Decision-consistent bias mediated by drift dynamics of human visual working memory

Authors: Hyunwoo Gu[1,2,3], Joonwon Lee[1], Sungje Kim[1], Jaeseob Lim[1], Hyang-Jung Lee[1], Heeseung Lee[1], Minjin Choe[1], Dong-Gyu Yoo[1], Jun Hwan (Joshua) Ryu[2,3], Sukbin Lim[4,5,6,*], and Sang-Hun Lee[1,*]

Affiliation:
[1] Department of Brain and Cognitive Sciences, Seoul National University, 1 Gwanak-ro, Seoul, 08826, Republic of Korea
[2] Department of Psychology, Stanford University, Stanford, CA 94305, United States
[3] Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305, United States
[4] Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, NYU Shanghai, 567 West Yangsi Road, Shanghai, 200126, People's Republic of China
[5] Neural Science, NYU Shanghai, 567 West Yangsi Road, Shanghai, 200126, People's Republic of China
[6] NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, 3663 Zhongshan Road North, Shanghai, 200062, People's Republic of China

Author contributions: J.Lee., S.K., J.Lim, H.-J.L., M.C., D.Y., and S.-H.L. designed research; H.G., S.L., and S.-H.L. performed research; H.G., H.-J.L., J.Lee., S.K., J.Lim, D.Y., and J.H.R. contributed unpublished reagents/analytic tools; H.G. wrote the first draft of the paper; H.G., S.L., and S.-H.L. edited and wrote the paper.

Acknowledgment: This research was supported by the Seoul National University Research Grant 339-20220013 and by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and Information and Communications Technology Grant No. NRF-2021R1F1A1052020. S. L. was supported by STI2030-Major Projects, No.2021ZD0203700/2021ZD0203705. S. L. also acknowledges the support of the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and the NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai.

The authors declare no competing financial interests.

* Correspondence should be addressed to Sukbin Lim at sukbin.lim@nyu.edu and Sang-Hun Lee at visionsl@snu.ac.kr.

Abstract

To adapt to dynamic surroundings, we need to reliably maintain sensory experiences while making accurate decisions about them. Nonetheless, humans tend to bias their ongoing actions toward their past decisions, a phenomenon dubbed decision-consistent bias. Efforts to explain this seemingly irrational bias have been limited to the sensory readout account. Here, by putting the bias in the context of mnemonic maintenance, we uncover its previously unidentified source: the interplay of decision-making with the drift dynamics of visual working memory. By taking behavioral snapshots of human visual working memory while concurrently tracking their cortical signals during a prolonged delay, we show that mnemonic representations transition toward a few stable points while initially biasing decisions and continuously drifting afterward in the direction consistent with the decisional bias. Task-optimized recurrent neural networks with drift dynamics reproduce the human data, offering a neural mechanism underlying the decision-consistent bias.

Introduction

Adapting to the world engages us in various perceptual tasks such as discrimination and estimation[1,2]. Sometimes, we need to carry out multiple tasks in succession, and in such a situation, the cognitive act for the preceding task may influence that for the subsequent task. A benchmark instance of this phenomenon is the decision-consistent bias, where the process underlying binary decision biases a later estimate of a stimulus toward the side consistent with the decision[2,3]. Understanding how this bias occurs can provide insights into how the brain flexibly uses and reuses its representations of the world's properties under different task demands. While extensive research has elucidated the process of reaching a decision based on sensory evidence[4–8], it remains unclear how that decision-related process influences the post-decisional process of maintaining the evidence for future use.

The decision-consistent bias had once been construed as a perceptual illusion[2,9], but it is reconceptualized to involve post-perceptual processes by recent empirical and computational studies[3,10,11]. Across these studies[2,3,9–11], as the sensory drive of stimulus is no longer available when performing the subsequent estimation tasks, the mnemonic representation of the stimulus should be formed and maintained in working memory (WM). Thus, WM is expected to play a role in mediating the decision-consistent bias. However, efforts to put the decision-consistent bias in the context of WM have been surprisingly scarce, particularly given WM's dynamic nature[13] and intimate relationship with decision-making (DM)[12–14].

