corresponds to the eccentricity of M-ring, splitting the voxels into 'L-prefer' and 'S-prefer' groups based on their preferred ring size. *F*, The differential of BOLD responses at peak between the small and large ring on the current trial. The vertical dashed line marks the eccentricity of M-ring. The horizontal red and blue lines mark the average BOLD signals of the L-prefer and S-prefer voxels, respectively. The vertical orange line quantifies the stimulus-driven gain of V1 responses. *G*,*I*, Time courses of the stimulus-driven gain of V1 responses to the current (*G*) and previous (*I*) stimuli. The stimulus duration and response window are demarcated by the light and dark gray bars demarcate (*G*,*I*). The 95% CIs of the mean across observers are indicated by the shaded areas (*F*) or by the vertical error bars (*B*,*C*,*G*,*I*). Asterisks indicate the statistical significance (*, $P < 0.05$; **, $P < 10^{-3}$; ***, $P < 10^{-4}$; *B*,*C*,*G*,*I*). The orange boxes and arrows are drawn to help the relationships between the panels (*E*,*F*,*G*).

PL difference ($-0.018$ ) did not significantly differ from zero ($t_{(18)} = -0.68, p = 0.50$) (Fig. 3*B*, right).

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

## Sensory adaptation in V1

As a first step toward the verification of the sensory-adaptation hypothesis, we defined the size-encoding signal in V1. As our group showed previously (Choe, Blake et al. 2014), the eccentricity-tuned BOLD responses in V1 (Fig. 3*D*) readily resolved the threshold-level differences in ring size, as anticipated by the retinotopic organization of the V1 architecture (Fig. 3*E*). Thus, the subtraction of the BOLD

responses at the voxels preferring S-ring to L-ring from those at the voxels preferring L-ring to S-ring (Fig. 3*F*) was significantly greater when $S_{(t)}$ was large than when small (the third and the fourth time points, $\beta = 0.11,\ t_{(18)} = 4.8,\ p = 1.5 \times 10^{-4}$ and $\beta = 0.13,\ t_{(18)} = 6.6,\ p = 3.7 \times 10^{-6}$; Fig. 3*G*).

Next, having defined the size-encoding signal in V1, which will be referred to as '$V1$', we sought evidence of sensory-adaptation in that signal. According to the previous work on sensory-adaptation (Clifford, Webster et al. 2007, Kohn 2007, Solomon and Kohn 2014, Weber, Krishnamurthy et al. 2019), we expected $V1$ to decrease following the large size and to increase following the small size due to the selective gain reduction at the sensory neurons tuned to previous stimuli. In line with this expectation, $V1$ indeed significantly decreased when preceded by L-ring than when preceded by S-ring (the fourth time point, $\beta = -0.45,\ t_{(18)} = -2.2,\ p = 0.040$; Fig. 3*H,I*). Although we rendered ineffective the autocorrelation between consecutive stimuli using an m-sequence (see Materials and Methods), we additionally checked the possibility that the observed adaptation of $V1$ might have spuriously occurred due to any imbalance in the ring size of the current stimuli. To do so, we first calculated the differences in $V1$ between the previous S- and L-rings separately for the three current stimuli and then averaged those three differences. We confirmed that the averaged $V1$ differences were smaller when preceded by L-ring than when preceded by S-ring (the fourth time point, $\beta = -0.44,\ t_{(18)} = -2.1,\ p = 0.049$).

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

## The variability of $V1$ associated with previous stimuli fails to contribute to the choice variability

Next, we verified the critical prediction of the sensory-adaptation hypothesis on repulsive bias. Below, we will define what this crucial prediction is and how we empirically examine that prediction.

Above, we confirmed that the ring size, not only on the

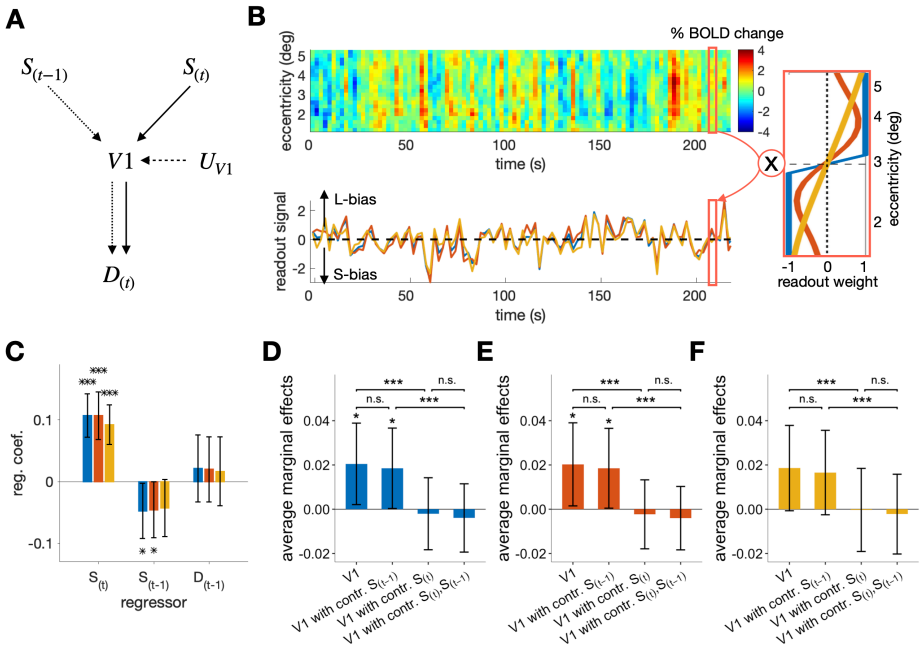


**Figure 4**. Origin of the covariation between the stimulus-encoding signal of V1 and the current choice. ***A***, The causal structure of the variables implied by the sensory-adaptation hypothesis. The stimulus-encoding signal of V1 ($V1$) is influenced by the current stimulus ($S_{(t)}$), the previous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

the left), respectively. Asterisks indicate the statistical significance (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$), and "n.s." stands for the non-significance of the test (*C-F*). The 95% CIs of the mean across participants are indicated by the vertical error bars (*C-F*).

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

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

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

The AME of $V1$ on $D_{(t)}$ was significant across participants ($\beta = 0.020, t_{(18)} = 2.3, p = 0.031$; Fig. 4*D*, the first bar). Importantly, it did not significantly decrease across participants when the influence of $S_{(t-1)}$ was controlled ($t_{(18)} = -1.6, p = 0.13$; Fig. 4*D*, the change of the first to second bars). Given the significant repulsive bias associated with $S_{(t-1)}$ presented on the 2-back trial, we also controlled $S_{(t-2)}$ in addition to $S_{(t-1)}$. Despite this additional control, the AME of $V1$ on $D_{(t)}$ did not significantly decrease ($t_{(18)} = -1.5,\ p = 0.15$). By contrast, the AME of $V1$ on $D_{(t)}$ substantially decreased across participants, almost to none, when the influence of $S_{(t)}$ was controlled ($t_{(18)} = -6.0,\ p = 1.1 \times 10^{-5}$; Fig. 4*D*, the change of the first to third bars). Likewise, the AME of $V1$ on $D_{(t)}$ with $S_{(t)}$ only controlled did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 1.4,\ p = 0.17$; Fig. 4*D*, the change of the fourth to third bars), whereas the AME of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ controlled was greater than that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 6.02,\ p = 1.1 \times 10^{-5}$; Fig. 4*D*, the change of the fourth to second bars). These results coherently indicate that the contribution of the previous stimuli to $D_{(t)}$ via $V1$ is absent or negligible, which is at odds with the sensory-adaptation hypothesis. The analyses above were carried out for $V1$ acquired at the fourth time point, where sensory adaptation was significant. However, an insignificant but substantial amount of sensory adaption occurred also at the preceding (third) time point (Fig. 3*I*). To check the possibility that the contribution of $S_{(t-1)}$ to $D_{(t)}$ via $V1$ might be

present if $V1$ is alternatively defined, we redefined $V1$ by averaging those acquired at the third and fourth points and repeated the same AME analyses as above. However, the contribution of the previous stimuli to $D_{(t)}$ via $V1$ is still absent or negligible: the AME of $V1$ on $D_{(t)}$ did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ controlled ($t_{(18)} = -1.4,\ p = 0.19$); the AME of $V1$ on $D_{(t)}$ with $S_{(t)}$ only controlled did not differ from that of $V1$ on $D_{(t)}$ with $S_{(t-1)}$ and $S_{(t)}$ both controlled ($t_{(18)} = 1.03,\ p = 0.32$).

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

### Repulsive bias in Experiment 2

Having failed to find the evidence supporting the sensory-adaptation hypothesis in Experiment 1, we conducted Experiment 2 to search the whole brain for the signal representing the class boundary and to test whether that signal relates to the previous stimuli and the

current choice in a manner consistent with the boundary-updating hypothesis. As mentioned earlier (see the first Results subsection), the experimental procedure in Experiment 2 was the same as in Experiment 1, except for the fMRI protocol.

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

### Bayesian model of boundary-updating (BMBU)

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

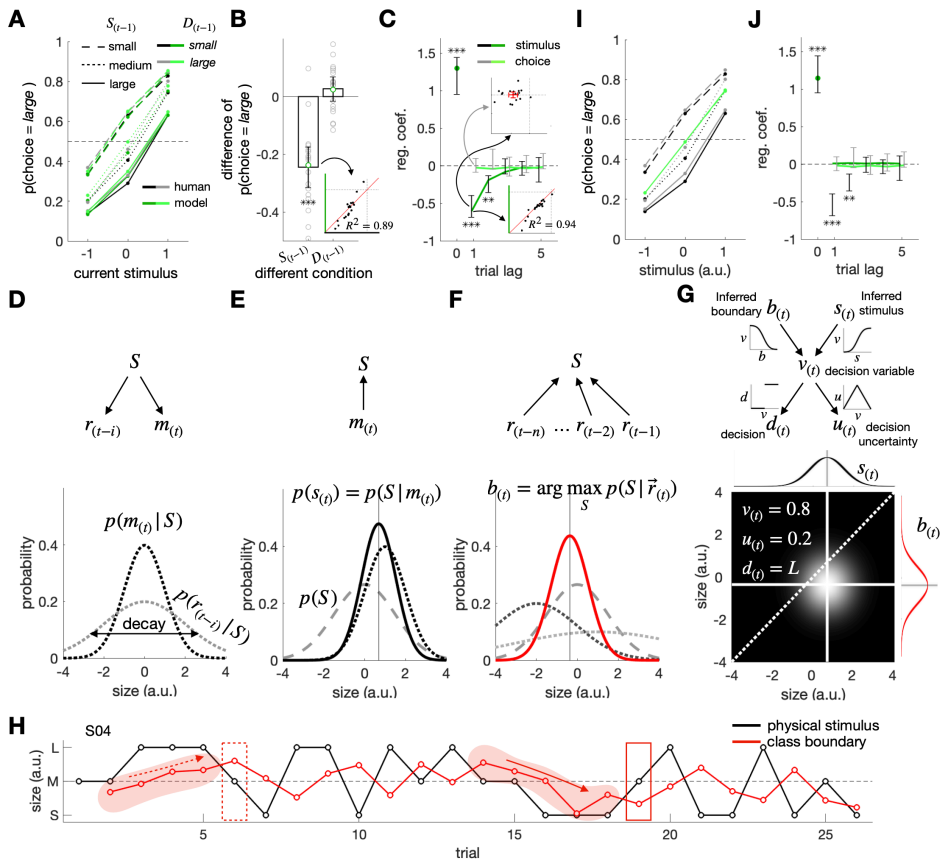


**Figure 5.** Repulsive bias in Experiment 2 and a Bayesian model of boundary updating (BMBU). ***A***–***C***, Repulsive bias in psychometric curves (***A***,***B***) and regression analysis (***C***). The formats were identical to those in the corresponding figure panels for Experiment 1 (**Fig. 3*A*–*C***), except that the *ex post* model simulation results (green lines and symbols) are added. In the bottom insets of ***B***, the observed (x-axis) and simulated (y-axis) average differences in the fractions of *large* choices between the trials in which the previous stimulus was L-ring and those in which it was S-ring are plotted against one another, where the red diagonal demarcates the identity line. In the bottom insets of ***C***, the observed (x-axis) and simulated (y-axis) regression coefficients for the previous stimulus ($S_{(t-1)}$) regressor are plotted against one another for individual observers, where the red diagonal demarcates the identity line. ***D***–***G***, The measurement generation (***C***), stimulus inference (***D***), class-boundary inference (***E***), and decision-variable deduction (***F***) processes of BMBU. BMBU posits that the Bayesian decision-maker has an internal causal model of how a physical stimulus size ($S$) engenders a current sensory measurement ($m_{(t)}$) and a retrieved memory measurement from $i$th preceding trial ($r_{(t-i)}$) (***D***, top), which specifies the probability distribution of $m_{(t)}$ and $r_{(t-i)}$ conditioned on $S$, respectively (***D***, bottom). In turn, $p(m_{(t)}|S)$ allows the Bayesian decision-maker to infer $S$ upon observing $m_{(t)}$ by combining

it with the prior knowledge about $S$, $p(S)$, to compute the posterior probability of $S$ given $m_{(t)}$, $p(S|m_{(t)})$ (*E*). Similarly, $p(r_{(t-i)}|S)$ allows for inferring the class boundary ($b_{(t)}$) upon retrieving the memory of previous sensory measurements ($\vec{r}_{(t)} = [r_{(t-1)}, r_{(t-2)}, \ldots]$) by combining it with $p(S)$ to compute the posterior probability of $S$ given $\vec{r}_{(t)}$, $p(S|\vec{r}_{(t)})$ (*F*). In *D*–*F*, black dotted curves, $p(m_{(t)}|S)$; gray dotted curves, $p(r_{(t-i)}|S)$—the darker the dotted curve is, the more recent the memory is; gray dashed curves, $p(S)$; black solid curve, $p(S|m_{(t)})$; red solid curve, $p(S|\vec{r}_{(t)})$. Finally, the inferred stimulus, $s_{(t)}$, and the inferred class boundary, $b_{(t)}$, allow for deducing the decision variable, $v_{(t)}$, the choice variable, $d_{(t)}$, and the uncertainty variable, $u_{(t)}$ (*G*, top), as illustrated in an example bivariate distribution of $s_{(t)}$ and $b_{(t)}$, from which $v_{(t)}$, $d_{(t)}$ and $u_{(t)}$ are derived (*G*, bottom). *H*, An example temporal trajectory of the class boundary inferred by BMBU in a single scan run. The black and red lines indicate the sizes of physical stimulus and the boundary inferred by BMBU, respectively. *I*,*J*, *Ex post* simulation results of the constant-boundary model. The formats are identical to those of A and B.

representing signal using a searchlight multivariate pattern analysis technique.

We developed BMBU by formalizing the binary classification task in terms of Bayesian decision theory (Knill and Richards 1996), a powerful framework for modeling human decision-making behavior under uncertainty. Binary classification is to judge whether the 'ring size on the current trial $t$ ($S_{(t)}$)' is larger or smaller than the 'the typical size of rings appearing across the entire trials ($\tilde{S}$).' Therefore, a classifier must infer them based on the measurements of stimulus size in the sensory and memory systems.

*The generative model*. On trial $t$, $S_{(t)}$ is randomly sampled from a probability distribution $p(S)$ and engenders a measurement in the sensory system $m_{(t)}$, which is a random sample from a probability distribution $p(m_{(t)}|S_{(t)})$ (black dotted curve of Fig. 5*D*, bottom). Critically, as $i$ trials elapse, $m_{(t)}$ is re-encoded into a mnemonic measurement in the working-memory system $r_{(t-i)}$, which is a random sample from a probability distribution $p(r_{(t-i)}|S_{(t)})$ (light-gray dotted curve in Fig. 5*D*, bottom). Here, we assumed that the width of $p(r_{(t-i)}|S_{(t)})$ increases as $i$ increases reflecting the working memory decay (Gorgoraptis, Catalao et al. 2011, Zokaei, Burnett

Heyes et al. 2015).

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

$$p(s_{(t)}) = p(S|m_{(t)}) \qquad \text{(Equation 1)}$$

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

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

$$b_{(t)} = \arg\max_{S} p(S|\vec{r}_{(t)}) \qquad \text{(Equation 2)}$$

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

*Making a decision with the inferred current stimulus size and the inferred class boundary*. Having estimated $s_{(t)}$ and $b_{(t)}$, the Bayesian classifier deduces a decision variable ($v_{(t)}$) from $s_{(t)}$ and $b_{(t)}$ and translating it into a binary decision ($d_{(t)}$) with a degree of uncertainty ($u_{(t)}$) (Fig. 5*G*). Here, $v_{(t)}$ is the probability that $s_{(t)}$ will be greater than $b_{(t)}$ ($v_{(t)} = p(s_{(t)} > b_{(t)})$); $d_{(t)}$ 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 ($u_{(t)} = p(s_{(t)} < b_{(t)}|d_{(t)} = large)$ or $p(s_{(t)} > b_{(t)}|d_{(t)} = small)$) (Sanders, Hangya et al. 2016).

In sum, BMBU models a human decision-maker as the Bayesian classifier who, over consecutive trials, continuously infers the class boundary ($b$) and the current stimulus size ($s$), deduces the decision variable ($v$) from $s$ and $b$, and makes a decision ($d$) with a

varying degree of uncertainty ($u$). As shown below, BMBU well predicts human participants' choices and reproduces their repulsive bias.

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

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

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

After equipping the models with their best-fit parameters, we assessed their reproducibility by making them simulate the decisions over the same sequence of ring sizes presented to the human participants (see Materials and Methods). From this simulation, we can also vividly appreciate how BMBU updates its class boundary ($b_{(t)}$) depending on the ring sizes encountered over a sequence of classification trials (Fig. 5*H*). As implied by Equation 2, BMBU continuously shifts $b_{(t)}$ toward the ring sizes shown in previous trials. Such attractive shifts are pronounced especially when streaks of S-ring (the solid arrow in Fig. 5*H*) or L-ring (the dashed arrow in Fig. 5*H*) appeared over trials. Importantly, we confirmed that such boundary-updating of BMBU reproduces the repulsive bias displayed by the human participants with a remarkable level of resemblance across participants, both for the psychometric curves (the R-squared of the effect of previous stimulus on PL between humans and BMBU

was 0.89; Fig. 5*A*,*B*) and for the coefficients of the stimulus and choice regressors (the R-squared of coefficients of the immediately preceding stimulus between humans and BMBU was 0.94; Fig. 5*C*). None of the simulated PLs and coefficients—a total of 17 points—fell outside the 95% confidence intervals of the corresponding human PLs and coefficients. Not surprisingly, the constant-boundary model failed to show any slightest hint of repulsive bias (Fig. 5*I*,*J*). Although we used m-sequences to prevent any auto-correlation among ring sizes, the failure of the constant-boundary model in reproducing repulsive bias reassures that the actual stimulus sequences used in the experiment do not contain any unwanted statistics that might induce spurious kinds of repulsive bias.

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

### Brain signals of the class boundary and the other latent variables

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

Then, with the trial-to-trial states of the simulated latent variables, we identified the brain signals of those variables with the

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

We incorporated these two points in our search of the brain signals of $b$, $s$ and $v$, as follows. Initially, we identified the candidate brain signals of $b$, $s$, and $v$ by localizing the patterns of activities that closely reflect the trial-to-trial states of $b$, $s$, and $v$. For localization, we used the support vector regressor decoding with the searchlight technique (Kahnt, Grueschow et al. 2011, Hebart, Schriever et al. 2014), which is highly effective in detecting the local patterns of population fMRI responses associated with the latent variables of computational models (Kriegeskorte, Goebel et al. 2006). Next, we put those candidate brain signals to a strong test of whether their trial-to-trial states satisfy the causal relationships with the other variables. Specifically, we converted those causal relationships into the empirically testable sets of regression models (Table 1), respectively for $b$ (14 regressions), $s$ (14 regressions) and $v$ (17 regressions) and checked whether all the regressors' coefficients derived from the brain signals were consistent with the regression models (see Materials and Methods). In what follows, we will describe how the regression tests for the brain signal of $b$ ($y_b$) were derived from the causal structure of the variables defined by BMBU (see Materials and Methods for those for the two remaining variables $s$ and $v$).

According to the causal relationship of $b$ with the latent

| name | cortical area | decoded variable | detected time from stimulus onset (s) | contiguous searchlights number | peak searchlight | |
|---|---|---|---|---|---|---|
| | | | | | MNI coordinate | GLMM p-value (right-tailed) |
| $IPL_{b1}$ | left inferior parietal lobe | $b_{(t)}$ | 1.1 | 15 (10) | $[-54,-27,48]$ ($[-54,-27,48]$) | $8.8\times10^{-8}$ ($7.4\times10^{-7}$) |
| $pSTG_{b3}$ | left posterior superior temporal gyrus | $b_{(t)}$ | 3.3 | 13 (7) | $[-45,-30,9]$ ($[-54,-27,12]$) | $5.8\times10^{-7}$ ($8.6\times10^{-8}$) |
| $pSTG_{b5}$ | left posterior superior temporal gyrus | $b_{(t)}$ | 5.5 | 18 (14) | $[-66,-21,9]$ ($[-66,-21,9]$) | $2.3\times10^{-7}$ ($2.6\times10^{-7}$) |
| $DLPFC_{s3}$ | left dorsolateral prefrontal cortex | $s_{(t)}$ | 3.3 | 33 (37) | $[-51,27,24]$ ($[-51,27,24]$) | $1.9\times10^{-7}$ ($7.7\times10^{-6}$) |
| $Cereb_{s5}$ | right cerebellum | $s_{(t)}$ | 5.5 | 36 (19) | $[36,-63,-21]$ ($[36,-69,-18]$) | $1.4\times10^{-6}$ ($5.0\times10^{-7}$) |
| $aSTG_{v5}$ | left anterior superior temporal gyrus | $v_{(t)}$ | 5.5 | 15 (4) | $[-60,-9,15]$ ($[-60,-9,15]$) | $4.9\times10^{-8}$ ($3.7\times10^{-7}$) |

**Table 2.** Specification of the brain signals of the latent variables of BMBU. The results outside of the parentheses indicate the main result obtained by using the searchlight composed of 123 voxels. The values inside of the parentheses are the results calculated by using different size of searchlight (87 voxels).

variables, $y_b$ must satisfy the following single linear regression models: $y_b$ must be positively regressed onto $b$ (#1) and be so even when the false discovery rate (Benjamini and Hochberg 1995) is applied (#2); $y_b$ must be positively regressed onto $b$ even when $b$ is orthogonalized to $v$ (#3) or $d$ (#4) because $y_b$ should reflect the variance irreducible to the offspring variables of $b$; $y_b$ must not be regressed onto $s$ because $b$ and $s$ are independent of one another

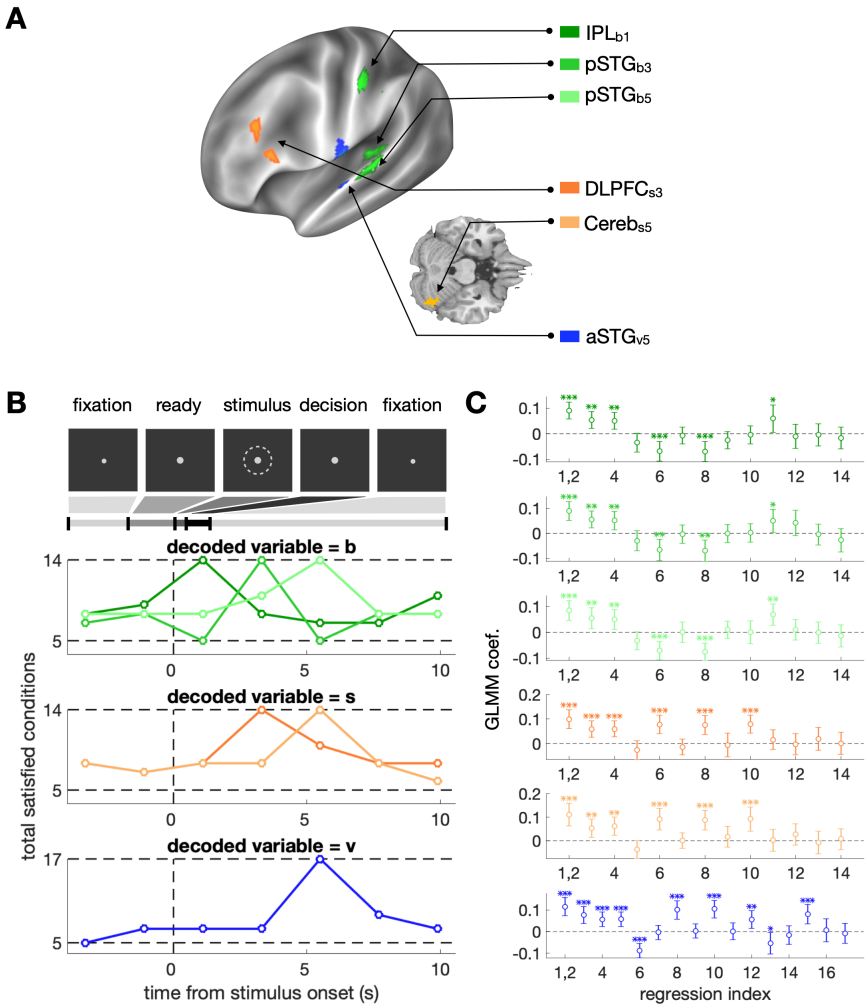


**Figure 6.** Brain signals of the latent variables of BMBU. *A*, Loci of the brain signals. The brain regions where BOLD activity patterns satisfied all the regressions implied by the causal structure of the variables in BMBU are overlaid on the inflated cortex and the axial view of the cerebellum of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

($b \nleftrightarrow s$ Fig. 5*G*; #5); $y_b$ must be negatively regressed onto $v$ ($b \rightarrow v$ Fig. 5*G*; #6) but not when $v$ is orthogonalized to $b$ because such

orthogonalization removes the influence of $b$ on $v$ (#7); $y_b$ must be negatively regressed onto $d$ ($b \rightarrow v \rightarrow d$ Fig. 5*G*; #8) but not onto $u$ because $u$ is not linearly correlated with $b$ ($b \rightarrow v \rightarrow u$ is blocked by the non-linear relationship between $u$ and $v$ Fig. 5*G*; #9). In addition, according to the causal relationship of the latent variables with the stimuli and choices (Fig. 5*D*–*G*), $y_b$ must satisfy the following multiple linear regression model defined by the observable variables $[S_{(t)}, S_{(t-1)}, S_{(t-2)}, D_{(t-1)}, D_{(t-2)}]$: $y_b$ must not be regressed onto the current stimulus (#10) because $b$ is independent of $S_{(t)}$; $y_b$ must be positively regressed onto the 1-back stimulus for sure (#11) because $b$ firmly shifts toward $S_{(t-1)}$; the regression of $y_b$ onto the 2-back stimulus must be weaker than that onto the 1-back stimulus (#12) due to memory decay (Fig. 5*D*) (accordingly, the sign of the regression coefficient of $S_{(t-2)}$ was defined as the complementary part of that of $S_{(t-1)}$); $y_b$ must not be regressed onto previous decisions because previous decisions do not have any influence on $b$ (#13,14). We did not include $D_{(t)}$ as a regressor in the multiple regression because $D_{(t)}$ may induce a spurious correlation between $b$ and $s$ by controlling the collider (common offspring) variable $v$ (Elwert and Winship 2014) ($b \rightarrow v \leftarrow s$ Fig. 5*G*) via its relationship with $v$ ($v \rightarrow d$ Fig. 5*G*).

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

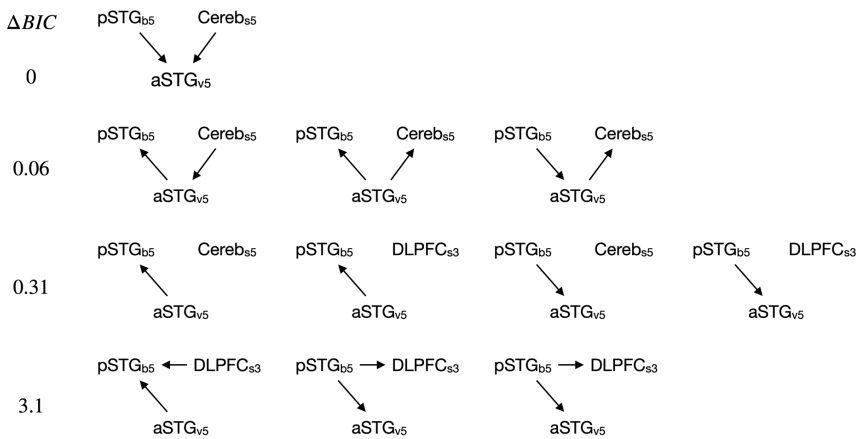


**Figure 7.** The probable causal structures between the brain signals of the latent variables in BMBU. For each row, the value in the left indicates the relative BIC scores of the causal structures in reference to the most probable one at the top.

searchlight size, we could detect the clusters that survived all regression tests around the six regions (Table 2).

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

The outcomes of BIC evaluation were consistent with BMBU. First, out of the 162 possible causal graphs, the smallest (best) BIC value was found for 'pSTG$_{b5}$→aSTG$_{v5}$ ←Cereb$_{s5}$' (Fig. 7). Second, We found that any graph with the causal arrows between $b_{(t)}$ and $s_{(t)}$ is significantly less likely than the best causal graph (BIC>2; shown at the bottom of Fig. 7) (Kass and Raftery 1995). The results indicate that the relationship between the identified brain signals faithfully

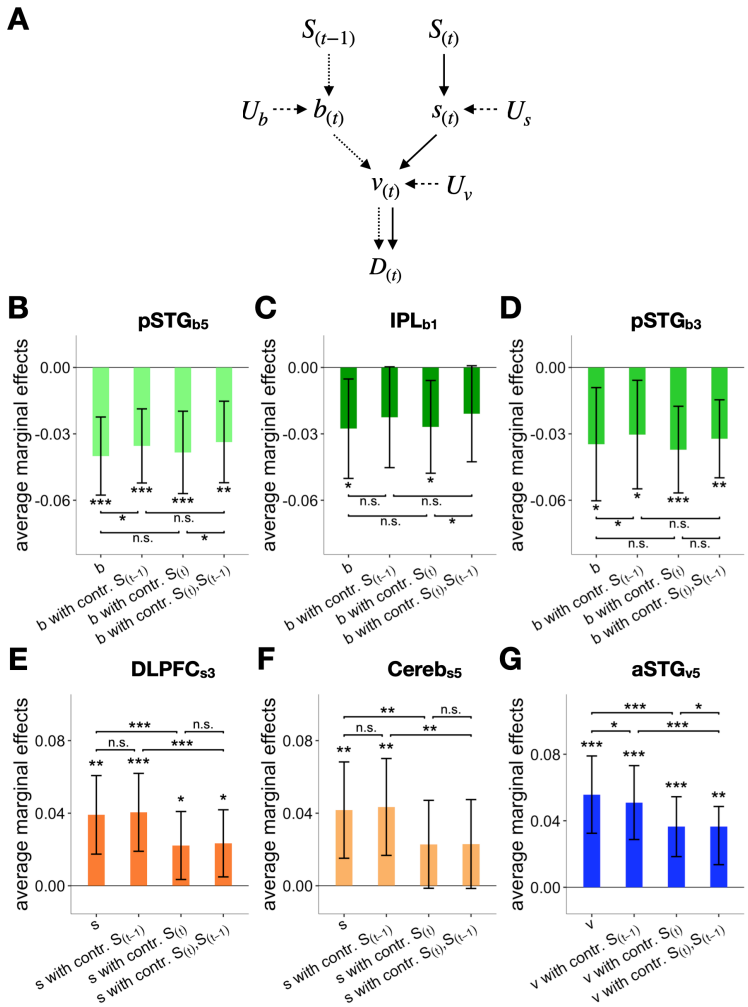


**Figure 8.** Origin of the covariation between the current choice and the brain signals of the latent variables in BMBU. ***A***, The causal structure of the variables implied by the boundary-updating hypothesis. The brain signal of the decision variable ($v_{(t)}$) is influenced by the brain signal of the inferred class criterion ($b_{(t)}$), brain signal of the inferred stimulus ($s_{(t)}$), and the unknown sources ($U_v$). In turn, $b_{(t)}$ is influenced by the previous stimulus ($S_{(t-1)}$) and the unknown sources ($U_b$) whereas $s_{(t)}$ is influenced by the current stimulus ($S_{(t)}$) and the unknown sources ($U_s$). Lastly, $v_{(t)}$ influences the current choice ($D_{(t)}$). If the boundary-updating hypothesis is true, part of the causal influence of $b_{(t)}$ on $D_{(t)}$ must originate from $S_{(t-1)}$, as indicated by the connected chain of the dotted arrows. ***B***-***G***, The average marginal effects (AMEs) of the brain signals on $D_{(t)}$, with the brain signals of $b_{(t)}$ from pSTG$_{b5}$ (***B***), IPL$_{b1}$ (***C***), and pSTG$_{b3}$ (***D***), $s_{(t)}$ from DLPFC$_{s3}$ (***E***), and Cereb$_{s5}$ (***F***), and $v_{(t)}$ from aSTG$_{v5}$ (***G***). In each panel, the influences of the given brain signal on $D_{(t)}$ that can be ascribed to $S_{(t-1)}$ and $S_{(t)}$ were assessed by checking i) whether the AME of the given brain signal on $D_{(t)}$ (left) is significantly reduced or

not after controlling the influence of $S_{(t-1)}$ (second from the left) and $S_{(t)}$ (second from the right), respectively, or ii) whether the AME of $V1$ on $D_{(t)}$ controlling the influence of both $S_{(t-1)}$ and $S_{(t)}$ (right) significantly increased or not after only controlling the influence of $S_{(t)}$ (second from the right) and $S_{(t-1)}$ (second from the left), respectively. The colors of the bars correspond to those of the brain regions shown in **Fig. 6*A***. Asterisks indicate the statistical significance (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$), and "n.s." stands for the non-significance of the test. The 95% CIs of the mean across participants are indicated by the vertical error bars.

reflects the causal relationship of the latent variables implied by BMBU.

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

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

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

As anticipated, the AME of $pSTG_{b5}$ on $D_{(t)}$ was negatively significant across participants ($t_{(17)} = -4.8, p = 1.7 \times 10^{-4}$; Fig. 8*B*, the first bar). Importantly, unlike the size-encoding signal in V1, the negative AME significantly weakened across participants when the contribution of $S_{(t-1)}$ was controlled ($t_{(17)} = 2.8, p = 0.012$; Fig. 8*B*, the change of the first to second bars). On the other hand, controlling

$S_{(t)}$ did not affect the AME of $pSTG_{b5}$ on $D_{(t)}$ at all ($t_{(17)} = 0.29, p = 0.77$; Fig. 8*B*, the change of the first to third bars), which is consistent with the absence of the contribution of $S_{(t)}$ on $b$ in the causal relationship defined by BMBU (Fig. 5*G*). Likewise, the null effect of $S_{(t)}$ on the AMEs of $pSTG_{b5}$ on $D_{(t)}$ was confirmed by the insignificant difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($t_{(17)} = -0.31, p = 0.77$; Fig. 8*B*, the change of the fourth to second bars). Also, the effect of $S_{(t-1)}$ on the AMEs of $pSTG_{b5}$ on $D_{(t)}$ was confirmed by the significant difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($t_{(17)} = -2.7, p = 0.014$; Fig. 8*B*, the change of the fourth to third bars).