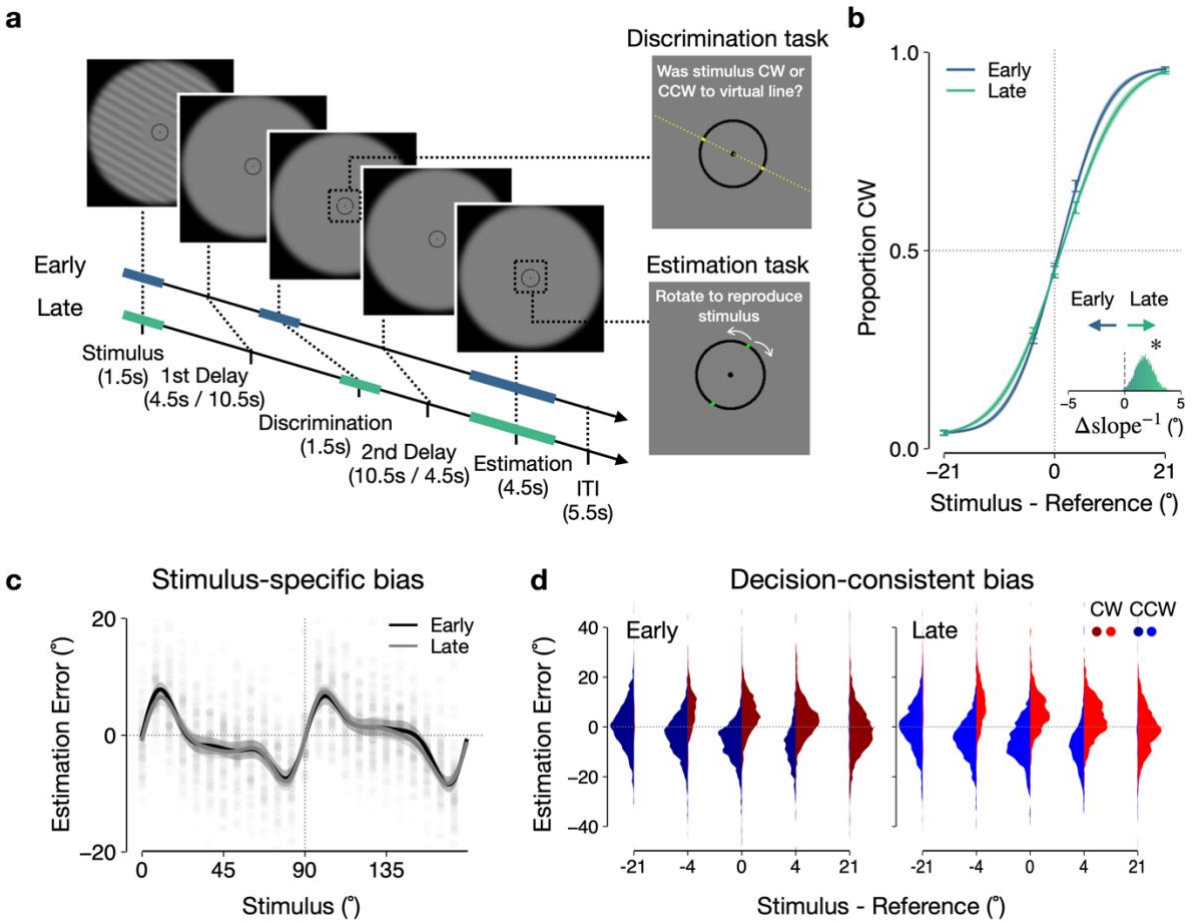


Fig. 1 | Stimulus-specific and decision-consistent biases in discrimination and estimation tasks with intervening delays. a, Task paradigm. Participants inside an fMRI scanner remembered the stimulus orientation. After the 1st delay, they were asked to make a discrimination decision about whether the stimulus orientation was CW or CCW to the reference (the virtual line connecting two yellow dots). Following another delay (2nd delay), they were asked to adjust the orientation of the response frame to match the stimulus orientation in memory. The DM timing ('early' versus 'late') was randomized on a trial-to-trial basis. **b**, Discrimination psychometric curves. The proportions of CW discrimination pooled across participants are plotted separately as a function of stimulus-reference difference for the early and late DM conditions. The curve was steeper in the early DM condition than in the late DM condition (inset; bootstrap test, $p = 0.030$, 95% CI = $[0.220°, 3.499°]$). Vertical error bars denote ± s.e.m.s **c**, Stimulus-specific bias in the early (black) and late (gray) DM conditions. Each point represents the mean of stimulus-conditioned errors for each participant. Curves are the smoothed lines of stimulus-conditioned errors pooled across participants. Shades denote ± s.e.m.s across participants. **d**, Decision-consistent bias in the early (left) and late (right) DM conditions. The densities of the DM-conditioned estimation errors pooled across participants are shown separately for the stimulus-reference differences. $*$, $p < 0.05$.

Here, to investigate how the decision-consistent bias unfolds over WM dynamics, we developed a paradigm where humans sequentially perform discrimination and estimation tasks with varying intervening delays based on a remembered orientation stimulus (**Fig. 1a**). We reasoned that the decision-consistent bias would undergo different time courses depending on the characteristics of the underlying WM dynamics. If WM only diffuses[17-22], the mnemonic representations become increasingly noisier but remain unbiased, whereas if WM not only diffuses but also drifts toward a few stable points over a stimulus space[15–17], the mnemonic representations are also biased. In the latter case, the mnemonic biases in the stimulus space can influence both pre-decisional and post-decisional processes, as the WM and DM are yoked to the target stimulus in our paradigm.

With the paradigm and rationale outlined above, we analyzed the behavioral and brain (fMRI) responses of fifty participants to determine what type of dynamics govern human visual WM and to investigate how the decision-consistent bias unfolds under that dynamics. The stimulus-specific bias increased over time, both in the DM behavior and in the WM representations decoded from the early visual cortex, indicating that the slow drift dynamics govern human visual WM of orientation. When decisions were made earlier than later, both behavioral and brain responses showed greater post-decision biases, the quality of the decision-consistent bias expected in the context of the drift WM dynamics. To gain insights into potential circuit mechanisms conferring that dynamic quality on the decision-consistent bias, we trained recurrent neural networks (RNNs) with the same paradigm performed by human participants. When enforced to display drift dynamics by varying the variability of inputs, the RNNs could reproduce the decision-timing-dependent decision-consistent bias observed in human data. Our findings highlight the importance of considering the dynamic nature of WM when accounting for the mutual influences between WM and DM.

Results

Stimulus-specific and decision-consistent biases in WM of orientation

On each trial, participants had to memorize the orientation of a briefly (1.5 s) viewed grating and report it after a prolonged (16.5 s) delay. To probe the decision-consistent bias, we also asked participants to perform a discrimination task (**Fig. 1a,** right top) before the estimation task (**Fig. 1a,** right bottom). In the discrimination task, upon the appearance of a reference, participants had to decide, with modest time pressure (1.5 s), whether the remembered orientation was tilted clockwise or counter-clockwise relative to the reference, the orientation of which was randomly determined such that it deviates from the target orientation by $-21°$, $-4°$, $0°$, $4°$, or $21°$, where the negative sign indicates the counter-clockwise offset.

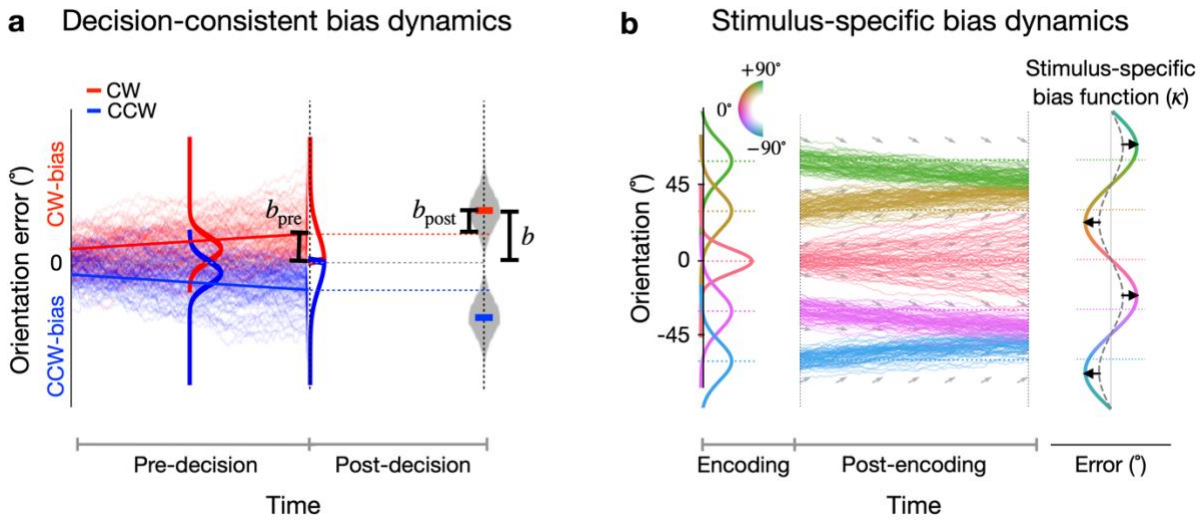


Fig. 2 | Impact of WM dynamics on decision-consistent and stimulus-specific biases. a, Decision-consistent bias under diffusion dynamics. Each trajectory represents a single-trial temporal evolution of memory states in the orientation space, subject to white Gaussian noise. Individual trajectories are color-labeled by the discrimination choice, which is determined by whether the memory state at the time of DM lies in the CW or CCW side of the reference. Across-trial probability densities conditioned on the choices are shown as solid curves. Orientation estimates are determined by the memory states at the end of the delay period. The decision-consistent bias (the mean of the estimation errors conditioned on each choice, b) can be decomposed into the component built up to the choice (pre-decision bias, b_{pre}) and the one occurring after the choice (post-decision bias, b_{post}). **b**, Stimulus-specific bias under drift dynamics. Each trajectory represents a single-trial temporal evolution of memory states during the period posterior to stimulus encoding, subject to drift dynamics (arrows in the background), being color-labeled with five different stimulus orientations. The efficient encoding scheme confers the initial stimulus-specific bias on memory states during the encoding period ('encoding bias,' depicted by the dotted curve on the right). Then, the drift dynamics of WM further potentiate the bias ('memory bias,' represented by the arrows on the right). These two sources of bias both contribute to the stimulus-specific bias in estimation errors (κ, depicted by the solid curve on the right).

Critically, to examine the dynamic aspects of the stimulus-specific and decision-consistent biases, we varied the timing of the discrimination task. The discrimination task onset was 4.5 s and 10.5 s following the stimulus offset, respectively, in the early DM and late DM conditions (**Fig. 1a,** bottom). As anticipated by the difference in pre-decision delay (4.5 s versus 10.5 s), the discrimination performance was better in the early DM condition, as indicated by a sharper psychometric curve (**Fig. 1b**).

The stimulus-specific and decision-consistent biases were evident in the errors in the estimation task, regardless of when the decision was made. When the estimation errors were plotted as a function of stimulus orientation, they were repulsed from the cardinal orientations

while being attracted to the oblique orientations (**Fig. 1c**), a commonly observed pattern in orientation estimation tasks[18–21]. When conditioned on the discrimination direction, the estimation errors were biased toward the discrimination direction (**Fig. 1d**).

Notably, as shown in the previous work[2,3], the decision-consistent bias was pronounced in the trials where the difference between the reference and stimulus orientations was small ($-4°$, $0°$, $4°$ on the x-axes in **Fig. 1d**). Furthermore, the marginal (unconditioned) distribution of estimation errors was broader in the near-reference ($-4°$, $0°$, $4°$) trials than the far-reference ($-21°$, $21°$) trials, which has been taken as evidence for the decision-*induced* bias in orientation estimation[3]. Consistently, we confirmed this pronounced error variability in the near-reference trials in the various datasets where the decision-consistent bias was reported (**Supplementary Fig. 1**).

Relating the dynamics of WM to the time courses of the biases

Central to the current work is the rationale that the stimulus-specific and decision-consistent biases undergo distinct time courses depending on whether WM has only diffusion or both diffusion and drift dynamics. To derive from this rationale concrete predictions in the current experimental paradigm, we first built a model class that can reproduce the observed stimulus-specific and decision-consistent biases at a phenomenal level with minimal assumptions about the sensory encoding process and the impact of DM on WM (**Figs. 2**,**3**).

To capture the stimulus-specific bias (**Fig. 1c**), we assumed that the distributions of orientation measurements at the sensory-encoding stage are constrained by the efficient encoding scheme[28,31] (**Fig. 2b**, left). To capture the pronounced decision-consistent bias for the near-reference trials (**Fig. 1d; Supplementary Fig. 1**), we also assumed that the act of committing to a decision in the discrimination task induces a pulse-like shift in mnemonic representation to the direction consistent with the decision ('decision-induced bias' in **Fig. 3a,b**). To be sure, these assumptions were introduced as possible ways of reproducing the observed patterns of estimation errors. Later, we will address their plausibility when describing the results of RNN simulations.

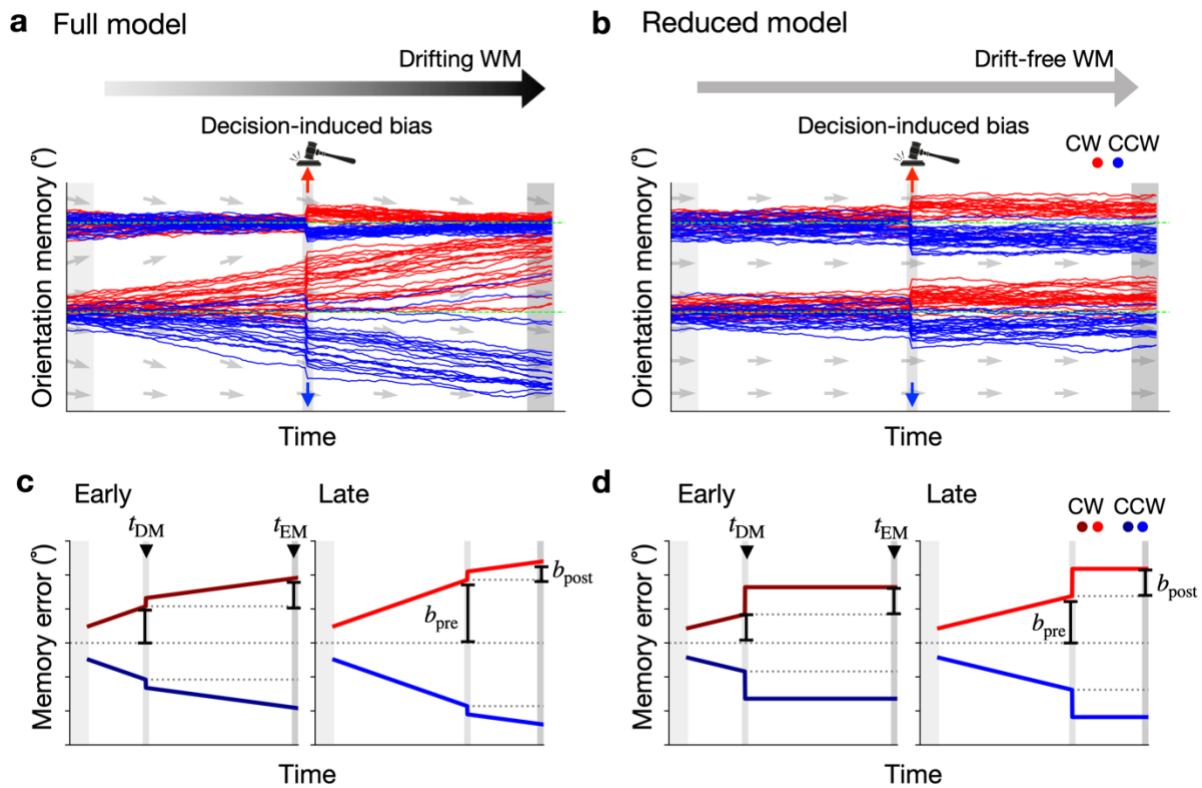


Fig. 3 | Decision-timing-dependent impact of WM drift on the decision-consistent bias. **a**,**b**, Schematic illustration of the WM model with diffusion and drift (full model, **a**) and the one only with diffusion (reduced model, **b**). Single-trial trajectories of memory states, each being color-labeled with the discrimination choice during the DM period (the gray shade in the middle), are shown for two stimulus orientations. Right after the stimulus offset (the gray shade on the left), the memory states in both models share the same encoding bias associated with the efficient encoding scheme. Afterward, they undergo different dynamics, diffusion-and-drift dynamics in the full model (**a**) and diffusion-only dynamics in the reduced model (**b**). In both models, the decision induces a pulse-like shift in memory states toward the direction consistent with itself (as indicated by the gavel and color-labeled vertical arrows in the middle). The color intensity in the large arrows at the top corresponds to the degree of decision-consistent bias. The small gray arrows in the background represent the direction of the drift. **c**,**d**, Evolution of decision-consistent biases in the early and late DM conditions predicted by the full (**c**) and reduced (**d**) models. The pre-decision bias (b_{pre}) increases as the DM time (t_{DM}) is delayed in both models, as indicated by the long b_{pre} bars in the late DM condition. By contrast, as the DM time (t_{DM}) is delayed, the post-decision bias (b_{post}) decreases in the full model, as indicated by the short b_{post} bar in the late DM condition, but remains unchanged in the reduced model, as indicated by the same-length b_{post} bars in the early and late DM conditions.