The same patterns were also observed for $IPL_{b1}$ and $pSTG_{b3}$ (Fig. 8*C*,*D*). Especially, the AMEs of $pSTG_{b3}$ and $IPL_{b1}$ on $D_{(t)}$ both weakened after controlling $S_{(t-1)}$ ($pSTG_{b3}$: $t_{(17)} = 2.2, p = 0.046$, Fig. 8*D*, the change of the first to second bars; $IPL_{b1}$: $t_{(17)} = 2.1, p = 0.0503$, Fig. 8*C*, the change of the first to second bars), but not after controlling $S_{(t)}$ ($pSTG_{b3}$: $t_{(17)} = -0.57, p = 0.58$, Fig. 8*D*, the change of the first to third bars; $IPL_{b1}$: $t_{(17)} = 0.22, p = 0.83$, Fig. 8*C*, the change of the first to third bars). The null effect of $S_{(t)}$ was confirmed by the insignificance difference between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($pSTG_{b3}$: $t_{(17)} = 0.43, p = 0.67$, Fig. 8*D*; $IPL_{b1}$: $t_{(17)} = -0.41, p = 0.69$, Fig. 8*C*, the change of the fourth to second bars). Also, the effect of $S_{(t-1)}$ was confirmed by the significant or marginally significant differences between the AME with $S_{(t-1)}$ controlled and that with $S_{(t)}$ and $S_{(t-1)}$ both controlled ($pSTG_{b3}$: $t_{(17)} = -1.9, p = 0.081$, Fig. 8*D*, the change of the fourth to third bars; $IPL_{b1}$: $t_{(17)} = -2.2, p = 0.045$, Fig. 8*C*, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of the class boundary to the current choice is significantly ascribed to the previous stimuli supporting the boundary-updating hypothesis on repulsive bias.

Having found the evidence supporting the boundary-updating hypothesis in the brain signals of $b$, we also carried out the same

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

On the contrary, the contribution of $v$ to $D_{(t)}$ must originate not only from $S_{(t-1)}$ but also from $S_{(t)}$ (Fig. 8*A*). In line with this implication, the AME of $\mathrm{aSTG}_{v5}$ on $D_{(t)}$ was significant ($t_{(17)} = 5.1, p = 9.7 \times 10^{-5}$; Fig. 8*G*, the first bar) and significantly decreased both after controlling $S_{(t-1)}$ ($t_{(17)} = -2.8, p = 0.012$; Fig. 8*G*, the change of the first to second bars) and after controlling $S_{(t)}$ ($t_{(17)} = -4.1, p = 7.5 \times 10^{-4}$) (Fig. 8*G*, the change of the first to third bars). Likewise, the AME of $\mathrm{aSTG}_{v5}$ on $D_{(t)}$ with controlled both $S_{(t)}$ and $S_{(t-1)}$

significantly increased both after controlling $S_{(t-1)}$ ($t_{(17)} = 4.1, p = 6.7 \times 10^{-4}$; Fig. 8*G*, the change of the fourth to second bars) and after controlling $S_{(t)}$ ($t_{(17)} = 2.8, p = 0.012$; Fig. 8*G*, the change of the fourth to third bars). Put together, the AME analyses suggest that the contribution of the decision variable to the current choice is significantly ascribed to both current and the previous stimuli supporting the boundary-updating hypothesis.

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

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

Furthermore, to test the sensory-adaptation hypothesis, we examined whether $VC_{s3}$ carries the stimulus signal via which the previous stimulus ($S_{(t-1)}$) contributes to the current choice ($D_{(t)}$). However, the AME of $VC_{s3}$ on $D_{(t)}$ did not decrease when the contribution of $S_{(t-1)}$ was controlled ($t_{(17)} = 0.28,\ p = 0.78$). Likewise, the AME of $VC_{s3}$ on $D_{(t)}$ with $S_{(t)}$ controlled did not differ from that with both $S_{(t-1)}$ and $S_{(t)}$ controlled ($t_{(17)} = 0.70, p = 0.49$). These

results corroborate the AME analyses on $V1$ in Experiment 1 (Fig. 4*D–F*), confirming that the previous stimulus is unlikely to contribute to the current choice via the stimulus-related signals in the early visual cortex.

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

## 2.4 Discussion

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

**Dissociation between sensory-adaptation in V1 and repulsive bias**

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

From a general perspective, our findings demonstrate a dissociation between the impact of previous decision-making

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

### The representations of the class boundary in IPL and pSTG

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

To our best knowledge, this issue has so far been addressed by one fMRI work (White, Mumford et al. 2012), which reported the

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

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

significant negative regression coefficient of the previous decision in Experiment 1 (Fig. 3*C*, 5*A*) could have been reflective of the negative influence of the previous motor response, as previously suggested (Zhang and Alais 2020).

### The representations of *inferred* stimuli in DLPFC and cerebellum

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

It is notable that BMBU treats $s_{(t)}$ as the random variable that has the noise variability in addition to being influenced by the physical stimulus (Fig. 8*A*). This means that the brain signal of $s_{(t)}$ is supposed to be associated with the choice even when the current stimulus was controlled because the noise variability can also influence the current choice, as captured by the concept of 'choice probability' (Macke and Nienborg 2019). However, unlike DLPFC and cerebellum, the AME of $V1$ on the current choice disappeared after controlling the current stimuli, which disqualifies $V1$ as the brain signal of $s_{(t)}$. In line with this, the AME of $VC_{s3}$ on $D_{(t)}$ also disappeared after $S_{(t)}$ was controlled in Experiment 2, which again disqualifies $VC_{s3}$ as the valid brain signal of $s_{(t)}$.

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

its subjective representation (Renart and Machens 2014), consistent with the previous findings of our group (Lee, Blake et al. 2007, Choe, Blake et al. 2014).

### The representation of the decision variable in aSTG

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

# Chapter 3

# Boundary updating as a source of history effect on decision uncertainty

Boundary updating, a process of adapting the decision boundary to previous stimuli, can exert a historical influence on binary choices. However, its influence on decision uncertainty as a historical source has been overlooked. Here, we show that boundary updating also confers a history effect on decision uncertainty, elevating decision uncertainty as current choices become increasingly congruent with previous stimuli, as evidenced by changes in its behavioral, neural, and physiological correlates.

## 3.1 Introduction

Stimulus and boundary are the two pillars of binary decision-making, both affecting observers' choices (Treisman and Williams 1984, Lages and Treisman 1998, Morgan, Watamaniuk et al. 2000, Benjamin, Diaz et al. 2009, Lages and Treisman 2010, Raviv, Lieder et al. 2014, Norton, Fleming et al. 2017, Hachen, Reinartz et al. 2021, Lee, Lee et al. 2023). One pillar, the boundary, is seldom presented explicitly at the time of decision and, thus, needs to be internally formed by observers (Treisman and Williams 1984, Lages and Treisman 1998, Raviv, Lieder et al. 2014, Hachen, Reinartz et al. 2021, Lee, Lee et al. 2023). This internal boundary tends to be continually updated to shift toward previous stimuli (Treisman and Williams 1984, Lages and Treisman 1998, Raviv, Lieder et al. 2014, Hachen, Reinartz et al. 2021, Lee, Lee et al. 2023) (Fig. 9*A*-1), making binary choices dependent on what has come before. This boundary updating accounts for a history effect on choice called 'repulsive bias,' where the current choice is biased away from

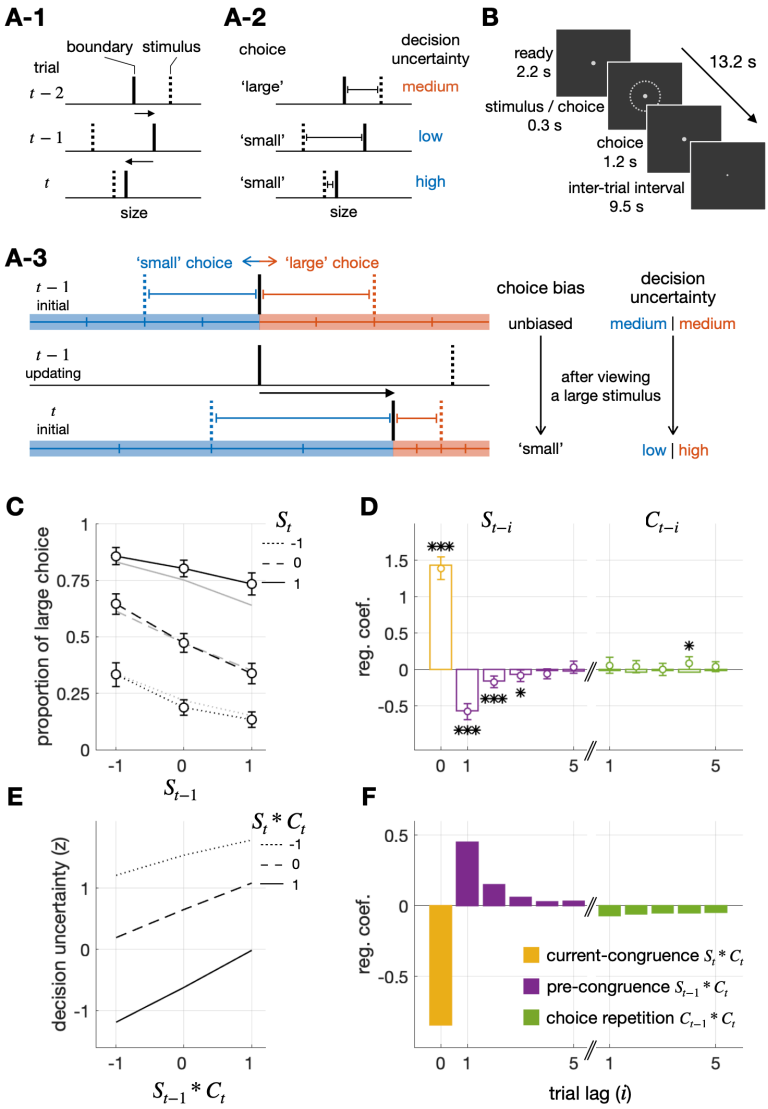


**Figure 9.** The pre-congruence effect on decision uncertainty. ***A,*** Schematic illustration of boundary updating (***A*-1**), decision uncertainty (***A*-2**), and the pre-congruence effect in decision uncertainty (***A*-3**). According to the boundary updating account, the boundary (solid vertical black bars) is drawn towards previous stimuli (dotted vertical black bars) (***A*-1**). Making a decision is more difficult as the boundary and the stimulus is closer (***A*-2**). The boundary attraction to previous stimuli brings the pre-congruence effect to decision uncertainty (***A*-3**). The horizontal baseline bar indicates the stimulus ranges that are smaller (blue) or larger (orange) than the boundary. The dotted blue and orange vertical bars indicate the average stimulus sizes smaller and larger than the boundary, respectively. The boundary is initially unbiased on the trial $t-1$ (***A*-3**, top). As a result, distances from the boundary to the average stimulus sizes of 'large' and 'small' choices are equivalent. However, if a large stimulus was sampled on the

trial $t-1$ (*A*-3, middle), then the boundary is biased to the large side at the trial $t$ (*A*-3, bottom). As a result, the distance between the boundary and the average 'large' stimulus gets shorter, while the distance to the average 'small' stimulus becomes longer. This leads to an increase in decision uncertainty for 'large' choices and a decrease in decision uncertainty for 'small' choices after seeing a large stimulus in previous trials. *B*, Task structure. *C*,*D*, The repulsive bias demonstrated in psychometric curves (*C*) and logistic regressions (*D*). The proportion of 'large' choices are plotted against previous stimuli ($S_{t-1}$) conditioned on current stimuli ($S_t$) in black lines for humans and in gray for BMBU (*C*). The negative slopes of the curves indicate the repulsive bias. The coefficients in the multiple logistic regression of current choices ($C_t$) onto current and previous stimuli ($S_{t-i}$) and previous choices ($C_{t-i}$) are plotted in circles and bars for humans and BMBU, respectively (*D*). The significant coefficients of previous stimuli, not previous choices, indicate the repulsive bias. *E*,*F*, The pre- and the current-congruence effects in decision uncertainty that are simulated by BMBM, shown in psychometric curves (*E*) and linear regressions (*F*) based on simulated choices. The simulated decision uncertainty increases with the congruence between current choices and previous stimuli ($S_{t-1} * C_t$) but decreases with the congruence between current choices and current stimuli ($S_t * C_t$), the former and latter implying the pre- and the current-congruence effects on decision uncertainty, respectively (*E*). The coefficients in the multiple linear regression of the simulated decision uncertainty onto the congruences of current choices with previous stimuli ($S_{t-i} * C_t$) and with previous choices ($C_{t-i} * C_t$) (*F*). The pronounced coefficients of the congruence between current choices and past stimuli (purple bars) implies the pre-congruence effect. Here and thereafter, error bars indicate 95% confidence interval of the means, and asterisks indicate $P$ values of two-sided Student's $t$-test: $*< 0.05$, $**< 0.01$, $***< 0.001$.

previous stimuli on a trial-to-trial basis (Treisman and Williams 1984, Lages and Treisman 1998, Bosch, Fritsche et al. 2020, Hachen, Reinartz et al. 2021, Lee, Lee et al. 2023) (Fig. 9*A*-3).

Critically, this boundary updating is supposed to confer a history effect also on 'decision uncertainty,' a cognitive quantity that increases as the stimulus and the boundary come nearer (Grinband, Hirsch et al. 2006, Kepecs, Uchida et al. 2008, White, Mumford et al. 2012, Hebart, Schriever et al. 2014) (Fig. 9*A*-2). It should be noted that decision uncertainty differs from 'sensory uncertainty,' which is influenced solely by the noise present in stimuli (Pouget, Drugowitsch et al. 2016). Suppose the boundary was neutral, and a large stimulus was presented in the last trial (Fig. 9*A*-3, top and

middle). The boundary would then be reset on the large side according to the boundary updating (Treisman and Williams 1984, Lages and Treisman 1998, Raviv, Lieder et al. 2014, Hachen, Reinartz et al. 2021, Lee, Lee et al. 2023), bringing itself nearer to the stimuli associated with the 'large' choice but farther from those associated with the 'small' choice (Fig. 9*A*-3, bottom). Consequently, having previously viewed large stimuli, observers' 'large' choices will be accompanied by high uncertainty, whereas their 'small' choices by low uncertainty. Generally put, *the more congruent the current choice is with previous stimuli, the more uncertain it is*. We shall call this boundary-induced history effect on decision uncertainty 'pre-congruence effect.'

The pre-congruence effect is straightforward but has not been established empirically. The neglect of this obvious source of the variability in decision variability is odd, given the fundamental roles of decision uncertainty in adaptive human behavior such as volatility monitoring (Behrens, Woolrich et al. 2007), learning from errors(Drugowitsch, Mendonça et al. 2019), and executive control (Shenhav, Cohen et al. 2016), and its wide-ranging associations with behavioral (response time (Grinband, Hirsch et al. 2006, Urai, Braun et al. 2017, Braun, Urai et al. 2018, Fietz, Pöhlchen et al. 2022)), physiological (pupil-size (Urai, Braun et al. 2017, Fietz, Pöhlchen et al. 2022)), and neural (salience network (Brown and Braver 2005, Grinband, Hirsch et al. 2006, Behrens, Woolrich et al. 2007, Urai, Braun et al. 2017, Fietz, Pöhlchen et al. 2022)) measures. Here, we probed human response time (RT), pupil size, and neural activity to ascertain the presence of the pre-congruence effect and assess its contribution to the variability in decision uncertainty.

## 3.2 Materials and Methods

### Behavior data acquisition

In the main results, we utilized two previously published datasets from our lab (Choe, Blake et al. 2014, Choe, Blake et al. 2016, Lee, Lee et al. 2023)The fMRI dataset (Lee, Lee et al. 2023) consisted of

18 participants (9 females) aged 20–30 years, while the eye tracking dataset (Choe, Blake et al. 2014, Choe, Blake et al. 2016) comprised 23 participants (11 females) aged 18–36 years. For the supplementary results, we used another dataset from our lab (Lee, Lee et al. 2023) that included 30 participants (13 females) aged 18–30 years. The experimental procedures were approved by the Research Ethics Committee of Seoul National University, and all participants provided informed consent and were unaware of the study's objectives.

### Task procedure

The main data set used the following task procedure (Fig. 9*B*): Participants were instructed to fixate at the center of the screen and classify a brief (0.3s) ring-shaped stimulus as either 'small' or 'large' within 1.5s after stimulus onset by pressing the left or right key, respectively. The timing and identity of each key press were recorded. Trials were separated by 13.2s and participants were given feedback on their performance after each run of 26 trials. Before the main runs, participants completed 54 practice trials followed by 180 threshold-calibration trials. During the threshold-calibration trials, which were separated by 2.7s, participants received trial-to-trial feedback based on the boundary with a radius of 2.84°. A Weibull function was fit to the psychometric curves obtained from the threshold-calibration trials using a maximum-likelihood procedure. The size threshold $\Delta°$ (i.e., the size difference between medium ring and large or small ring) associated with a 70.7% correct proportion was estimated from the fitted Weibull function. The mean and standard deviation of the estimated size threshold were 0.023 and 0.0078, respectively. One of three estimated ring sizes ($2.84 - \Delta°, 2.84°, 2.84 + \Delta°$) was shown on each trial. Participants had extensive training on the task before participating in the main experiments.

Two of the three supplementary datasets were collected from 58 participants, who performed pitch and ring-size classification tasks on separate sessions, respectively, with 315 trials using fine-

grained stimuli randomly sampled from normal distributions with an inter-trial interval of 2s and without trial-to-trial feedback for each task. The other dataset was collected from 30 participants, who performed the same classification task over 5 daily sessions with 1,700 trials using ring stimuli of 5 discrete sizes ($3.84°, 3.92°, 4.00°, 4.08°, 4.16°$), with trial-to-trial feedback and an inter-trial interval of 2.5s.