Next, we incorporated the two possible scenarios of WM dynamics into the above platform to predict what distinct time courses of the stimulus-specific and decision-consistent

biases would occur under the corresponding scenarios. As for the stimulus-specific bias, it is predicted to remain unchanged in the diffusion-only scenario (**Fig. 3b**) but to grow in the diffusion-and-drift scenario (**Fig. 3a**) because the across-trial averages of mnemonic orientation representations steadily approach a few stable points over orientation space in the latter scenario (**Fig. 2b**). As for the decision-consistent bias, the influences of WM dynamics on the decision-consistent bias differ between its pre-decision (b_{pre}) and post-decision (b_{post}) components. b_{pre} is predicted to grow as the decision is delayed under both scenarios (**Fig. 3c,d**) because the diffusion dynamic alone leads to a greater separation between the decision-conditioned distributions of mnemonic orientation representations (**Fig. 2a**). By contrast, b_{post} is predicted to be unaffected by when the decision is made in the diffusion-only scenario (**Fig. 3d**) but to decrease as the decision is delayed in the diffusion-and-drift scenario (**Fig. 3c**) because the drifts toward a few stable points over orientation space not only cause the stimulus-specific bias in DM but also further the bias in the same directions.

In what follows, with the above distinct predictions under the two scenarios of WM dynamics, we determined which scenario better accounts for the time courses of the stimulus-specific and decision-consistent biases in participants' behavior and brain activity.

Growth of the biases in behavior

We began by comparing the decision-consistent bias in the near-reference trials between the early and late DM conditions. It was greater in the late DM condition than in the early DM condition (**Fig. 4a,b**; paired samples t-test, $t(49) = 2.929$, $p = 0.005$, 95% confidence interval (CI) =$[0.204°, 1.093°]$). These results are consistent with the diffusion-only scenario because the gradual noise growth makes the two choice-conditioned error distributions increasingly distant from each other as the 'choice-conditioning' time is delayed. While the diffusion-and-drift scenario does not necessarily predict a specific outcome, it is compatible with the observed results as long as the drift rate does not unrealistically increase over time.

Next, we turn to the model predictions differing between the two scenarios, those regarding the stimulus-specific bias. To examine whether the stimulus-specific bias grows over time, we separately estimated its magnitudes at the time of DM for the early and late DM conditions. The magnitudes of the stimulus-specific bias were estimated in the following procedure. First, for each stimulus orientation for each participant, we derived the two psychometric curves, respectively, from the early and late discrimination performances. Next, we fitted the weighted von Mises basis function $w \cdot \kappa(\theta)$ to the points of subjective equality of those psychometric curves across the stimulus orientations θs (**Supplementary Fig. 2**; see **Methods** for details). Lastly, we quantified the magnitudes of the stimulus-specific bias with the best-fit values of the bias weight parameter w. The stimulus-specific bias was greater at the time of late (10.5 s after stimulus offset) discrimination than at the time of early (4.5 s after stimulus offset) discrimination (**Fig. 4c,d**; paired samples t-test, $t(49) = 3.244$, $p = 0.002$, 95% CI =$[0.070, 0.297]$).

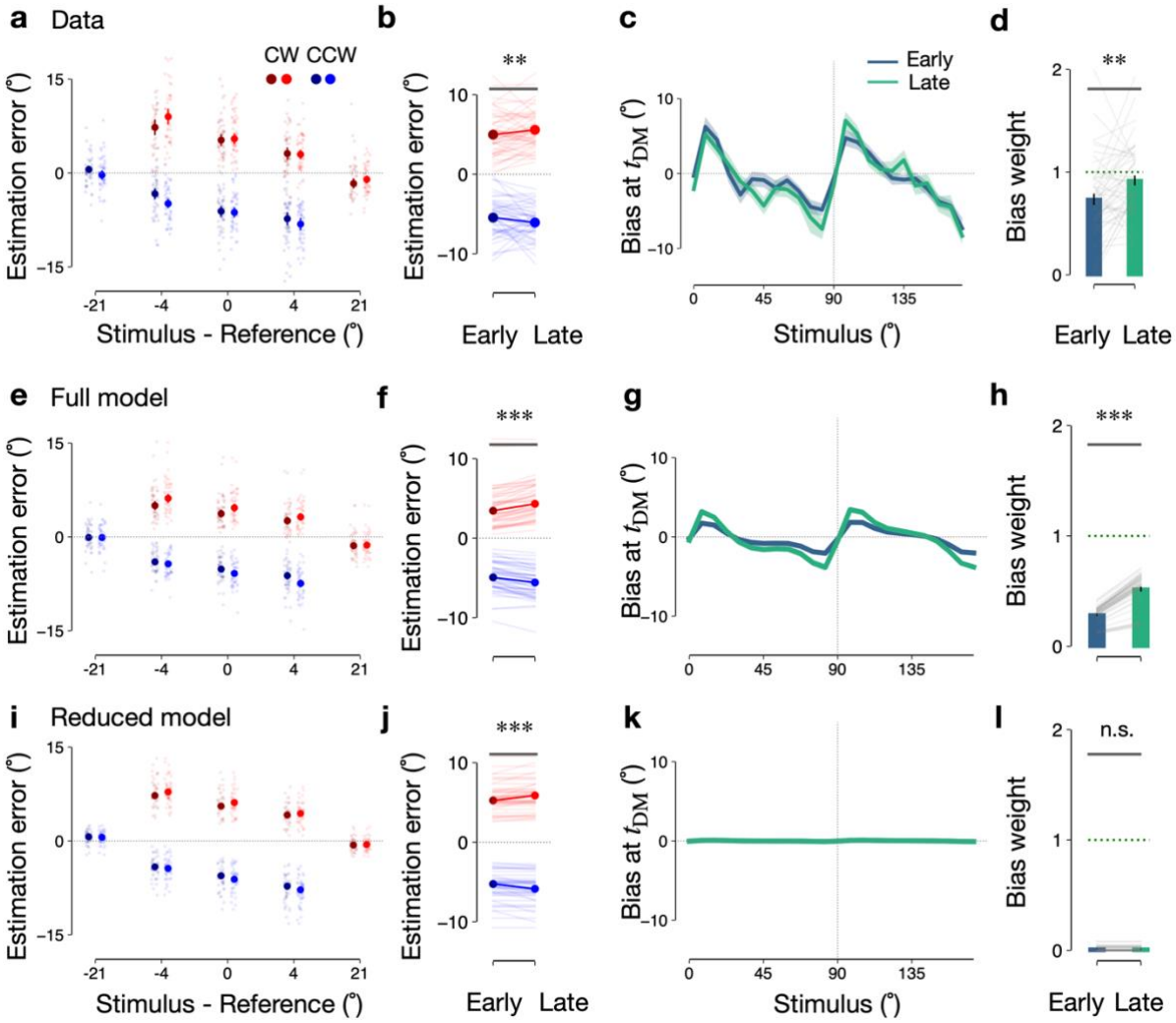


Fig. 4 | Growth of the biases in behavior. (Continued on the following page.)

Fig. 4 | Growth of the biases in behavior. a,**b**, Decision-consistent bias and its growth in human estimation behavior. **a**, Decision-conditioned estimation errors in the early (dark-colored dots) and late (light-colored dots) DM conditions. Small transparent dots represent the data points from individual participants. Solid dots with error bars denote the across-participant means $\pm$ s.e.m.s. The data for CW and CCW choices were not shown for the $21°$ and $-21°$ reference conditions, respectively, because they were too noisy due to the small number of trials (**Fig. 1d**). **b**, Comparison in the decision-consistent bias for the near-reference trials between the early and late DM conditions. Dots denote the across-participant medians, while lines represent individual participants' data. The decision-consistent bias was greater in the late DM condition than in the early DM condition (paired samples t-test, $t(49) = 2.929$, $p = 0.005$, 95% CI =$[0.204°, 1.093°]$). **c**,**d**, Stimulus-specific bias and its growth in human decision behavior. **c**, Changes of the points of subjective equality in participants' decisions as a function of stimulus orientation, shown separately for the early and late DM conditions. Colored lines with shades denote the means ± s.e.m.s based on bootstrapping. **d**, Comparison in the stimulus-specific bias in DM between the early and late DM conditions. Bars denote the across-participant means of the relative amplitudes of stimulus-specific bias function $\kappa(\theta)$ in DM, while lines denote individual participants' data. Error bars represent ± s.e.m.s. The stimulus-specific bias was greater in the late DM condition than in the early DM condition (paired samples t-test, $t(49) = 3.244$, $p = 0.002$, 95% CI =$[0.070, 0.297]$). **e-h**, Decision-consistent (**e,f**) and stimulus-specific (**g,h**) biases of the full model. Same as **a-d**, but for the *ex-post* model simulations based on the best-fitting parameters for individual participants. Paired samples t-tests, $t(49) = 11.601$, $p = 1.161 \times 10^{-15}$, 95% CI=$[1.371°, 1.946°]$ (**f**), $t(49) = 22.501$, $p = 1.748 \times 10^{-27}$, 95% CI=$[0.212, 0.254]$ (**h**). **i-l**, Same as **e-h**, but for the reduced model. Paired samples t-tests, $t(49) = 4.232$, $p = 1.013 \times 10^{-4}$, 95% CI=$[0.299°, 0.840°]$ (**j**), $t(49) = 1.083$, $p = 0.284$, 95% CI=$[-6.137 \times 10^{-18}, 2.050 \times 10^{-17}]$ (**l**). $***$, $p < 0.001$, $**$, $p < 0.01$, n.s., $p > 0.05$.

The results above lend qualitative support for drift dynamics in WM of orientation. Seeking quantitative support for drift dynamics, we fitted the models with or without drift to the behavioral data (**Methods**). The model with drift was superior to the one without drift for all participants when the goodness of fit was compared with the Bayesian Information Criterion (**Supplementary Fig. 3a**). Notably, the fitted parameters of drift rate suggest that WM of orientation drifts in a quite slow regime (less than $1°/\text{sec}$; w_K in **Supplementary Fig. 3c**). Moreover, in the *ex-post* model simulation, the prediction of the model based on its best-fit parameters, the model with drift reproduced not only the observed growth and shape of the stimulus-specific bias (**Fig. 4g,h**) but also those of the decision-consistent bias **(Fig. 4e,f**; **Supplementary Fig. 4a,b**). By contrast, the model without drift could reproduce only the observed patterns of the decision-consistent bias (**Fig. 4i,j**) but not those of the stimulus-specific bias (**Fig. 4k,l; Supplementary Fig. 4c**). Furthermore, the amplitude of a 'decision-induced bias' was significantly positive only in the full model, which was required to capture the near-reference variability (**Supplementary Figs. 3**,**5**).

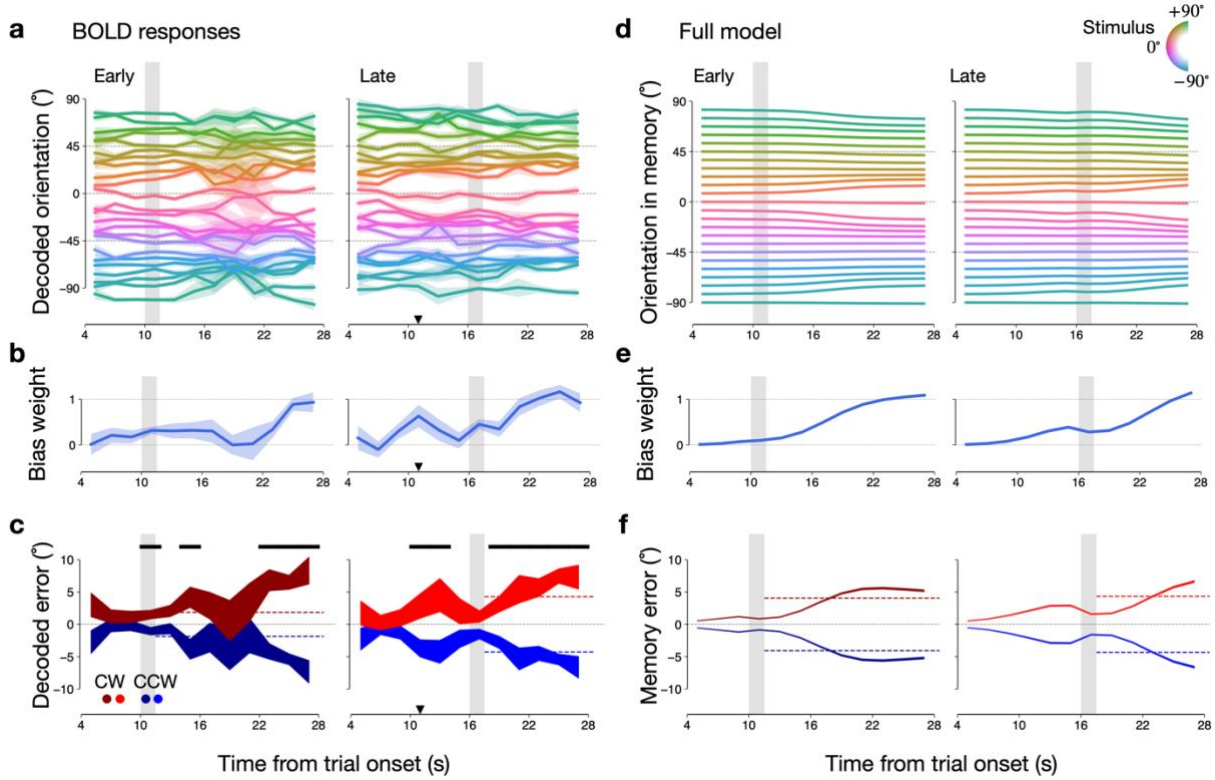


Fig. 5 | Growth of the biases in the cortical signal of WM of orientation. a,**b**, Evolution of the stimulus-specific bias in the WM signals of orientation in the human early visual cortex (V1, V2, and V3). The DM period is demarcated by gray vertical bars, taking into account a hemodynamic delay. **a**, Temporal trajectories of the brain signals of orientation, pooled across participants. Colors represent stimulus orientations. Shades are ± s.e.m.s based on bootstrapping. Inverted triangle, the pre-decision time point (t=11s from trial onset) in the late DM condition. **b**, Growth of the bias weight (κ, the amplitude of stimulus-specific bias). The bias weight of value 1 is equivalent to the amplitude of the behavioral bias. The inverted triangle demarcates the pre-decision time point at which the bias weight significantly deviated from zero (One-sample t-test, $t(49) = 2.543$, $p = 0.014$, 95% CI=$[0.133, 1.131]$). Shades denote ± s.e.m.s across participants. **c**, Growth of decision-consistent bias, pooled across participants. Black horizontal bars at the top demarcate the time points at which the two decision-conditioned means of the brain signal of orientation significantly differed from each other ($p < 0.05$, bootstrap test, Bonferroni corrected for the number of time points). Shades are ± s.e.m.s based on bootstrapping. The inverted triangle demarcates the pre-decision time point at which the decision-consistent bias was significant. Colored dashed lines mark the degree of the decision-consistent bias at the onset of the post-decision epoch (see **Methods**). **d-f**, Growth of biases predicted by the full model. Same as **a-c**, but for the *ex-post* model simulations based on the best-fitting parameters for individual participants and the hemodynamic convolution on the polar space (see **Methods**).