### Estimation of decision uncertainty

By fitting the model parameters of BMBU ($\Theta = \{\mu_0, \sigma_0, \sigma_m, \kappa\}$) separately for each human participant, we were able to create a Bayesian observer tailored to each individual. By conducting the experiment again with these Bayesian observers using the same stimulus sequences presented to their human counterparts, we obtained a sufficient number ($10^6$ repetitions) of simulated choices, $c_t$, and decision uncertainty values, $u_t$. These values were determined based on the corresponding number of stimulus estimates, $s_t$, and boundary estimates, $b_t$, for each Bayesian observer. Finally, we took the averages across those $10^6$ simulations as the final outcomes. When estimating $u_t$ for the observed choice $C_t$, we only included simulation outcomes where the simulated choice $c_t$ matched the observed choice $C_t$.

### Definition of congruence

The term 'congruence' refers to the product of the variable of interest and the current choice, where choices were represented as 1 for 'large' and -1 for 'small' and stimuli were represented as 1, 0, and -1 for large, medium, and small sizes, respectively. For instance, if the current choice was 'large (1)' and the previous stimulus was a 'small' size (-1), then the congruence would be -1 ($1 * -1$). In the supplementary data where there were five stimuli, the stimuli were encoded as [-2, -1, 0, 1, 2].

### Definition of the signed R-squared

The signed R-squared is a measure that takes into account the

direction of the relationship between a variable of interest and other variables in a multiple regression model. It is calculated by multiplying the sign of the regression coefficient with the uniquely explained variance of the variable. The uniquely explained variance is obtained through variance partitioning analysis(Borcard, Legendre et al. 1992, Perquin, Heed et al. 2022), which involves comparing the R-squared values of the full regression model that includes all variables and the reduced regression model that excludes the variable of interest. For instance, to calculate the signed R-squared of variable $x$, we first calculate the R-squared of the full model ($R^2_{xyz}$) that includes variables $x$, $y$, and $z$. Next, we compute the R-squared of the reduced model ($R^2_{yz}$) that includes variables $y$ and $z$ only. The uniquely explained variance of $x$ ($R^2_{x,yz}$) is obtained by subtracting $R^2_{yz}$ from $R^2_{xyz}$. Finally, the signed R-squared of $x$ is defined as $sign(\beta_x)R^2_{x,yz}$, where $sign(\beta_x)$ is the sign of the regression coefficient $\beta_x$ of the multiple regression model with variables $x$, $y$, and $z$ as regressors.

BOLD signal

MRI data were preprocessed as the previous section except that the spatial smoothing is additionally applied with 8×8×8mm full-width half-maximum Gaussian kernel after normalizing the image.
To explore the brain regions that showed a correlation between BOLD signals and $u_t$, we conducted a regression analysis for each participant and voxel. The response variable for the regression model was the preprocessed BOLD signals concatenated across runs. We used three explanatory vectors in the regression model. The first vector was created by convolving the canonical hemodynamic function with $u_t$ for each trial. The second vector was created by convolving the canonical hemodynamic function with a constant value of 1, and both vectors were standardized. The third vector contained only a single value of 1. The first regression coefficient indicated how well $u_t$ predicted the BOLD signal for each voxel, and by calculating this coefficient for every voxel, we generated a map of

regression coefficients for each participant and the entire brain.

To determine whether the average coefficient of $u_t$ across participants significant, we performed a two-sided Student $t$-test on each voxel. We then corrected the $P$ values of the entire brain for false discovery rate (FDR) (Benjamini and Hochberg 1995). The total number of voxels in the entire brain was 90,481, and we defined the brain areas significantly correlated with $u_t$ as the voxel clusters covering a region larger than 15 contiguous voxels and having FDR-corrected $P$ values less than 0.05. For the ROI analysis, we averaged the preprocessed BOLD signals across individual voxels within an ROI.

## Pupillometry

Stimuli were presented in a dimly lit room on a gamma-linearized 22-inch CRT monitor (Totoku CV921X CRT monitor) operating at vertical refresh rate of 180Hz and a spatial resolution of 800×600 pixels. Stimuli were generated using MATLAB (MathWorks) in conjunction with MGL (http://justingardner.net/mgl) on a Macintosh computer. Observers viewed the monitor at a distance of 90cm while their binocular eye positions were sampled at 500Hz by an infrared eye tracker (EyeLink 1000 Desktop Mount, SR Research; instrument noise, 0.01° RMS). The LED illuminator and camera were positioned side by side, at a distance of 65cm from the observer, and angled toward the observer's face to ensure that infrared light illuminated both eyes and was being reflected from both eyes and imaged on the camera sensor.

When measuring gaze position using video-based methods, eye blinks can interfere with the accuracy of the data, as pupil and gaze information are not available during these periods. We identified eye blinks based on three criteria: (i) missing pupil data for either eye, (ii) pupil-size measurements with unrealistically large fluctuations (>50 units per sample), or (iii) substantial deviation (>20°) of gaze position from the screen center. Data collected immediately before and after an eye blink (±200 ms) were likely contaminated and therefore we linearly interpolated the pupil-sizes

and the three gaze positions ($x$, $y$, and $d$ $(=\sqrt{x^2+y^2})$) before and after the blink events.

To minimize confounding effects from gaze position on pupil-size measurements, we processed the blink-free samples of pupil-size and three gaze positions time series by linear detrending, band-pass filtering (0.01Hz to 4Hz cut-off frequency with a Butterworth filter), and resampling to 10Hz. The resampled time series of pupil-size was then orthogonalized from the three resampled gaze positions up to their fourth powers (total 12 regressors) to extract any confounding in the pupil-size originating from the gaze position(Choe, Blake et al. 2016). The orthogonalized pupil-size time series was then standardized by subtracting the mean and dividing the standard deviation of the time series. The pupil-sizes of the eyes were averaged, and trials were epoched and baseline corrected by subtracting the baseline pupil-size averaged across 0ms to 500ms from the stimulus onset. We chose this baseline time window because we assumed that the decision-related pupil signal would be initiated between -200ms to 300ms from the stimulus onset and because the latency of the empirical responses was around 200ms(Korn and Bach 2016). Finally, the baseline corrected pupil-sizes were aligned by the stimulus onset or the choice following previous studies(de Gee, Knapen et al. 2014, Urai, Braun et al. 2017).

## 3.3 Results

We recruited 41 human participants and asked them to perform a binary classification task. The task involved classifying a series of rings presented in a pseudo-randomized order (Buracas and Boynton 2002) as either ‘small’ or ‘large,’ one at a time, under moderate time pressure (Fig. 9*B*). As previously reported (Treisman and Williams 1984, Lages and Treisman 1998, Raviv, Lieder et al. 2014, Hachen,

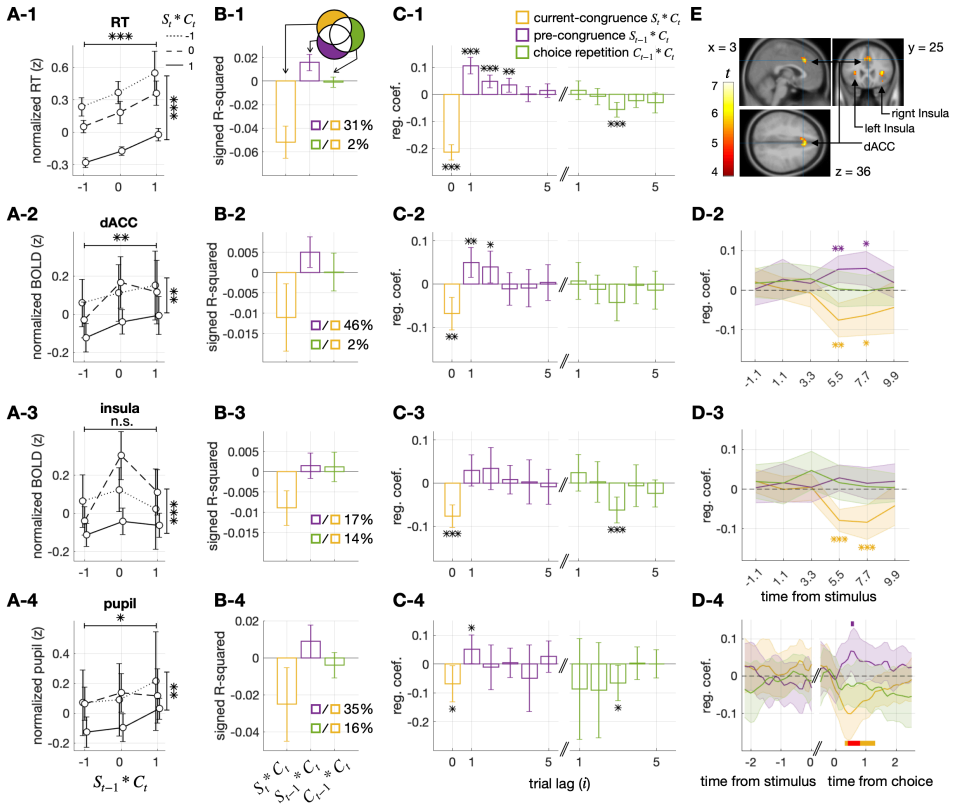


**Figure 10.** The pre- and the current-congruence effects on RT, cortical activity, and pupil-size. *A*, The pre- and the current-congruence effects captured in chronometric (1), neurometric (2,3), and pupillometric (4) curves. Normalized measures of RT (1), dACC (2) and insula (3) BOLD signals acquired at the fourth time frame within a trial, and pupil-size (4) taken at 0.5 sec after button press are plotted against the congruences between current choices and immediately preceding stimuli ($S_{t-1}*C_t$) and the congruences between current choices and current stimuli ($S_t*C_t$). In each panel, the asterisks on top and side indicate the significant contributions of the pre-congruence effect ($S_{t-1}*C_t$) and the current-congruence effect ($S_t*C_t$) to the measures of interest, respectively. Their significant contributions are ascertained by the multiple regressions of the measures of interest onto $S_{t-1}*C_t$, $S_t*C_t$, and $C_{t-1}*C_t$. The presence of the asterisks on top and side indicate that the RT, dACC activity, and pupil-size are under the influences of both the pre- and the current-congruence effects, whereas the insula activity is under the influence of only the current-congruence effect. *B*, Variances uniquely explained by the congruences of current choices with current stimuli (yellow), with immediately preceding stimuli (purple), and with immediately preceding choices (green). The percentages quantify the ratio of the explained variance sizes between the yellow bar and each of the others. *C*, The pre- and the current-congruence effects and the effect of previous choices captured in multiple regression analyses. The coefficients in the multiple linear regression of

the measures onto the congruences of current choices with current stimuli (yellow), past stimuli (purple), and past choices (green) are plotted in bars. The asterisks indicate the significant contributions of a given regressor to the variability of the measures. Note that the significant influence of the pre-congruence effect ($S_{t-i}*C_t$) for the RT (**1**), dACC activity (**2**), and pupil-size (**4**), but not for the insula activity (**3**). ***D***, Time courses of the coefficients in the multiple regression of the dACC (**2**) and insula activity (**3**), and pupil-size (**4**) onto the congruences of current choices with current stimuli (yellow; $S_t*C_t$), with immediately preceding stimuli (purple; $S_{t-1}*C_t$), and with immediately preceding choices (green; $C_{t-1}*C_t$). The colored horizontal bars indicate *P* values of two-sided Student's *t*-test of the coefficients (purple and yellow: * < 0.05; red: ** < 0.01) (**4**). The emergence of the pre-congruence effect is aligned with the time frames associated with decision-making for both dACC and pupil-size. ***E***, *t*-statistics of voxel clusters whose BOLD signals significantly predict the simulated decision uncertainty (two-sided Student's *t*-test $P < 0.05$ after controlling the false discovery rate; the minimum number of contiguous voxels is 15).

Reinartz et al. 2021, Lee, Lee et al. 2023), the participants displayed the repulsive bias, making less of the 'large (small)' choice following the large (small) stimulus in the previous trial (Fig. 9*C*; $\beta_{S_{t-1}} = -0.57$ ($P = 2.87 \times 10^{-14}$), $\beta_{S_t} = 1.30$ ($P = 4.9 \times 10^{-22}$), $\beta_{C_{t-1}} = 0.12$ ($P = 0.036$), where $S_{t-i}$, stimulus on $i$-th preceding trial; $C_{t-i}$, choice on $i$-th preceding trial). Also, the repulsive bias was not constrained to the immediately preceding stimulus but extended to the stimulus at the lag of three trials (Fig. 9*D*).

To simulate how the boundary updating relates to the pre-congruence effect, we employed a Bayesian model for boundary updating (BMBU). Before proceeding, we ensured that BMBU's choices readily captured the repulsive bias in the current data set, which is evident in the close correspondence of human and BMBU's psychometric functions (Fig. 9*C*) and regression coefficients (Fig. 9*D*). This implies that the participants update their internal boundary by shifting it toward previous stimuli as BMBU does. Then, we turned to the simulated decision uncertainty to see whether the boundary updating induces the pre-congruence effect. As we intuited early on (Fig. 9*A*-3), the more congruent BMBU's current choice was with previous stimuli, the more uncertain it was (Fig. 9*E*). BMBU also

lets us foresee that the pre-congruence effect gradually diminishes as trials further lag (the purple bars in Fig. 9*F*). In addition to capturing this newly recognized history effect, BMBU also accounts for the well-established 'current-congruence effect'—a phenomenon where decision uncertainty decreases as the current choice becomes congruent with current stimuli (Kepecs, Uchida et al. 2008, Sanders, Hangya et al. 2016, Urai, Braun et al. 2017) (as indicated by the progressive elevation of the lines in Fig. 9*E* and the yellow bar in Fig. 9*F*).

Moving on, as a first step to establish the empirical presence of the pre-congruence effect, we probed RT (N=41), a well-known behavioral correlate of decision uncertainty (Grinband, Hirsch et al. 2006, Urai, Braun et al. 2017, Fietz, Pöhlchen et al. 2022). As predicted, participants' RTs displayed both the pre-congruence effect ($\beta_{S_{t-1}*C_t} = 0.11$ $(P = 9.2 \times 10^{-9})$) and the current-congruence effect ($\beta_{S_t*C_t} = -0.21$, $(P = 2.7 \times 10^{-18})$) (Fig. 10*A*-1). Notably, the size of the pre-congruence effect was substantial and tantamount to 31% of that of the current-congruence effect (Fig. 10*B*-1). The pre-congruence effect on decision uncertainty could be traced back to three trials (Fig. 10*C*-1), which is consistent with the history effect on choice—the repulsive bias (Fig. 9*D*).

We further confirmed the robustness of the pre-congruence and current-congruence effects using the RTs collected under different conditions, where fine-grained auditory (Fig. 11-1, $\beta_{S_{t-1}*C_t} = 0.12$ $(P = 2.0 \times 10^{-11})$, $\beta_{S_t*C_t} = -0.35$, $(P = 4.5 \times 10^{-34})$) or ring (Fig. 11-2, $\beta_{S_{t-1}*C_t} = 0.16$ $(P = 3.1 \times 10^{-18})$, $\beta_{S_t*C_t} = -0.27$, $(P = 2.0 \times 10^{-28})$) stimuli were classified with a short inter-trial interval (2 sec; N=58) and where five ring stimuli were used with trial-to-trial feedback and a shorter inter-trial interval (2.5 sec, N=30) (Fig. 11-3, $\beta_{S_{t-1}*C_t} = 0.050$ $(P = 7.6 \times 10^{-4})$, $\beta_{S_t*C_t} = -0.11$, $(P = 1.1 \times 10^{-6})$). These findings demonstrate that the effects of pre- and current-congruence are applicable to various sensory modalities, feedback conditions, and inter-trial intervals.

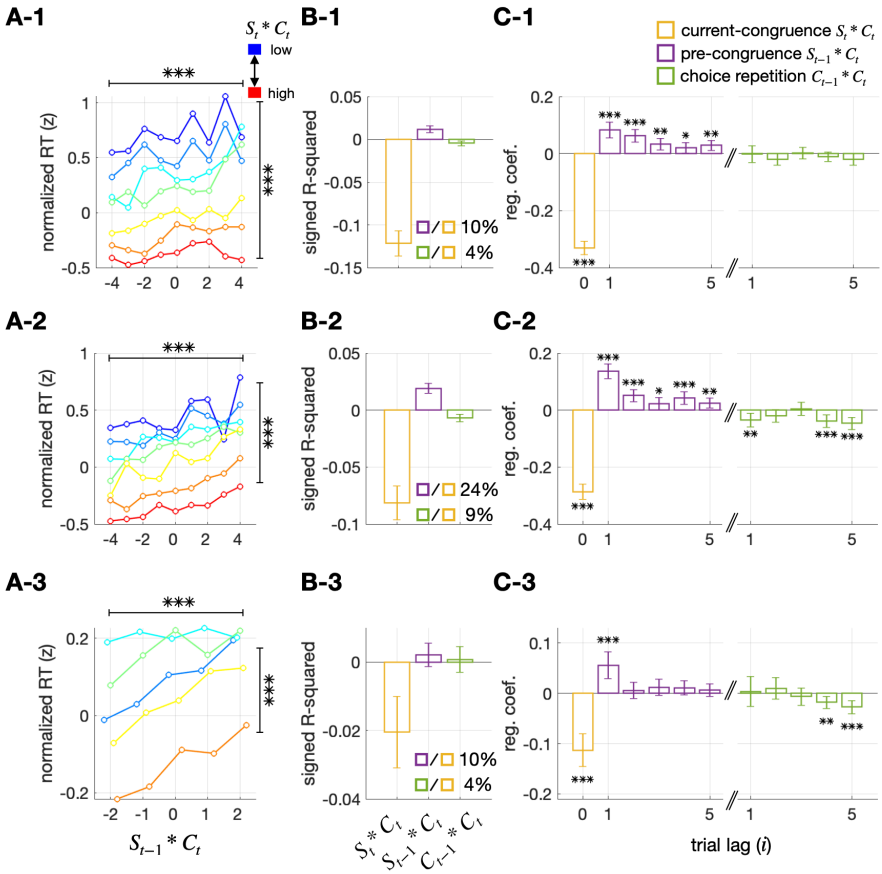


**Figure 11.** The pre- and current-congruence effects on reaction time (RT) in other task conditions. The pre- and current-congruence effects captured in chronometric curves for fine-grained pitch (1) or ring (2) stimuli and five ring stimuli with trial-to-trial feedback (3). (1,2) To better illustrate the data, continuous stimuli were discretized into nine bins. The lowest two current-congruence trials (error trials on extremely small and large rings or low and high pitches) were not shown because the trials were so rare that the data points are unreliable. *A*, RT is plotted against the congruences between current choices and the immediately preceding stimuli ($S_{t-1} * C_t$) and the congruences between current choices and current stimuli ($S_t * C_t$). The significance of the multiple regression coefficients for predicting RT by $S_{t-1} * C_t$, $S_t * C_t$, and $C_{t-1} * C_t$ are indicated by asterisks. The horizontal and vertical lines indicate the significance of the regression coefficients of $S_{t-1} * C_t$ and $S_t * C_t$, respectively. The pre-congruence effect is statistically significant in RT. *B*, the proportions of variance that are uniquely explained by the congruences of current choices to current stimuli (yellow), to immediately preceding stimuli (purple), and to immediately preceding choices (green). The percentages indicate the ratio between the sizes of the uniquely explained variances. *C*, the multiple regression coefficients of the congruences of current choices to current stimuli (yellow), previous stimuli

(purple), and previous choices (green) for predicting RT.

| cortical area | contiguous voxel (N) | peak voxel | |
|---|---|---|---|
| | | MNI coordinate | $P$ |
| dorsal anterior cingulate cortex | 64 | $[6, 27, 36]$ | $8.2 \times 10^{-6}$ |
| left insula | 27 | $[-30, 21, 9]$ | $5.7 \times 10^{-6}$ |
| right insula | 66 | $[36, 18, 6]$ | $7.5 \times 10^{-7}$ |

**Table 3.** The brain regions where there was a significant correlation between BOLD signals and simulated decision uncertainty. We defined these regions based on two conditions: (1) the regression coefficient between the simulated decision uncertainty and the BOLD signal was significant (two-sided Student's $t$-test $P < 0.05$, after controlling the false discovery rate), and (2) there were at least 15 contiguous voxels that met the first condition.