Growth of the biases in cortical signals of mnemonic orientation

The behavioral analysis probed the "snapshots" of the biases at the moments of discrimination

and estimation. To expand our examination of the biases outside those snapshot moments, we decoded the mnemonic signal of stimulus orientation from the BOLD data acquired simultaneously with behavior and tracked the time courses of the biases in that decoded signal. We targeted the early visual cortex, V1, V2, and V3, for this signal, considering the availability of high-fidelity WM of orientation there[22,23].

The time courses of the brain signal of WM were consistent with the diffusion-and-drift scenario in the following aspects. First, the temporal trajectories of the brain signals conditioned on stimulus orientation drifted away from the cardinal orientations and towards the oblique orientations in both early and late DM conditions (**Fig. 5a**). Second, to track the gradual growth of the stimulus-specific bias, we quantified the strength of the bias at each time point by finding the best-fit bias weight parameter with the von Mises basis function defined for each participant in the behavioral data, $\hat{\kappa}(\theta)$. The bias weight was initially low and increased to a level close to that quantified in the behavioral estimation errors (**Fig. 5b**). Third, to track the gradual growth of the decision-consistent bias, we estimated the means of the brain signals conditioned on discrimination direction at each time point for the near-reference trials. The means of these DM-conditioned distributions of the brain signal gradually bifurcated away from target orientation towards the direction consistent with DM (**Fig. 5c**).

There are two caveats in inferring the WM dynamics from the BOLD signals. The BOLD signals may appear to change more gradually than the actual neural activities due to the hemodynamic delay in neurovascular coupling. Another issue is that the limited spatial resolution of BOLD signals may cause interference between the stimulus and reference orientations during the DM epoch. Indeed, the brain signals seemed transiently attracted towards the near references, especially in the late DM condition (**Fig. 5a**, right), leading to a transient drop in stimulus-specific bias (**Fig. 5b**, right), and in decision-consistent bias (**Fig. 5c**, right) as well, during the DM epoch. Notably, despite such transient interferences, both stimulus-specific and decision-consistent biases were greater in the post-decision epoch than in the pre-decision epoch (as indicated by the inverted triangle and dashed lines in **Fig. 5b-c**).

For a more straightforward comparison between the predicted (by the diffusion-and-drift model) and observed (in BOLD) trajectories of WM, we addressed the second caveat by incorporating the attraction of the mnemonic representation of target orientation towards the reference input into the model prediction—as done similarly in the well-established event-related analysis[24] and the first caveat by convoluting the predicted trajectories of WM with the canonical hemodynamic impulse response function. We stress that the model predictions were made with the parameters best fit to the behavioral data except for those regarding the attraction towards the reference (see **Methods** and **Supplementary Fig. 6** for further details). As a result, the WM trajectories predicted by the diffusion-and-drift model (**Fig. 5d-f**) closely resembled those observed in the BOLD signals of WM (**Fig. 5a-c**). For the majority of participants, the resemblance to the observed trajectories of the BOLD signals was better for the diffusion-and-drift model than for the diffusion-only model (**Supplementary Fig. 6**). The diffusion-only model was quite limited, especially in reproducing the continual growth of the stimulus-specific bias found in the BOLD signals (**Supplementary Fig. 7**).

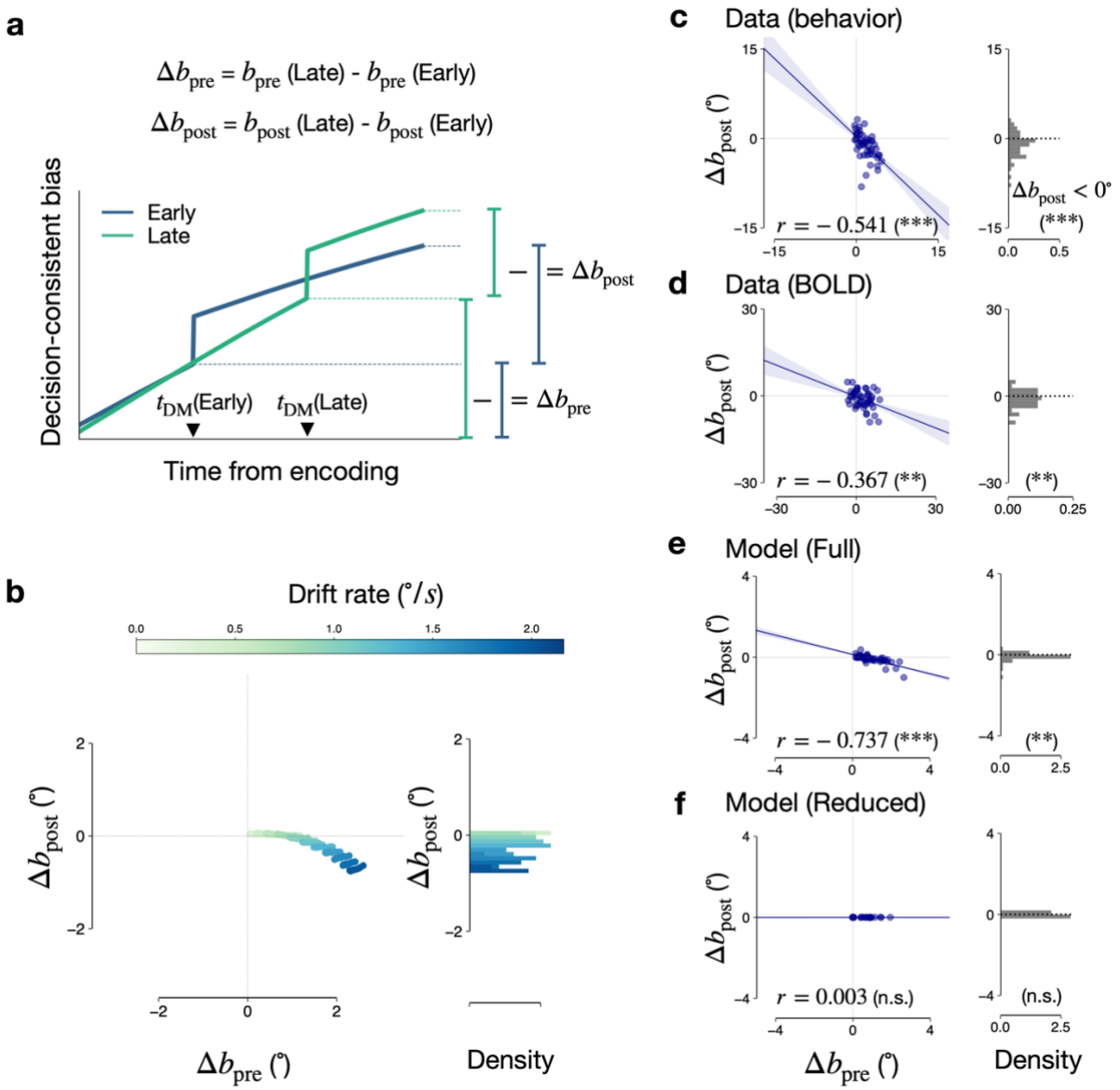


Fig. 6 | Decision-timing-dependent post-decision bias. (Continued on the following page.)