Next, by regressing the fMRI data (N=18) onto BMBU's decision uncertainty, we identified the neural correlates of decision uncertainty in the dorsal anterior cingulate cortex (dACC) and insula (Fig. 10*E*; Table. 3), consistent with previous studies (Grinband, Hirsch et al. 2006, Shenhav, Straccia et al. 2014, Fietz, Pöhlchen et al. 2022). At the time point with their highest correlation with BMBU's decision uncertainty (Fig. 12-1), the dACC's responses showed the pre-congruence and current-congruence effects (Fig. 10*A*-2; $\beta_{S_{t-1}*C_t} = 0.053$, $(P = 0.0037)$, $\beta_{S_t*C_t} = -0.076$ $(P = 0.0016)$). The pre-congruence effect in the dACC was equivalent in size to 46% of that of the current-congruence effect (Fig. 10*B*-2), traced back to two trials (Fig. 10*C*-2), and pronounced at the time points aligned with decision-making (Fig. 10*D*-2). In contrast, the insula exhibited only the current-congruence effect (Fig. 10*A*3-*D*3, 12-2; $\beta_{S_{t-1}*C_t} = 0.028$ $(P = 0.083)$, $\beta_{S_t*C_t} = -0.079$ $(P = 1.4 \times 10^{-5})$).

Finally, the pupil-size data (N=23) also confirmed the pre-congruence effect. At the time point with its highest correlation with BMBU's decision uncertainty (Fig. 12-3), the pupil size displayed both pre-congruence and current-congruence effects (Fig. 10*A*-4; $\beta_{S_{t-1}*C_t} = 0.053$ $(P = 0.0037)$, $\beta_{S_t*C_t} = -0.076$ $(P = 0.0016)$). The pre-congruence effect in pupil size was equivalent in size to 35% of that of the current-congruence effect (Fig. 10*B*-4), traced back to the

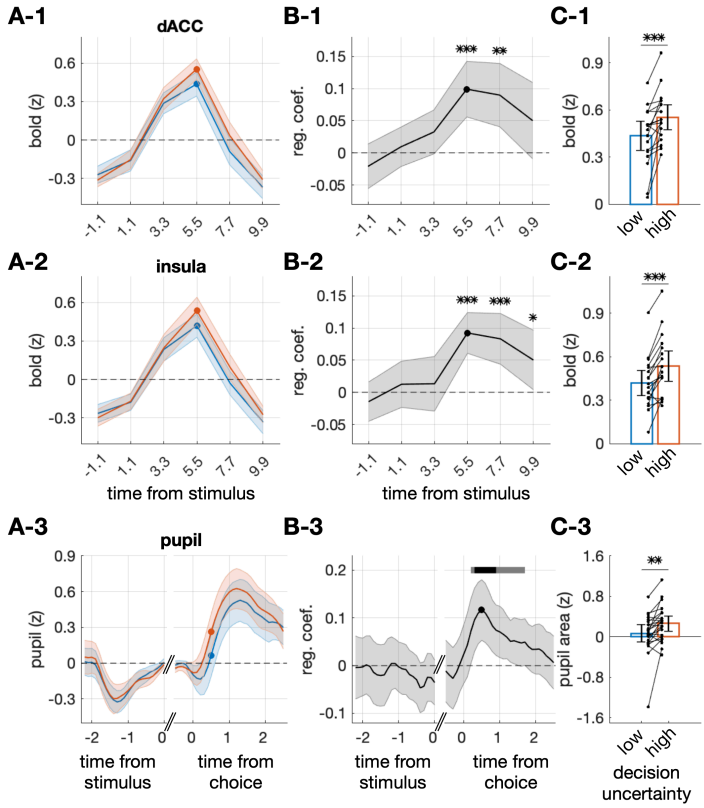


**Figure 12.** The time courses of dACC, insula, and pupil-size and their linear regressions to decision uncertainty. ***A***, the time courses of (1) dACC, (2) insula, and (3) pupil-size, which were conditioned by the simulated decision uncertainty into lower (blue) and higher (red) decision uncertainty conditions. ***B***, the linear regression coefficient of the simulated decision uncertainty to predict dACC, insula or pupil-size for each time frame. The *P* values of two-sided Student's *t*-test of the coefficients are indicated by the horizontal bar (gray: $*< 0.05$; black: $**< 0.01$). ***C***, the BOLD signals and pupil-sizes are shown for the time frame with the most significant coefficient, which is indicated by the circles in the time courses (panels ***A*** and ***B***). Each dot pair and line represent a single participant.

immediately preceding trial (Fig. 10*C*-4), and pronounced at the time points aligned with decision-making (Fig. 10*D*-4).

## 3.4 Discussion

The pre-congruence effect demonstrated here differs considerably from earlier discoveries of the history effects on decision

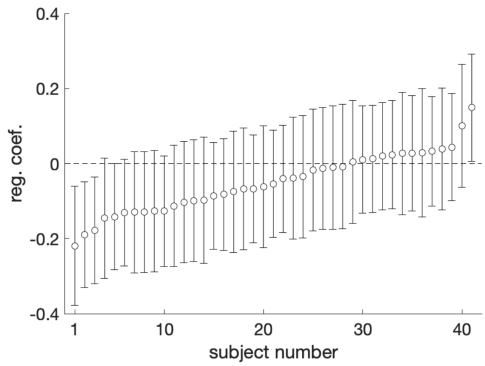


**Figure 13.** The regression coefficient between consecutive decision uncertainties. The circles represent the linear regression coefficients of the decision uncertainty of the previous trial used to predict the decision uncertainty of the current trial. The decision uncertainty was simulated using BMBU. Only a single participant showed a significantly positive correlation between consecutive decision uncertainties

uncertainty. First, it differs from the history effect on RTs in tasks with an autocorrelated stimulus sequence (Cho, Nystrom et al. 2002) because the stimuli were independent across trials in our task (Buracas and Boynton 2002). Second, it differs from the 'priming effect (Kristjánsson and Campana 2010, Galluzzi, Benedetto et al. 2022),' where RTs become faster as current stimuli become congruent with previous stimuli, an effect opposite to the pre-congruence effect, where RTs become slower under an identical situation. Third, it differs from the phenomenon where the congruence between subsequent choices biases RT (Urai, De Gee et al. 2019) because the pre-congruence effect is founded on the congruence between previous stimuli and current choices, not between previous and current choices. Fourth, it differs from the so-called 'confidence leak (Rahnev, Koizumi et al. 2015),' where confidence reports are correlated across trials, because the confidence leak does not entail any historical effect on choices at all, whereas the pre-congruence effect entails the history effect on choice, the repulsive bias. Fifth, it differs from the so-called 'post-error slowing (Jentzsch and Dudschig 2009),' a tendency to slow down after committing an error on previous trials, because the pre-

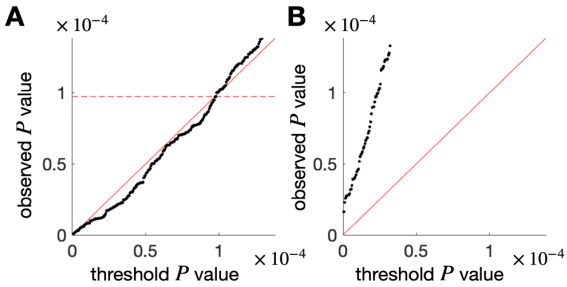


**Figure 14.** Controlling the false discovery rate (FDR) of the *P* values of the whole-brain analysis. ***A***,***B***, The lowest P values of each voxel in the whole brain for predicting decision uncertainty in BMBU (***A***) and the constant-boundary model (***B***), respectively, plotted against the FDR threshold *P* values. The critical *P* value, which is the maximum *P* value lower than the FDR threshold (indicated by the horizontal dashed line), determines the significant voxels after controlling for FDR. For BMBU, 177 voxels were significant, whereas no voxel was significant for the constant-boundary model.

congruence effect does not entail error feedback. Furthermore, the simulated decision uncertainty of BMBU does not show positive dependences between consecutive trials (Fig. 13), which must be present both in the confidence leak and the post-error slowing. Lastly, our work should not be confused with the history effects on choices (Urai, Braun et al. 2017, Braun, Urai et al. 2018) which were investigated using RT, pupil size, and confidence reports because they were not about decision uncertainty modulation but about choice bias.

Our work has important implications for the cognitive neuroscience of decision uncertainty and history effects. We could explain more of the trial-to-trial variability of decision uncertainty by discovering its historical source—boundary updating. This will significantly facilitate discerning the 'genuine' neural substrate of decision uncertainty. For example, the dACC and the insula appeared similarly involved in signaling task difficulty in previous studies (Seeley 2019). However, we could point to the dACC as a more genuine locus of decision uncertainty by demonstrating both the pre-congruence and current-congruence effects in the dACC but only the current-congruence effect in the insula. In the same vein, we could

no longer find the significant neural correlate of decision uncertainty without incorporating the boundary updating into the definition of decision uncertainty (Fig. 14). Our discovery of the historical source of decision uncertainty also calls for a need to reinterpret previous findings. For example, in metacognition research, the presence of history effects in confidence reports has been considered a source of metacognitive inefficiency (Shekhar and Rahnev 2021). According to the boundary updating, however, the absence, not the presence, of history effects in confidence reports may indicate metacognitive inefficiency.

Recent studies have been reporting diverse history effects on choice (Fründ, Wichmann et al. 2014, Fritsche, Mostert et al. 2017, Braun, Urai et al. 2018, Bosch, Fritsche et al. 2020, Lak, Hueske et al. 2020) but largely neglected the history effects on decision uncertainty. Decision uncertainty is considered one of the core cognitive quantities that enable human intelligence (Behrens, Woolrich et al. 2007, McGuire, Nassar et al. 2014, Shenhav, Cohen et al. 2016, Drugowitsch, Mendonça et al. 2019). Our work points to a need to extend the current scope of studying the history effects on choice further, including decision uncertainty, to understand the nature of human intelligence better.

# Chapter 4

# General Discussion

## 4.1 A review of history effects of previous stimuli and previous choices

In perceptual decision-making tasks where trial-to-trial feedback is absent, researchers have explored how previous stimuli and previous choices contribute to the history effects on current choices. By studying the influence of these factors, we gain insights into the complex dynamics of decision-making.

### The effect of previous stimuli

Numerous studies have delved into the effects of previous stimuli on current decision-making processes. These investigations have focused on three main paradigms: serial dependence, boundary-updating, and sensory adaptation.

In the realm of serial dependence, researchers (Fischer and Whitney 2014, Liberman, Fischer et al. 2014, Bliss, Sun et al. 2017, Cicchini, Mikellidou et al. 2018, Pascucci, Mancuso et al. 2019, Fritsche, Spaak et al. 2020, Ceylan, Herzog et al. 2021, Murai and Whitney 2021, Sheehan and Serences 2022) have commonly employed a reproduction task paradigm. During this task, participants are briefly presented with a stimulus, followed by a reproduction cue. Their goal is to adjust the cue to replicate a specific feature of the original stimulus, such as its orientation. The observed trend in this paradigm is an attractive bias, where current reproductions tend to be drawn towards previous stimuli.

In contrast, the boundary-updating studies (Norton, Fleming et al. 2017, Hachen, Reinartz et al. 2021) typically employ a two-alternative forced choice task (2AFC) with a single stimulus per trial.

Participants are asked to determine whether a particular feature of the stimulus is greater or smaller than a specific decision boundary in a one-dimensional feature space. The outcome here shows a repulsive bias, wherein current choices tend to move away from previous stimuli.

Additionally, the sensory adaptation paradigm (Kohn 2007, Nakashima and Sugita 2017, Weber, Krishnamurthy et al. 2019, Fritsche, Solomon et al. 2022) involves a comparison task where subjects must choose the larger of two stimuli. However, one of these stimuli has been previously adapted by a long-term adaptor within the same trial. The perception of the adapted stimulus in this paradigm shows a repulsion effect, causing it to be perceived as different from the quantity of the adaptor.

In summary, the effects of previous stimuli on current choices can be either attractive, as observed in serial dependence studies, or repulsive, as seen in boundary-updating and sensory adaptation studies. These insights shed light on the intricate dynamics of decision-making processes influenced by stimulus history.

### The effect of previous choices

Furthermore, researchers have examined the effects of previous choices on decision-making processes, focusing on the phenomenon known as choice repetition or alternation bias (Akaishi, Umeda et al. 2014, Urai, Braun et al. 2017, Braun, Urai et al. 2018, Urai and Donner 2022). These investigations typically employ a two-alternative forced choice (2AFC) task with a single stimulus within each trial. Participants are tasked with determining whether the stimulus is located on one side or the other of a decision boundary.

Previous studies have explored the relationship between current choices and the choices made in preceding trials, yielding varied results. Some studies have identified a prominent choice repetition bias across subjects (Akaishi, Umeda et al. 2014, Urai, Braun et al. 2017), indicating a tendency to choose the same option as in the previous trial. However, other studies have shown inconsistent choice biases across subjects, with no clear pattern of

choice repetition (Braun, Urai et al. 2018, Urai and Donner 2022).

In summary, the effects of previous choices on current decision-making processes have been investigated through the examination of choice repetition or alternation bias. The findings from these studies reveal a diverse range of outcomes, with some studies highlighting a significant bias towards choice repetition across subjects, while others demonstrate no consistent bias. These investigations shed light on the intricate relationship between previous choices and current decision-making in 2AFC tasks.

### Disentangling the effects of previous stimuli and previous choices

As I mentioned earlier, the effects of previous stimuli and previous choices have been extensively studied over the past decade. However, the results from these studies are not consistent with each other. The association between previous stimuli, previous choices, and current choices can be either positive or negative depending on the specific task conditions. This lack of coherence in findings may stem from methodological issues in previous studies, as they often failed to disentangle the effects of previous stimuli and previous choices. Instead, they focused on investigating one factor while neglecting the influence of the other. This oversight can lead to confusion in interpreting the results, as stimuli and choices are strongly correlated, making it easy to attribute the effects of one to the other without proper consideration.

Recent studies have addressed this methodological issue by concurrently examining the effects of previous stimuli and previous choices while disentangling their individual impacts (Fornaciai and Park 2019, Pascucci, Mancuso et al. 2019, Bosch, Fritsche et al. 2020, Moon and Kwon 2022, Sheehan and Serences 2022). Strikingly, regardless of the specific task paradigm used (2AFC, reproduction, or comparison tasks), these recent studies have reported a consistent finding. They have found that current choices are repelled from previous stimuli and attracted towards previous choices.

This observed pattern of stimulus repulsion and choice attraction contradicts the findings of serial dependence studies,

where current choices are attracted towards previous stimuli. However, in the reproduction task, Moon and Kwon successfully disentangled the effects of previous stimuli and previous choices. Their study revealed that all 32 subjects exhibited consistent history effects, wherein the current reproduction was repelled from previous stimuli but attracted towards previous choices. Consequently, the assimilative effect of previous stimuli, as reported in previous serial dependence studies, was actually attributed to the assimilative effect of previous choices.

These recent investigations shed light on the interplay between previous stimuli and previous choices, providing a more comprehensive understanding of their respective influences on current decision-making processes.

### Unresolved question

However, the origin of the repulsive bias associated with previous stimuli during the reproduction task remains unclear. While my study showed that the repulsive bias during classification tasks can be explained by boundary-updating rather than sensory adaptation, it becomes challenging to explain the repulsive bias in the reproduction task (Pascucci, Mancuso et al. 2019, Moon and Kwon 2022). This is because the reproduction task does not require the presence of a decision boundaryTherefore, the existence of the repulsive bias in the reproduction task seems to suggest the involvement of sensory adaptation as the underlying mechanism.

However, both our study and another brain imaging study do not fully support sensory adaptation as the origin of the repulsive bias (Sheehan and Serences 2022). While we found sensory adaptation effects in visual cortices, we were unable to establish a direct causal link between sensory adaptation and behavior. Sheehan and Serences (2022) conducted a modeling approach and proposed that the bias induced by sensory adaptation might be compensated for by downstream areas. Consequently, the findings from neuroimaging studies cast doubt on the notion that the repulsive bias during the reproduction task can be solely attributed to sensory

adaptation.