Fig. 6 | Decision-timing-dependent post-decision bias. a, Predicted time courses of the decision-consistent bias under drift dynamics. The differences between the early and late DM conditions in the pre-decision and post-decision components of the decision-consistent bias (Δb_{pre} and Δb_{post}) are defined in equations at the top and graphically illustrated on the right. **b**, Results of the *ex-ante* model simulation of Δb_{pre} and Δb_{post} with different drift and diffusion rates. Δb_{post} is plotted against Δb_{pre} in the left panel, where dot colors represent the drift rates used in given simulations (as indicated by the color bar at the top). The Δb_{post} histogram is shown in the right panel, with the same color scheme. **c-f**, Across-individual correlations between Δb_{pre} and Δb_{post} (left) and Δb_{post} histogram (right) for the behavior data (**c**), the brain signals of WM orientation (**d**), the *ex-post* simulation data with the full (**e**) and reduced (**f**) models. Same as **b**, except that dots represent individual participants; lines in the left panels are based on a linear regression model, and the colored shades represent ± s.e.m.s of the regression lines. r, Pearson correlation coefficient. Left panels, $p = 5.066 \times 10^{-5}$, 95% CI=$[-0.712, -0.309]$ (**c**), $p = 8.723 \times 10^{-3}$, 95% CI=$[-0.586, -0.099]$ (**d**), $p = 1.015 \times 10^{-9}$, 95% CI=$[-0.843, -0.578]$ (**e**), $p = 0.986$, 95% CI=$[-0.276, 0.281]$ (**f**). Right panels, one-sided t-tests under the alternative hypothesis $\Delta b_{\text{post}} < 0°$, $t(49) = -3.612$, $p = 3.576 \times 10^{-4}$ (**c**), $t(49) = -2.485$, $p = 8.203 \times 10^{-3}$ (**d**), $t(49) = -3.181$, $p = 1.272 \times 10^{-3}$ (**e**), $t(49) = -0.592$, $p = 0.278$ (**f**). $***$, $p < 0.001$, $**$, $p < 0.01$, n.s., $p > 0.05$.

Post-decision bias depends on decision timing

As previously explained, the post-decision component of the decision-consistent bias (b_{post}) decreases only in the drift dynamics as the decision is delayed, whereas the pre-decision component of the bias (b_{pre}) increases in both scenarios (**Fig. 6a**). We quantified these delay-dependent changes by subtracting the values of b_{pre} and b_{post} in the early DM condition from those in the late DM condition, respectively, indicated by Δb_{pre} and Δb_{post}. To preview how Δb_{pre} and Δb_{post} are distributed under the drift dynamics, we simulated the discrimination and estimation behaviors using the diffusion-and-drift model while varying the drift and diffusion rate parameters within the range identified by the model fitting (w_K and w_D in **Supplementary Fig. 3c**). The simulation confirmed that Δb_{pre} and Δb_{post} fall in the positive (b_{pre} in the late DM condition > b_{pre} in the early DM condition) and negative (b_{post} in the late DM condition < b_{post} in the early DM condition) regimes, respectively. Further, the absolute amounts of Δb_{pre} and Δb_{post} increase as the drift rate increases, resulting in a negative correlation between them (**Fig. 6b**).