Interestingly, the boundary-updating explanation still holds promise as a potential reason for the repulsive bias in the reproduction task. It is possible that individuals may update natural boundaries, such as cardinal orientations in space, which could induce a repulsive bias (Gibson and Radner 1937). Therefore, future studies are needed to further investigate and clarify the true origin of the repulsive bias during the reproduction task.

In summary, the origin of the repulsive bias associated with previous stimuli during the reproduction task remains uncertain. While the boundary-updating explanation may not fully account for it, findings from neuroimaging studies challenge the idea that sensory adaptation is the sole underlying mechanism. The possibility of natural boundary updating influencing the repulsive bias warrants further exploration. Future research endeavors are necessary to shed light on the underlying mechanisms and provide a comprehensive understanding of the repulsive bias during the reproduction task.

## 4.2 A new perspective on the system-level neural processing of perceptual decision-making of the relative structure

I developed a formal framework for classification inferring the boundary and examined its implications for choices and decision uncertainty, particularly in relation to the history effect. Through my research, I discovered that the neural signals related to the boundary are represented in the inferior parietal lobule (IPL) and superior temporal gyrus (STG), while the decision variable signal is represented in the STG. In this concluding section, I will present a fresh perspective on PDM and elaborate on how this perspective enhances our understanding of previous viewpoints on PDM.

Imagine yourself as a hunter living in prehistoric times who encounters a wild horse in the woods (Fig. 15). Here, you are

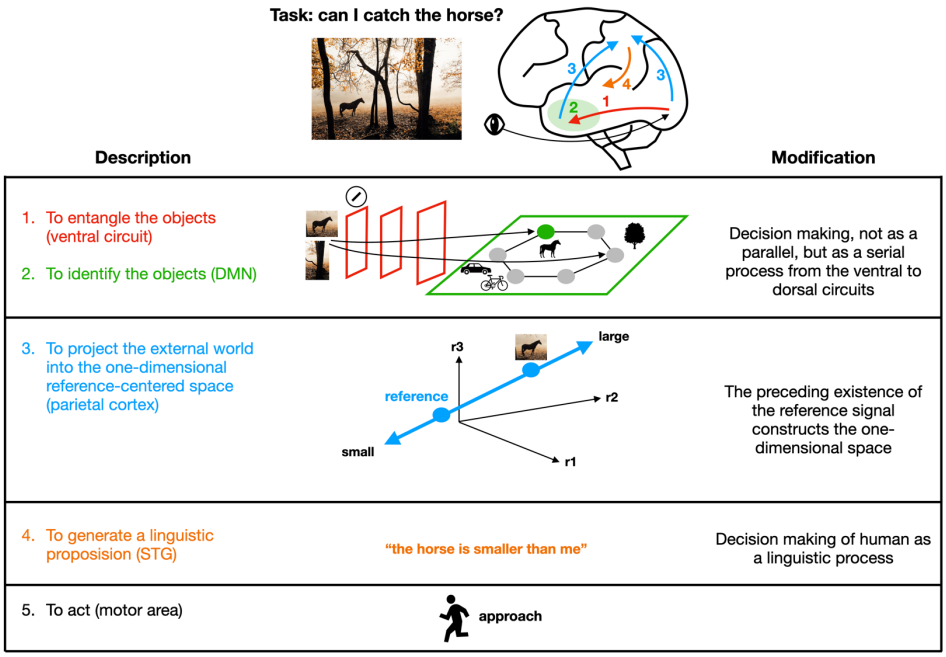


**Figure 15.** A schematic overview of the visual processing for decision-making. The visual processing is composed of five stages. The left and the middle columns indicate the descriptions of the stages. The right column indicate the modification of the proposed perspective from previous ones.

presented with a decision: Should you approach and capture the horse or conserve your energy by not doing so? This scenario exemplifies perceptual decision-making, where the outcome of your decision directly influences subsequent actions. The question arises: how does the human brain carry out such decision-making processes?

### The ventral stream: entangling raw sensory signals into separable categorical signals

V1 is the first cortical region external visual stimuli activate (DiCarlo, Zoccolan et al. 2012). V1 responds to basic features of visual objects, such as orientation (Hubel and Wiesel 1962) and spatial frequency (Tootell, Silverman et al. 1981). The neural responses in V1 are then transmitted to the next cortical region called V2, where the dorsal and ventral visual streams begin to differentiate (Kravitz, Saleem et al. 2011).

Representations of object identity emerge within the ventral visual stream (DiCarlo, Zoccolan et al. 2012). Specifically, in both humans and monkeys, object identity representations are found in the inferior temporal cortex (IT), a region in the ventral visual stream (Kriegeskorte, Mur et al. 2008). The crucial evidence supporting the role of IT in representing object identities is that the population neural responses in IT maintain object identities regardless of changes in viewing conditions (Hung, Kreiman et al. 2005, Quiroga, Reddy et al. 2005).

The question of how IT represents object identities has been a longstanding challenge (DiCarlo and Cox 2007). The neural mechanism responsible for object recognition must disentangle the overlapping neural response patterns associated with different objects. Recent studies have shown that this disentanglement can be achieved through the use of deep convolutional neural networks (CNNs), non-linear, multi-layered, feedforward neural networks (Yamins, Hong et al. 2014). The success of CNNs suggests that the complex task of disentangling neural population responses can be accomplished by combining feedforward and non-linear computations without the need for extensive recurrent communication, attention mechanisms, or complex coding schemes involving precise spike timing or synchrony (DiCarlo, Zoccolan et al. 2012).

In summary, the ventral visual circuit seems dedicated to disentangling the complex neural population responses to visual stimuli by separating them into distinct neural population responses based on object identities using multi-layered non-linear computations.

### The default mode network: to confer the meaning on the untangled population responses

The ventral visual stream plays a crucial role in transforming the compound representation of orientations and spatial frequencies, as encoded by neurons in V1, into a categorical identity representation by neurons in the inferior temporal cortex (IT). While the ventral stream effectively separates neural responses based on object

identities, it is important to consider the relationship between different identities when representing identity (Behrens, Muller et al. 2018). In other words, we expect that the neural representation of the semantic meaning of a car is more similar to that of a bicycle than that of a tree, reflecting the similarity in semantic meaning. Research suggests that the neural responses of the default mode network (DMN) construct a highly structured concept space that resembles the one constructed by the grid-cell organization found in the hippocampus (Constantinescu, O'Reilly et al. 2016, Garvert, Dolan et al. 2017). Based on this, it is conjectured that the categorical signal in IT may induce a population neural response in the concept space of the DMN, thereby conferring reliable semantic meaning to objects irrespective of the observer's viewing perspective.

However, although the ventral information processing is useful for recognizing object identities, it imposes a fundamental constraint when estimating specific features of an object. For instance, the ventral processing would yield a similar representation for a horse regardless of its size (Fig. 15). An accurate estimation of the horse's size is required to effectively determine whether the horse is capturable for a hunter. Therefore, an additional neural mechanism, located in the parietal cortex, is needed to estimate the magnitude of specific features of an object.

### The parietal cortex: to project the external world into the one-dimensional reference-centered space

The parietal cortex, which occupies a significant portion of the dorsal stream, serves two major functions: 1) representing visual objects in a one-dimensional space (Ganguli, Bisley et al. 2008, Fitzgerald, Freedman et al. 2013, Summerfield, Luyckx et al. 2020) and 2) representing the relationships between objects (Chafee, Averbeck et al. 2007, Bottini and Doeller 2020, Summerfield, Luyckx et al. 2020).

The one-dimensional space representation refers to the fact that neurons in the parietal cortex have similar preferences for visual stimuli. For example, when neurons learn associations between visual stimuli in the lateral intra-parietal cortex (LIP), their

preferences become nearly identical (Fitzgerald, Freedman et al. 2013). The parietal cortex is also known for its proportional responses to the magnitude of stimuli such as space, time, and size (Walsh 2003, Luyckx, Nili et al. 2019). This suggests that the parietal cortex reduces external environments to a one-dimensional magnitude space.

Additionally, the parietal cortex encodes the relationships between objects. Patients with parietal cortex damage struggle to accurately reproduce the spatial relationships of stimuli (Black and Strub 1976). Parietal neurons represent the relative position of an object to a reference rather than the absolute position of the object (Chafee, Averbeck et al. 2007, Sheahan, Luyckx et al. 2021). Furthermore, for spatial navigation, the parietal cortex represents the location of objects relative to the self (Colby and Goldberg 1999, Schindler and Bartels 2013).

However, I propose that these two properties reflect different facets of a unified process: the parietal cortex represents the one-dimensional magnitude of a target feature relative to a reference (Fig. 15). The magnitude representation function reflects a specific condition when the reference point is fixed to a stable point, while the relational representation function reflects a specific condition when one of the objects serves as the reference.

One important question is how the brain represents the reference. My work partially addresses this question by showing that the reference is explicitly represented in the parietal cortex before stimulus presentation (Fig. 15). I speculate that this preceding reference signal updates the prior reference point in the one-dimensional space within the parietal cortex. Subsequently, the stimulus is projected onto this modified one-dimensional space.

In summary, I suppose that after identifying the identity of the target object through the ventral circuit and DMN, the brain constructs a one-dimensional feature space centered on a reference within the parietal cortex. The target object is then projected onto this space to extract the magnitude of the feature of interest. For example, after a hunter recognizes a horse in the woods through the

ventral and DMN circuits, the relative size of the horse compared to the hunter is estimated by projecting the sensory signal of the horse onto the one-dimensional size space in the parietal cortex, which is normalized based on the size of the hunter. If the horse is smaller than the hunter, the hunter will approach to capture it.

Is the parietal cortex the region where the final decision output, determining whether the horse is larger than the hunter, is generated? The parietal cortex appears to possess all the necessary elements for making the decision, as it represents the relative size of the horse. Single-cell studies in non-human primates have indeed found the parietal cortex to be the locus for representing the decision variable (Roitman and Shadlen 2002, Gold and Shadlen 2007, Zhou and Freedman 2019). Surprisingly, there are few reports on the decision variable in the human parietal cortex (Kahnt, Grueschow et al. 2011, Hebart, Schriever et al. 2014). I actually found the decision variable outside of the parietal cortex, specifically in the superior temporal gyrus (STG). I will discuss the implications of why I found the decision variable in the STG of the human brain.

### STG: to generate a language-based decision

I propose that the discrepancy between humans and non-human primates in representing the decision variable arises from humans' language-based decision-making processes (Fodor 1975, Frankland and Greene 2020). Specifically, I suggest that human perceptual decision-making involves a commitment to linguistic propositions about the stimulus, whereas non-human primates' decision-making does not involve a linguistic process. Non-human primates are trained in the task by learning the stimulus-decision contingency through reward-based reinforcement learning. In contrast, humans can learn the task through language-based instructions even when no rewards are involved. As a result, I expect humans to rely heavily on language-based processing for performing the task, while non-human primates would rely solely on the reinforced stimulus-choice contingency.

Interestingly, the superior temporal gyrus (STG), where I

found the decision variable, is well-known for its contribution to language processing (Willems, Özyürek et al. 2009). In language-based tasks, STG not only specifies the semantic meaning of individual words but also specifies the relational meaning between words by representing the target object (e.g., patient) and the reference object (e.g., agent) in a sentence (Frankland and Greene 2015, Frankland and Greene 2020). Therefore, even in a perceptual decision-making task, I speculate that STG represents not only the magnitudes of the target and reference objects separately but also their relational meaning. Thus, I propose that STG may be the region where decisions are generated in the human brain.

Non-human primate studies have not reported the presence of the decision variable in STG. Thus, further research is needed to examine the involvement of the STG in decision-making among non-human primates.

In summary, I propose a system-level, five-stage mechanism for making a perceptual decision (Fig. 15). First, a complex visual scene consisting of an entangled manifold of objects is untangled based on object identities through the ventral visual stream. Second, the identities of the untangled visual signals are recognized through interaction with the cognitive map in DMN. Third, the target feature of the identified object is projected onto the reference-centered one-dimensional space in the parietal cortex. Fourth, a decision is generated as a linguistic proposition in STG. Finally, the decision is put into action via the motor execution system.

## Modifications

In this section, I will clarify the modifications I made to the previous perspectives on perceptual decision-making.

First, I propose that the ventral and dorsal visual circuits are not parallel but interact serially. Conventionally, the ventral and dorsal pathways have been considered parallel circuits, with the ventral pathway dedicated to object identity recognition and the dorsal pathway involved in representing the structure between objects. This view suggests that the two pathways operate

independently (Kravitz, Saleem et al. 2013, Bottini and Doeller 2020, Summerfield, Luyckx et al. 2020). However, I suggest that these two pathways work together through a sequential process to serve a single purpose. I believe that the ventral stream identifies the external world and provides the parietal cortex with a relevant context for the current condition. Using this context identified by the ventral stream, the parietal cortex constructs a one-dimensional reference-centered space.

Second, I propose that the signal of an internal reference that is present before a stimulus may play a crucial role in constructing the one-dimensional reference-centered space in the parietal cortex. Previous studies have reported that the parietal cortex represents the position of a stimulus in the one-dimensional space relative to a reference point, but it remains unclear how the reference-centered space is constructed in the parietal cortex (Colby and Goldberg 1999, Chafee, Averbeck et al. 2007, Schindler and Bartels 2013, Sheahan, Luyckx et al. 2021). Therefore, I believe that the preceding reference signal is a crucial factor in flexibly reconstructing the subspaces of the parietal cortex based on the current context.

Third, I propose that the decision-making processes in the human brain differ from those in non-human primates' brains because humans rely on language-based processes to make decisions. This conjecture is supported by previous findings that the brain signals representing the decision variable are not well detected in the parietal cortex of humans but are readily detected in that of non-human primates (Kahnt, Grueschow et al. 2011, Hebart, Schriever et al. 2014). Additionally, I found that the decision variable is detected in the superior temporal gyrus (STG), a region involved in the language processing (Frankland and Greene 2015, Frankland and Greene 2020).

However, I must note that all three modifications mentioned above require further empirical evidence. It is my hope that future research will thoroughly investigate and refine the hypotheses proposed in my thesis.

## 4.3 Limitations

However, I must note that all three modifications mentioned above require further empirical evidence.

### Generalizability of the identified brain regions

It remains uncertain whether the identified network is also observed in other stimulus domains and modalities. It is possible that a similar network can be identified in perceptual classification tasks, as the identified regions are not specific to a particular modality. In particular, the parietal cortex has been known to encode the magnitude of stimuli, regardless of whether it is related to time, space, numerosity, or reward (Walsh 2003, Luyckx, Nili et al. 2019). However, it is unclear whether the same network is involved in decision-making processes related to factors other than perceptual quantity, such as value, as the brain region responsible for encoding value is distinct from that involved in processing perceptual quantities (Padoa-Schioppa and Assad 2006).

### The role of ventral stream

Furthermore, the role of ventral stream should also be more examined. In a study conducted by Milner, Harvey, and Pritchard in 1998, a patient who had experienced a right hemisphere stroke primarily affecting the occipito-temporal area was found to exhibit under-perception of object size. However, this abnormality did not extend to the processing of size for visuomotor control (Milner, Harvey et al. 1998). The findings suggest that size estimation impairments can occur as a result of damage to the ventral stream without significant disruption of the dorsal stream. In other words, the results indicate the critical role of the ventral stream in size estimation.

While it is widely accepted that the parietal lobe plays a crucial role in visuospatial neglect symptoms, as noted by Milner and colleagues (Milner, Harvey et al. 1998), it would be simplistic to

attribute magnitude estimation solely to the parietal cortex. Further research is needed to investigate the involvement of the ventral stream in processing visual magnitude.

### Validation of the decoding results by using other methods

I employed the searchlight approach to detect the brain signals associated with the latent variables of BMBU. The searchlight approach operates under the assumption that local patterns of brain activity encode the target representations (Haynes 2015). However, it is important to acknowledge that the searchlight approach relies on several assumptions. If the brain representation of the target variables violates these assumptions, it may lead to potentially misleading results (Etzel, Zacks et al. 2013). Consequently, it is necessary to validate the findings using alternative methodological approaches.

One promising alternative approach is predictive modeling (Woo, Chang et al. 2017). In contrast to the searchlight approach, predictive modeling utilizes all available brain data to generate the best possible prediction of the target variable. This approach avoids multiple comparisons and increases statistical power (Reddan, Lindquist et al. 2017). Consequently, predictive modeling provides a comprehensive weight map of the entire brain, indicating the contribution of each voxel in predicting the latent variable (Woo et al., 2017).

However, the effectiveness of applying predictive models to cognitive science studies remains uncertain. Predictive models were originally developed for clinical purposes, where they are trained using leave-one-subject-out cross-validation and subsequently tested by evaluating their performance on new, out-of-sample subjects (Chang, Gianaros et al. 2015, Woo, Chang et al. 2017). This procedure implies that predictive models have typically been developed with the assumption that a shared pattern across subjects underlies clinical symptoms. This assumption may not align with the goals of cognitive science studies, which aim to identify brain regions encoding target variables regardless of the presence of shared

patterns across subjects.

### A distinction between the similarity-based and the boundary-based decision-making scenarios

I propose a distinction should be made between two distinct decision-making scenarios. The first scenario involves conventional decision-making based on similarity-based identity (Gold and Shadlen 2007). In this scenario, observers must acquire knowledge of the distribution of items within specific categories (Rosch 1973, Kamp and Partee 1995, Seger and Miller 2010, Nosofsky 2011, Douven, Decock et al. 2013). For instance, when determining whether an animal is a *cat* or a *dog*, it is necessary to develop an understanding of the typical visual characteristics exhibited by samples from each category and make a judgment about the extent of similarity between the present animal and cats or dogs (Fig. 16*A*). In this particular scenario, it is possible to describe the visual attributes of cats and dogs without explicitly establishing a boundary appearance that demarcates the two categories.

On the other hand, the second scenario involves decision-making based on boundary-based identity. In this context, observers are required to acquire knowledge about the specific location of category boundaries within a one-dimensional magnitude space that is relevant to the task at hand (Rips and Turnbull 1980, Kennedy 2007, Tribushinina 2011, Sheahan, Luyckx et al. 2021). For example, when deciding whether a dog is *small* or *large*, individuals need to possess knowledge of the typical size range associated with dogs and then determine whether the size of the given animal falls below or exceeds that range (Fig. 16*B*). In this particular scenario, the distinction between small and large dogs cannot be defined without the establishment of a size boundary.

Interestingly, this semantic distinction between similarity-based and boundary-based identities, which has long been

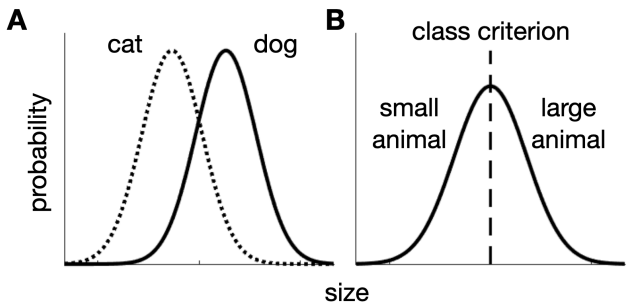


**Figure 16.** The two decision-making scenarios. ***A,*** A similarity-based decision-making scenario. The distributions of cat and dogs are maintained, respectively. ***B,*** A boundary-based decision-making scenario. The class criterion divides the single distribution of dogs into small and large dogs.

appreciated by linguists (Solt 2015), has often been overlooked within decision-making research, where similarity-based models have predominantly prevailed. For instance, signal detection theory (SDT), a widely accepted model of binary perceptual decision-making (Green and Swets 1966, Gold and Shadlen 2007, Rahnev and Denison 2018), frames binary decision-making as an assessment of the degree of similarity between two category items and the present stimulus. However, SDT, by its very nature, may not be ideally suited for capturing the intricacies of boundary-based identification, as it does not explicitly incorporate the concept of category boundaries into its framework.

While the differentiation between the boundary-based and similarity-based decision-making models appears intuitive, the necessity of the boundary-based model in explaining the observed repulsive bias in my research remains uncertain. It is plausible that the similarity-based model alone could sufficiently account for the repulsive bias. Notably, a prior modeling study indicated that the similarity-based model appeared capable of explaining the repulsive bias (Norton, Fleming et al. 2017). Consequently, further investigations are warranted to ascertain whether the boundary-based model is indispensable or if the similarity-based model alone is satisfactory for elucidating the repulsive bias.

It is my hope that future research will thoroughly investigate and refine the hypotheses proposed in my thesis.

# Bibliography


Akaishi, R., K. Umeda, A. Nagase and K. Sakai (2014). "Autonomous mechanism of internal choice estimate underlies decision inertia." Neuron 81(1): 195-206.

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

Apresjan, J. D. (1974). "Regular polysemy."

Ashburner, J. (2007). "A fast diffeomorphic image registration algorithm." Neuroimage 38(1): 95-113.

Baumann, O., R. J. Borra, J. M. Bower, K. E. Cullen, C. Habas, R. B. Ivry, M. Leggio, J. B. Mattingley, M. Molinari and E. A. Moulton (2015). "Consensus paper: the role of the cerebellum in perceptual processes." The Cerebellum 14(2): 197-220.

Behrens, T. E., T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld and Z. Kurth-Nelson (2018). "What is a cognitive map? Organizing knowledge for flexible behavior." Neuron 100(2): 490-509.

Behrens, T. E., M. W. Woolrich, M. E. Walton and M. F. Rushworth (2007). "Learning the value of information in an uncertain world." Nature neuroscience 10(9): 1214.

Benjamin, A. S., M. Diaz and S. Wee (2009). "Signal detection with criterion noise: Applications to recognition memory." Psychological review 116(1): 84.

Benjamini, Y. and Y. Hochberg (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.

Bierwisch, M. (1989). "The semantics of gradation." Dimensional adjectives 71(261): 35.

Bierwisch, M. (2009). The semantics of gradation, Universitätsbibliothek Johann Christian Senckenberg.

Black, F. W. and R. L. Strub (1976). "Constructional apraxia in patients with discrete missile wounds of the brain." Cortex 12(3): 212-220.

Bliss, D. P., J. J. Sun and M. D'Esposito (2017). "Serial dependence is absent at the time of perception but increases in visual working memory." Scientific Reports 7(1): 14739.

Borcard, D., P. Legendre and P. Drapeau (1992). "Partialling out the spatial component of ecological variation." Ecology 73(3): 1045-1055.

Bosch, E., M. Fritsche, B. V. Ehinger and F. P. de Lange (2020). "Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias." Journal of vision 20(12): 9-9.

Bottini, R. and C. F. Doeller (2020). "Knowledge across reference frames: Cognitive maps and image spaces." Trends in Cognitive Sciences 24(8): 606-619.

Braun, A., A. E. Urai and T. H. Donner (2018). "Adaptive history biases result from confidence-weighted accumulation of past choices." Journal of Neuroscience 38(10): 2418-2429.

Bromiley, P. (2003). "Products and convolutions of Gaussian probability density functions." Tina-Vision Memo 3(4): 1.
Brown, J. W. and T. S. Braver (2005). "Learned predictions of error likelihood in the anterior cingulate cortex." Science 307(5712): 1118-1121.
Buracas, G. T. and G. M. Boynton (2002). "Efficient design of event-related fMRI experiments using M-sequences." Neuroimage 16(3): 801-813.
Carlson, J. M., D. Foti, L. R. Mujica-Parodi, E. Harmon-Jones and G. Hajcak (2011). "Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: a combined ERP and fMRI study." Neuroimage 57(4): 1608-1616.
Carter, C. S., T. S. Braver, D. M. Barch, M. M. Botvinick, D. Noll and J. D. Cohen (1998). "Anterior cingulate cortex, error detection, and the online monitoring of performance." Science 280(5364): 747-749.
Cavanagh, J. F. and M. J. Frank (2014). "Frontal theta as a mechanism for cognitive control." Trends in cognitive sciences 18(8): 414-421.
Ceylan, G., M. H. Herzog and D. Pascucci (2021). "Serial dependence does not originate from low-level visual processing." Cognition 212: 104709.
Chafee, M. V., B. B. Averbeck and D. A. Crowe (2007). "Representing spatial relationships in posterior parietal cortex: single neurons code object-referenced position." Cerebral Cortex 17(12): 2914-2932.
Chang, L. J., P. J. Gianaros, S. B. Manuck, A. Krishnan and T. D. Wager (2015). "A sensitive and specific neural signature for picture-induced negative affect." PLoS biology 13(6): e1002180.
Chase, D. P. (2019). Aristotle: Nicomachean Ethics, e-artnow.
Cho, R. Y., L. E. Nystrom, E. T. Brown, A. D. Jones, T. S. Braver, P. J. Holmes and J. D. Cohen (2002). "Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task." Cognitive, Affective, & Behavioral Neuroscience 2(4): 283-299.
Choe, K. W., R. Blake and S.-H. Lee (2014). "Dissociation between neural signatures of stimulus and choice in population activity of human V1 during perceptual decision-making." Journal of Neuroscience 34(7): 2725-2743.
Choe, K. W., R. Blake and S.-H. Lee (2016). "Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation." Vision research 118: 48-59.
Chomsky, N. (2014). Aspects of the Theory of Syntax, MIT press.
Cicchini, G. M., K. Mikellidou and D. C. Burr (2018). "The functional role of serial dependence." Proceedings of the Royal Society B 285(1890): 20181722.
Clifford, C. W., M. A. Webster, G. B. Stanley, A. A. Stocker, A. Kohn, T. O. Sharpee and O. Schwartz (2007). "Visual adaptation: Neural, psychological and computational aspects." Vision research 47(25): 3125-3131.
Colby, C. L. and M. E. Goldberg (1999). "Space and attention in parietal cortex." Annual review of neuroscience 22(1): 319-349.
Collins, T. (2021). "Serial dependence occurs at the level of both features and integrated object representations." Journal of Experimental Psychology: General.

Constantinescu, A. O., J. X. O'Reilly and T. E. Behrens (2016). "Organizing conceptual knowledge in humans with a gridlike code." Science 352(6292): 1464-1468.
Cox, R. W. (1996). "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages." Computers and Biomedical research 29(3): 162-173.
de Gee, J. W., T. Knapen and T. H. Donner (2014). "Decision-related pupil dilation reflects upcoming choice and individual bias." Proceedings of the National Academy of Sciences 111(5): E618-E625.
Dehaene, S. and J.-P. Changeux (2011). "Experimental and theoretical approaches to conscious processing." Neuron 70(2): 200-227.
DiCarlo, J. J. and D. D. Cox (2007). "Untangling invariant object recognition." Trends in cognitive sciences 11(8): 333-341.
DiCarlo, J. J., D. Zoccolan and N. C. Rust (2012). "How does the brain solve visual object recognition?" Neuron 73(3): 415-434.
Douven, I., L. Decock, R. Dietz and P. Égré (2013). "Vagueness: A conceptual spaces approach." Journal of Philosophical Logic 42(1): 137-160.
Drugowitsch, J., A. G. Mendonça, Z. F. Mainen and A. Pouget (2019). "Learning optimal decisions with confidence." Proceedings of the National Academy of Sciences 116(49): 24872-24880.
Dyjas, O., K. M. Bausenhart and R. Ulrich (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(8): 1819-1841.
Elwert, F. and C. Winship (2014). "Endogenous selection bias: The problem of conditioning on a collider variable." Annual review of sociology 40: 31-53.
Engel, S. A., D. E. Rumelhart, B. A. Wandell, A. T. Lee, G. H. Glover, E.-J. Chichilnisky and M. N. Shadlen (1994). "fMRI of human visual cortex." Nature.
Etzel, J. A., J. M. Zacks and T. S. Braver (2013). "Searchlight analysis: promise, pitfalls, and potential." Neuroimage 78: 261-269.
Fietz, J., D. Pöhlchen, F. P. Binder, B. W. Group, M. Czisch, P. G. Sämann and V. I. Spoormaker (2022). "Pupillometry tracks cognitive load and salience network activity in a working memory functional magnetic resonance imaging task." Human Brain Mapping 43(2): 665-680.
Fischer, J. and D. Whitney (2014). "Serial dependence in visual perception." Nature neuroscience 17(5): 738.
Fitzgerald, J. K., D. J. Freedman, A. Fanini, S. Bennur, J. I. Gold and J. A. Assad (2013). "Biased associative representations in parietal cortex." Neuron 77(1): 180-191.
Fodor, J. A. (1975). The language of thought, Harvard university press.
Fornaciai, M. and J. Park (2019). "Spontaneous repulsive adaptation in the absence of attractive serial dependence." Journal of vision 19(5): 21-21.
Frankland, S. M. and J. D. Greene (2015). "An architecture for encoding sentence meaning in left mid-superior temporal cortex." Proceedings of the National Academy of Sciences 112(37): 11732-11737.

Frankland, S. M. and J. D. Greene (2020). "Concepts and compositionality: in search of the brain's language of thought." Annual review of psychology 71(1): 273-303.
Friston, K. J., S. Williams, R. Howard, R. S. Frackowiak and R. Turner (1996). "Movement-related effects in fMRI time-series." Magnetic resonance in medicine 35(3): 346-355.
Fritsche, M., P. Mostert and F. P. de Lange (2017). "Opposite effects of recent history on perception and decision." Current Biology 27(4): 590-595.
Fritsche, M., S. G. Solomon and F. P. de Lange (2022). "Brief stimuli cast a persistent long-term trace in visual cortex." Journal of Neuroscience 42(10): 1999-2010.
Fritsche, M., E. Spaak and F. P. de Lange (2020). "A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception." Elife 9.
Fründ, I., F. A. Wichmann and J. H. Macke (2014). "Quantifying the effect of intertrial dependence on perceptual decisions." Journal of vision 14(7): 9-9.
Galluzzi, F., A. Benedetto, G. M. Cicchini and D. C. Burr (2022). "Visual priming and serial dependence are mediated by separate mechanisms." Journal of Vision 22(10): 1-1.
Ganguli, S., J. W. Bisley, J. D. Roitman, M. N. Shadlen, M. E. Goldberg and K. D. Miller (2008). "One-dimensional dynamics of attention and decision making in LIP." Neuron 58(1): 15-25.
Gao, J.-H., L. M. Parsons, J. M. Bower, J. Xiong, J. Li and P. T. Fox (1996). "Cerebellum implicated in sensory acquisition and discrimination rather than motor control." Science 272(5261): 545-547.
Garvert, M. M., R. J. Dolan and T. E. Behrens (2017). "A map of abstract relational knowledge in the human hippocampal–entorhinal cortex." Elife 6: e17086.
Gibson, J. J. and M. Radner (1937). "Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies." Journal of experimental psychology 20(5): 453.
Gold, J. I. and M. N. Shadlen (2007). "The neural basis of decision making." Annual review of neuroscience 30.
Gorgoraptis, N., R. F. Catalao, P. M. Bays and M. Husain (2011). "Dynamic updating of working memory resources for visual objects." Journal of Neuroscience 31(23): 8502-8511.
Green, D. M. and J. A. Swets (1966). Signal detection theory and psychophysics. Oxford, England, John Wiley.
Grinband, J., J. Hirsch and V. P. Ferrera (2006). "A neural representation of categorization uncertainty in the human brain." Neuron 49(5): 757-763.
Hachen, I., S. Reinartz, R. Brasselet, A. Stroligo and M. Diamond (2021). "Dynamics of history-dependent perceptual judgment." Nature communications 12(1): 1-15.
Hanks, T. D., C. D. Kopec, B. W. Brunton, C. A. Duan, J. C. Erlich and C. D. Brody (2015). "Distinct relationships of parietal and prefrontal cortices to evidence accumulation." Nature 520(7546): 220.
Haynes, J.-D. (2015). "A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives." Neuron 87(2): 257-270.

Hebart, M. N., Y. Schriever, T. H. Donner and J.-D. Haynes (2014). "The relationship between perceptual decision variables and confidence in the human brain." Cerebral Cortex 26(1): 118-130.
Holroyd, C. B., S. Nieuwenhuis, N. Yeung, L. Nystrom, R. B. Mars, M. G. Coles and J. D. Cohen (2004). "Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals." Nature neuroscience 7(5): 497.
Hubel, D. H. and T. N. Wiesel (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." The Journal of physiology 160(1): 106.
Hung, C. P., G. Kreiman, T. Poggio and J. J. DiCarlo (2005). "Fast readout of object identity from macaque inferior temporal cortex." Science 310(5749): 863-866.
Huttenlocher, J., E. T. Higgins and H. H. Clark (1971). "Adjectives, comparatives, and syllogisms." Psychological Review 78(6): 487.
Jenkinson, M., P. Bannister, M. Brady and S. Smith (2002). "Improved optimization for the robust and accurate linear registration and motion correction of brain images." Neuroimage 17(2): 825-841.
Jentzsch, I. and C. Dudschig (2009). "Short article: Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing." Quarterly Journal of Experimental Psychology 62(2): 209-218.
Kahnt, T., M. Grueschow, O. Speck and J.-D. Haynes (2011). "Perceptual learning and decision-making in human medial frontal cortex." Neuron 70(3): 549-559.
Kahnt, T., J. Heinzle, S. Q. Park and J.-D. Haynes (2011). "Decoding different roles for vmPFC and dlPFC in multi-attribute decision making." Neuroimage 56(2): 709-715.
Kamp, H. and B. Partee (1995). "Prototype theory and compositionality." Cognition 57(2): 129-191.
Kass, R. E. and A. E. Raftery (1995). "Bayes factors." Journal of the american statistical association 90(430): 773-795.
Katz, J. J. (1972). "Semantic theory."
Kennedy, C. (2007). "Vagueness and grammar: The semantics of relative and absolute gradable adjectives." Linguistics and philosophy 30(1): 1-45.
Kepecs, A., N. Uchida, H. A. Zariwala and Z. F. Mainen (2008). "Neural correlates, computation and behavioural impact of decision confidence." Nature 455: 227.
Kim, J.-N. and M. N. Shadlen (1999). "Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque." Nature neuroscience 2(2): 176.
Klein, E. (1980). "A semantics for positive and comparative adjectives." Linguistics and philosophy 4(1): 1-45.
Knapen, T., M. Rolfs, M. Wexler and P. Cavanagh (2010). "The reference frame of the tilt aftereffect." Journal of Vision 10(1): 8-8.
Knill, D. C. and W. Richards (1996). Perception as Bayesian inference, Cambridge University Press.
Kohn, A. (2007). "Visual adaptation: physiology, mechanisms, and functional benefits." Journal of neurophysiology 97(5): 3155-3164.

Korn, C. W. and D. R. Bach (2016). "A solid frame for the window on cognition: Modeling event-related pupil responses." Journal of vision 16(3): 28-28.
Kravitz, D. J., K. S. Saleem, C. I. Baker and M. Mishkin (2011). "A new neural framework for visuospatial processing." Nature Reviews Neuroscience 12(4): 217-230.
Kravitz, D. J., K. S. Saleem, C. I. Baker, L. G. Ungerleider and M. Mishkin (2013). "The ventral visual pathway: an expanded neural framework for the processing of object quality." Trends in cognitive sciences 17(1): 26-49.
Kriegeskorte, N., R. Goebel and P. Bandettini (2006). "Information-based functional brain mapping." Proceedings of the National Academy of Sciences 103(10): 3863-3868.
Kriegeskorte, N., M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka and P. A. Bandettini (2008). "Matching categorical object representations in inferior temporal cortex of man and monkey." Neuron 60(6): 1126-1141.
Kristjánsson, Á. and G. Campana (2010). "Where perception meets memory: A review of repetition priming in visual search tasks." Attention, Perception, & Psychophysics 72(1): 5-18.
Lages, M. and M. Treisman (1998). "Spatial frequency discrimination: visual long-term memory or criterion setting?" Vision research 38(4): 557-572.
Lages, M. and M. Treisman (2010). "A criterion setting theory of discrimination learning that accounts for anisotropies and context effects." Seeing and perceiving 23(5): 401-434.
Lak, A., G. M. Costa, E. Romberg, A. A. Koulakov, Z. F. Mainen and A. Kepecs (2014). "Orbitofrontal cortex is required for optimal waiting based on decision confidence." Neuron 84(1): 190-201.
Lak, A., E. Hueske, J. Hirokawa, P. Masset, T. Ott, A. E. Urai, T. H. Donner, M. Carandini, S. Tonegawa and N. Uchida (2020). "Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon." ELife 9: e49834.
Lassiter, D. and N. D. Goodman (2017). "Adjectival vagueness in a Bayesian model of interpretation." Synthese 194(10): 3801-3836.
Lau, H. and D. Rosenthal (2011). "Empirical support for higher-order theories of conscious awareness." Trends in cognitive sciences 15(8): 365-373.
Lee, H., H.-J. Lee, K. W. Choe and S.-H. Lee (2023). "Neural evidence for boundary updating as the source of the repulsive bias in classification." bioRxiv: 2023.2001. 2011.523692.
Lee, H.-J., H. Lee, C. Y. Lim, I. Rhim and S.-H. Lee (2023). "What humans learn from corrective feedback for perceptual decision-making." bioRxiv: 2023.2001. 2011.523567.
Lee, S.-H., R. Blake and D. J. Heeger (2007). "Hierarchy of cortical responses underlying binocular rivalry." Nature neuroscience 10(8): 1048.
Leeper, T. J., J. Arnold and V. Arel-Bundock (2018). "Margins: Marginal effects for model objects." R package version 0.3 23: 2018.
Lehrer, A. and K. Lehrer (1982). "Antonymy." Linguistics and philosophy: 483-501.

Liberman, A., J. Fischer and D. Whitney (2014). "Serial dependence in the perception of faces." Current Biology 24(21): 2569-2574.
Lueckmann, J.-M., J. H. Macke and H. Nienborg (2018). "Can serial dependencies in choices and neural activity explain choice probabilities?" Journal of Neuroscience 38(14): 3495-3506.
Luyckx, F., H. Nili, B. Spitzer and C. Summerfield (2019). "Neural structure mapping in human probabilistic reward learning." Elife 8: e42816.
Maat, H. P. (2006). "Subjectification in gradable adjectives." Subjectification. Various paths to subjectivity 31: 279-320.
Macke, J. H. and H. Nienborg (2019). "Choice (-history) correlations in sensory cortex: cause or consequence?" Current Opinion in Neurobiology 58: 148-154.
Marco-Pallarés, J., S. V. Müller and T. F. Münte (2007). "Learning by doing: an fMRI study of feedback-related brain activations." Neuroreport 18(14): 1423-1426.
Marcus, D., J. Harwell, T. Olsen, M. Hodge, M. Glasser, F. Prior, M. Jenkinson, T. Laumann, S. Curtiss and D. Van Essen (2011). "Informatics and data mining tools and strategies for the human connectome project." Frontiers in neuroinformatics 5: 4.
McGuire, J. T., M. R. Nassar, J. I. Gold and J. W. Kable (2014). "Functionally dissociable influences on learning rate in a dynamic environment." Neuron 84(4): 870-881.
McLelland, D., B. Ahmed and W. Bair (2009). "Responses to static visual images in macaque lateral geniculate nucleus: implications for adaptation, negative afterimages, and visual fading." Journal of Neuroscience 29(28): 8996-9001.
Milner, A., M. Harvey and C. Pritchard (1998). "Visual size processing in spatial neglect." Experimental Brain Research 123: 192-200.
Mize, T. D., L. Doan and J. S. Long (2019). "A general framework for comparing predictions and marginal effects across models." Sociological Methodology 49(1): 152-189.
Moon, J. and O.-S. Kwon (2022). "Attractive and repulsive effects of sensory history concurrently shape visual perception." BMC biology 20(1): 247.
Morgan, M. (2014). "A bias-free measure of retinotopic tilt adaptation." Journal of Vision 14(1): 7-7.
Morgan, M., S. Watamaniuk and S. McKee (2000). "The use of an implicit standard for measuring discrimination thresholds." Vision research 40(17): 2341-2349.
Murai, Y. and D. Whitney (2021). "Serial dependence revealed in history-dependent perceptual templates." Current Biology 31(14): 3185-3191. e3183.
Nahum, M., L. Daikhin, Y. Lubin, Y. Cohen and M. Ahissar (2010). "From comparison to classification: a cortical tool for boosting perception." Journal of Neuroscience 30(3): 1128-1136.
Nakashima, Y. and Y. Sugita (2017). "The reference frame of the tilt aftereffect measured by differential Pavlovian conditioning." Scientific reports 7: 40525.

Nestares, O. and D. J. Heeger (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**(5): 705-715.
Norton, E. H., S. M. Fleming, N. D. Daw and M. S. Landy (2017). "Suboptimal criterion learning in static and dynamic environments." PLoS computational biology **13**(1): e1005304.
Nosofsky, R. M. (2011). "The generalized context model: An exemplar model of classification." Formal approaches in categorization: 18-39.
Olman, C. A., S. Inati and D. J. Heeger (2007). "The effect of large veins on spatial localization with GE BOLD at 3 T: Displacement, not blurring." Neuroimage **34**(3): 1126-1135.
Orhan, A. E. and W. J. Ma (2019). "A diverse range of factors affect the nature of neural representations underlying short-term memory." Nature neuroscience **22**(2): 275-283.
Padoa-Schioppa, C. and J. A. Assad (2006). "Neurons in the orbitofrontal cortex encode economic value." Nature **441**(7090): 223-226.
Partee, B. H. (2007). "Compositionality and coercion in semantics: The dynamics of adjective meaning." Cognitive foundations of interpretation: 145-161.
Pascucci, D., G. Mancuso, E. Santandrea, C. Della Libera, G. Plomp and L. Chelazzi (2019). "Laws of concatenated perception: Vision goes for novelty, decisions for perseverance." PLoS biology **17**(3): e3000144.
Patterson, C. A., S. C. Wissig and A. Kohn (2013). "Distinct effects of brief and prolonged adaptation on orientation tuning in primary visual cortex." Journal of Neuroscience **33**(2): 532-543.
Pavan, A., R. B. Marotti and G. Campana (2012). "The temporal course of recovery from brief (sub-second) adaptations to spatial contrast." Vision research **62**: 116-124.
Perquin, M. N., T. Heed and C. Kayser (2022). "Variance (un) explained: Experimental conditions and temporal dependencies explain similarly small proportions of reaction time variability in perceptual and cognitive tasks." bioRxiv: 2022.2012. 2022.521656.
Pouget, A., J. Drugowitsch and A. Kepecs (2016). "Confidence and certainty: distinct probabilistic quantities for different goals." Nature neuroscience **19**(3): 366-374.
Quiroga, R. Q., L. Reddy, G. Kreiman, C. Koch and I. Fried (2005). "Invariant visual representation by single neurons in the human brain." Nature **435**(7045): 1102-1107.
Rahnev, D. and R. N. Denison (2018). "Suboptimality in perceptual decision making." Behavioral and Brain Sciences **41**.
Rahnev, D., A. Koizumi, L. Y. McCurdy, M. D'Esposito and H. Lau (2015). "Confidence leak in perceptual decision making." Psychological science **26**(11): 1664-1680.
Raviv, O., I. Lieder, Y. Loewenstein and M. Ahissar (2014). "Contradictory behavioral biases result from the influence of past stimuli on perception." PLOS Comput Biol **10**(12): e1003948.

Reddan, M. C., M. A. Lindquist and T. D. Wager (2017). "Effect size estimation in neuroimaging." JAMA psychiatry 74(3): 207-208.
Rees, G., G. Kreiman and C. Koch (2002). "Neural correlates of consciousness in humans." Nature Reviews Neuroscience 3(4): 261-270.
Renart, A. and C. K. Machens (2014). "Variability in neural activity and behavior." Current opinion in neurobiology 25: 211-220.
Rips, L. J. and W. Turnbull (1980). "How big is big? Relative and absolute properties in memory." Cognition 8(2): 145-174.
Roitman, J. D. and M. N. Shadlen (2002). "Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task." Journal of neuroscience 22(21): 9475-9489.
Rosch, E. H. (1973). "Natural categories." Cognitive psychology 4(3): 328-350.
Rotstein, C. and Y. Winter (2004). "Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers." Natural language semantics 12: 259-288.
Sanders, J. I., B. Hangya and A. Kepecs (2016). "Signatures of a statistical computation in the human sense of confidence." Neuron 90(3): 499-506.
Schindler, A. and A. Bartels (2013). "Parietal cortex codes for egocentric space beyond the field of view." Current Biology 23(2): 177-182.
Scutari, M. (2009). "Learning Bayesian networks with the bnlearn R package." arXiv preprint arXiv:0908.3817.
Seeley, W. W. (2019). "The salience network: a neural system for perceiving and responding to homeostatic demands." Journal of Neuroscience 39(50): 9878-9882.
Seger, C. A. and E. K. Miller (2010). "Category learning in the brain." Annual review of neuroscience 33: 203-219.
Sereno, M. I., A. Dale, J. Reppas, K. Kwong, J. Belliveau, T. Brady, B. Rosen and R. Tootell (1995). "Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging." Science 268(5212): 889-893.
Sheahan, H., F. Luyckx, S. Nelli, C. Teupe and C. Summerfield (2021). "Neural state space alignment for magnitude generalization in humans and recurrent networks." Neuron 109(7): 1214-1226. e1218.
Sheehan, T. C. and J. T. Serences (2022). "Attractive serial dependence overcomes repulsive neuronal adaptation." PLoS biology 20(9): e3001711.
Shekhar, M. and D. Rahnev (2021). "Sources of metacognitive inefficiency." Trends in Cognitive Sciences 25(1): 12-23.
Shenhav, A., J. D. Cohen and M. M. Botvinick (2016). "Dorsal anterior cingulate cortex and the value of control." Nature neuroscience 19(10): 1286-1291.
Shenhav, A., M. A. Straccia, J. D. Cohen and M. M. Botvinick (2014). "Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value." Nature neuroscience 17(9): 1249.
Shmuel, A., E. Yacoub, D. Chaimow, N. K. Logothetis and K. Ugurbil (2007). "Spatio-temporal point-spread function of fMRI signal in human gray matter at 7 Tesla." Neuroimage 35(2): 539-552.

Smith, A. M., B. K. Lewis, U. E. Ruttimann, Q. Y. Frank, T. M. Sinnwell, Y. Yang, J. H. Duyn and J. A. Frank (1999). "Investigation of low frequency drift in fMRI signal." Neuroimage 9(5): 526-533.
Smith, N. K. (2011). Immanuel Kant's critique of pure reason, Read Books Ltd.
Solomon, S. G. and A. Kohn (2014). "Moving sensory adaptation beyond suppressive effects in single neurons." Current Biology 24(20): R1012-R1022.
Solt, S. (2015). "Vagueness and imprecision: Empirical foundations." Annu. Rev. Linguist. 1(1): 107-127.
Soon, C. S., M. Brass, H.-J. Heinze and J.-D. Haynes (2008). "Unconscious determinants of free decisions in the human brain." Nature neuroscience 11(5): 543-545.
Stocker, A. A. and E. P. Simoncelli (2006). Sensory adaptation within a Bayesian framework for perception. Advances in neural information processing systems.
Summerfield, C., F. Luyckx and H. Sheahan (2020). "Structure learning and the posterior parietal cortex." Progress in neurobiology 184: 101717.
Tootell, R. B., M. S. Silverman and R. L. De Valois (1981). "Spatial frequency columns in primary visual cortex." Science 214(4522): 813-815.
Treisman, M. and T. C. Williams (1984). "A theory of criterion setting with an application to sequential dependencies." Psychological Review 91(1): 68.
Tribushinina, E. (2011). "Once again on norms and comparison classes." Linguistics 49(3): 525-553.
Urai, A. E., A. Braun and T. H. Donner (2017). "Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias." Nature communications 8: 14637.
Urai, A. E., J. W. De Gee, K. Tsetsos and T. H. Donner (2019). "Choice history biases subsequent evidence accumulation." Elife 8: e46331.
Urai, A. E. and T. H. Donner (2022). "Persistent activity in human parietal cortex mediates perceptual choice repetition bias." Nature communications 13(1): 1-15.
Wagemans, J., J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh and R. Von der Heydt (2012). "A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization." Psychological bulletin 138(6): 1172.
Walsh, V. (2003). "A theory of magnitude: common cortical metrics of time, space and quantity." Trends in cognitive sciences 7(11): 483-488.
Weber, A. I., K. Krishnamurthy and A. L. Fairhall (2019). "Coding principles in adaptation." Annu. Rev. Vis. Sci 5(1): 427-449.
Webster, M. A. and J. D. Mollon (1997). "Adaptation and the color statistics of natural images." Vision research 37(23): 3283-3298.
Wertheimer, M. (1912). "Experimentelle studien uber das sehen von bewegung." Zeitschrift fur psychologie 61: 161-165.
White, C. N., J. A. Mumford and R. A. Poldrack (2012). "Perceptual criteria in the human brain." Journal of Neuroscience 32(47): 16716-16724.

Willems, R. M., A. Özyürek and P. Hagoort (2009). "Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language." Neuroimage 47(4): 1992-2004.
Williams, R. and A. Jorgensen (2023). "Comparing logit & probit coefficients between nested models." Social Science Research 109: 102802.
Woo, C.-W., L. J. Chang, M. A. Lindquist and T. D. Wager (2017). "Building better biomarkers: brain models in translational neuroimaging." Nature neuroscience 20(3): 365-377.
Wutz, A., R. Loonis, J. E. Roy, J. A. Donoghue and E. K. Miller (2018). "Different levels of category abstraction by different dynamics in different prefrontal areas." Neuron 97(3): 716-726. e718.
Yamins, D. L., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert and J. J. DiCarlo (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the national academy of sciences 111(23): 8619-8624.
Zhang, H. and D. Alais (2020). "Individual difference in serial dependence results from opposite influences of perceptual choices and motor responses." Journal of Vision 20(8): 2-2.
Zhou, Y. and D. J. Freedman (2019). "Posterior parietal cortex plays a causal role in perceptual and categorical decisions." Science 365(6449): 180-185.
Zokaei, N., S. Burnett Heyes, N. Gorgoraptis, S. Budhdeo and M. Husain (2015). "Working memory recall precision is a more sensitive index than span." Journal of Neuropsychology 9(2): 319-329.

## 국문 초록

# 판단 경계의 업데이트가 선택과 선택 불확실성의 역사 효과에 미치는 영향

서울대학교 대학원
자연과학대학 뇌인지과학과
이 희 승

사람이 대상을 인식할 때 독특한 특징은 대상이 지닌 속성의 크기를 ‘크다/작다’ 와 같이 상대적인 범주로 단순하게 인식한다는 것이다. 언어학자와 인지과학자들은 사람이 이렇게 대상을 상대적인 범주로 인식할 때 어떻게 대상을 양쪽 두 범주 중 어떤 한 범주로 인식하는 것인지 그 작동 방식을 연구해 왔다. 그들은 사람은 판단의 기준이 되는 경계 값을 기억에 지니고 있으며 그 판단경계보다 대상이 큰지 작은지에 따라 대상의 범주가 정해지는 것이라고 분석하였다.

그러나 사람이 실제로 판단경계를 활용하여 대상의 범주를 판단한다는 설명은 현재까지 명확이 입증된 것이 아니다. 특히 뇌가 실제로 판단경계를 표상하는지가 보여진 적이 없다. 따라서 본 연구에서는 사람이 링의 크기를 ‘크다’ 혹은 ‘작다’로 판단할 때

사람의 뇌에서 판단기준이 실제로 표상되는지 알아보았다.

이 목적을 위해 이 연구에서는 판단기준의 특징으로 여겨지는 ‘기준 업데이트’라는 현상을 이용하였다. 기준 업데이트란 판단 기준이 최근에 본 대상의 크기와 비슷해진다는 가설을 말한다. 즉 큰 자극을 보면 판단 기준은 큰 자극과 비슷한 크기로 커진다는 것이다. 이 가설은 결국 다음에 내리는 대상에 대한 범주 판단이 이전에 본 대상의 크기와 반대 쪽으로 편향된다는 밀침편향 예측한다. 본 연구는 뇌에서 판단 기준이 표상된다면 그 판단 기준은 실제로 최근에 본 대상의 크기와 비슷해져야 한다고 전제하고 이 전제를 만족하는 뇌 자기공명영상 신호가 존재하는지 조사해 보았다. 그리고 실제로 왼쪽 두정엽과 측두엽에서 관찰된 판단기준의 신호가 이전에 본 자극과 비슷해 진다는 것을 확인 하였다.

아울러 기준업데이트가 선택 행위 영향을 준다는 것은 알려져 왔지만 판단불확실성에 주는 영향은 이제껏 연구되지 않았다. 따라서 본 연구는 기준업데이트가 판단불확실성에는 어떤 영향을 끼치는지도 조사하였다. 본 연구는 판단불확실성과 관련 깊다고 알려진 판단시간, 전방대상피질의 활동, 동공크기의 변화를 조사하였고 이 세 가지 모두 기준업데이트에 따라 체계적으로 영향을 받는다는 것을 밝혀 냈다. 이 발견은 기준업데이트가 선택 행위 뿐만 아니라 판단불확실성에도 영향을 주는 요소라는 것을 보여준다.

이 학위논문에서 이뤄진 발견은 사람의 뇌가 어떻게 대상을 상대적인 범주로 인식하는지 그 작동 방식에 대한 이해를 확장하였다. 그것은 사람의 뇌는 두정엽이란 부분에서 이전에 봤던 자극들을 이용해서 현재 자극이 나타나기 전에 판단 기준을 표상하며 그 표상된 판단 기준을 이용해 사람의 언어처리를 담당한다고 알려진 측두엽에서 대상에 대한 판단이 이뤄진다는 것이다.

**주요어 :** 기준업데이트, 역사효과, 밀침편향, 의사결정, 판단불확실성
**학 번 :** 2017-38129