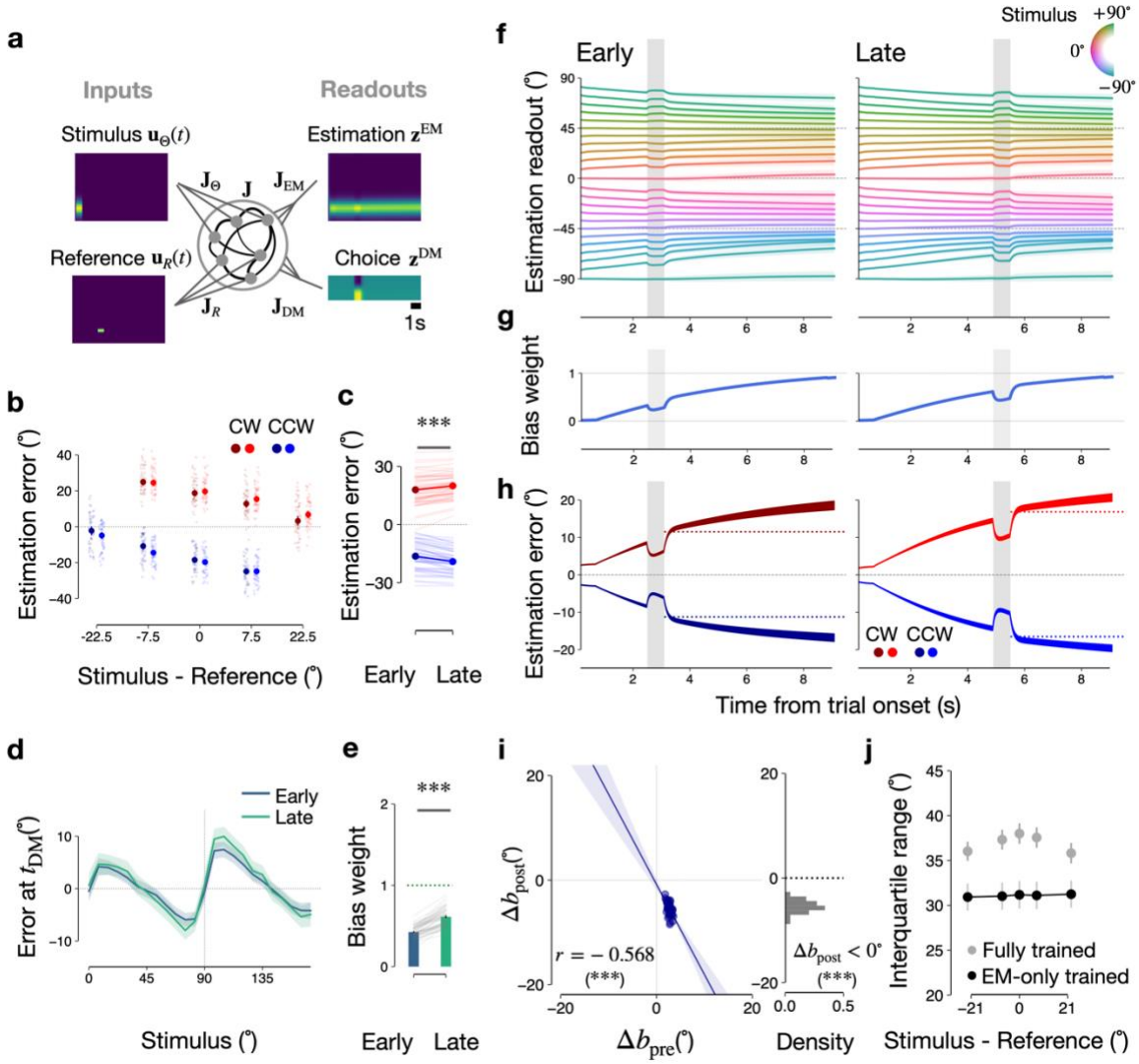


Fig. 7 | Reproduction of human data by task-optimized RNNs. a, Architecture of RNNs. RNN receives time-varying inputs of stimulus $\Theta(t)$ and reference $R(t)$ and generates two time-varying outputs, orientation estimation output z^{EM} through readout projection J_{EM} and binary decision output $\boldsymbol{z}^{\mathrm{DM}}$ through readout projection $\boldsymbol{J}_{\mathrm{DM}}$. **b-e**, Decision-consistent bias (**b**), stimulus-specific bias (**d**), and their growth (**c**,**e**) in the trained RNNs. Same as **Fig. 4a-d**, but for RNNs' behavior. Paired samples t-tests, $t(49) = 17.692$, $p = 6.407 \times 10^{-23}$, 95% CI=$[4.250°, 5.339°]$ (**c**), $t(49) = 16.019$, $p = 4.108 \times 10^{-21}$, 95% CI=$[0.164, 0.211]$ (**e**). **f-h**, Growth of the biases in RNNs' orientation readouts. Same as **Fig. 5a-c**, but for RNNs' orientation readouts. **i**, Across-RNN correlations between Δb_{pre} and Δb_{post} (left) and Δb_{post} histogram (right). Same as Fig. 6**c**, but with RNN's behavior. Left panel, $p = 1.695 \times 10^{-5}$, 95% CI=$[-0.731, -0.344]$. Right panel, $t(49) = -30.922$, $p = 4.161 \times 10^{-34}$. **j**, Interquartile ranges across relative reference conditions in the marginal distribution of the estimation errors, shown separately for the original RNNs (gray dots) and the RNNs trained with the loss function penalizing only for estimation errors (black dots). Bars are ± s.e.m.s across individual RNNs. $***$, $p < 0.001$.

The behavior (**Fig. 6c**) and BOLD (**Fig. 6d**) results matched the simulation results, both in the signs of Δb_{pre} and Δb_{post} and in the correlation between the two. We stress that Δb_{pre} and Δb_{post} were quantified in a model-free way without relying on any models proposed elsewhere in this paper (**Methods**). The *ex-post* model simulation using the best-fitting parameters confirmed that the model with drift (**Fig. 6e**), but not the one without drift (**Fig. 6f**), reproduced the observed results. The relationship between the pre-decision and post-decision components of the decision-consistent bias and how they are affected by decision timing and drift provides further support for the drift dynamics in WM of orientation.

RNN with drift dynamics reproduces the human data

So far, we have shown the significance of drift dynamics in capturing the temporal dependence of the biases by analyzing the data and WM dynamic models. We utilized phenomenological drift-diffusion models with several simplifying assumptions of WM and DM interactions, for instance, pulse-like change in the activity states with DM. Next, we used RNN[5,25,26] training to infer potential circuit mechanisms further.

Using a task paradigm equivalent to what human participants performed, we trained 50 independent RNNs with a joint loss function that penalizes them for both decision and estimation errors (**Fig. 7a**). Considering the spatial separation of the stimulus and reference in the human task, we fed stimulus and reference inputs to two separate populations of RNNs. We generated the RNN stimulus inputs by sampling from an orientation distribution that follows the previously suggested parametric form of variability[27] to model the initial encoding distribution that considers the orientation-specific variability found in natural scenes[20,27]. This enforced the stimulus-specific drift dynamics of RNNs towards the orientations with higher input variability, thereby reducing the estimate errors.

The trained RNNs exhibited several features characteristic of the human data. Specifically, they showed the decision-consistent and stimulus-specific biases, which grew as the decision was delayed (**Fig. 7b-e**). The growths were gradual, just like the BOLD data (**Fig. 7f-h**). Notably, RNNs also exhibited the effects of sensory drive associated with the reference presentation during the discrimination-task epoch. This is because all neurons of RNNs, including those that receive the reference, contribute to the estimation readout (**Fig. 7a**). Consistent with both human and diffusion-and-drift model data, RNNs showed the negative values for Δb_{post}, which were negatively correlated with Δb_{pre} (**Fig. 7i**). Lastly, RNNs showed broad distributions of estimation errors when the reference was close to the target stimulus, a behavioral feature diagnostic of the decision-*induced* bias (gray symbols in **Fig. 7j**). Importantly, such increases of estimation errors in the near-reference conditions were not observed when RNNs were trained with a loss function that penalizes them for only estimation errors (black symbols in **Fig. 7j**). This indicates that DM plays a crucial role in inducing the near-reference variability. Put together, these observations indicate that training RNNs to minimize both discrimination and estimation errors while conferring drift dynamics on them is sufficient for

RNNs to display the key features of the decision-consistent and stimulus-specific biases observed in the human data.

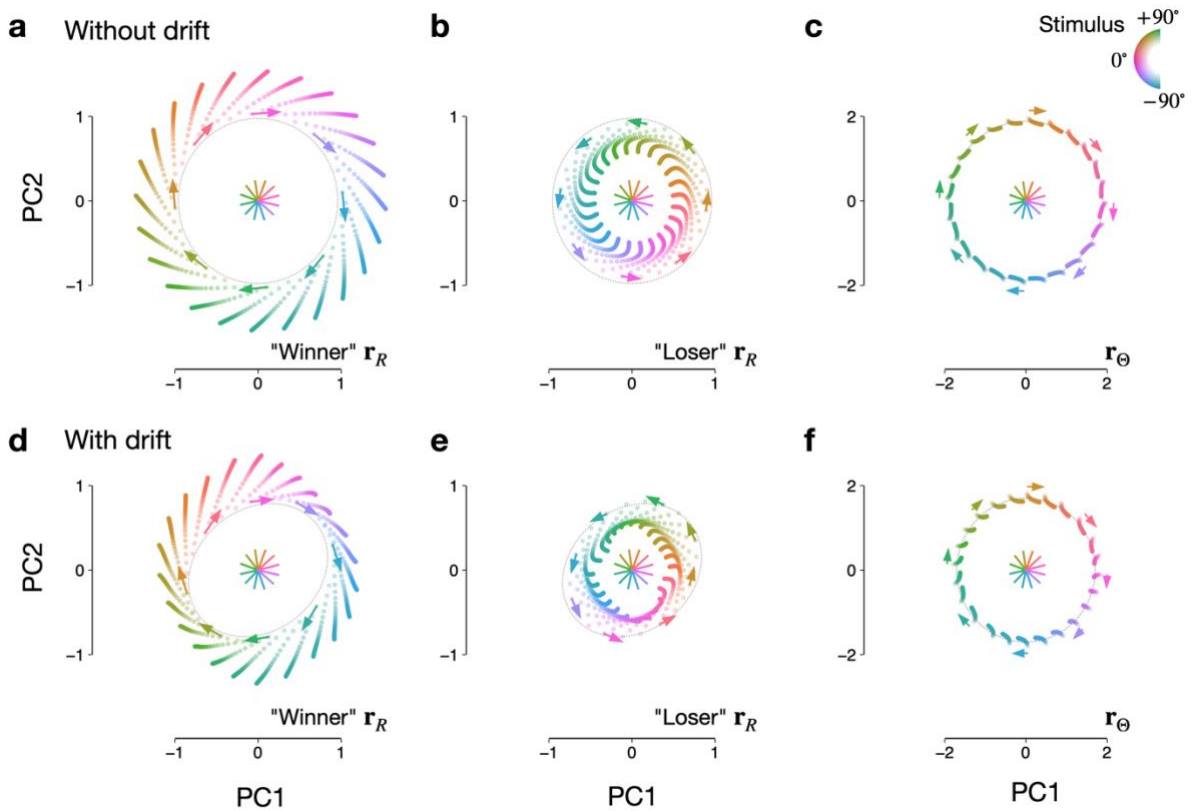


Fig. 8 | State space analysis on RNNs. a-c, Projection of the population dynamics of homogeneous (without drift) RNNs during the DM epoch on the common principal component space spanned by the activities of $\boldsymbol{r}_{\boldsymbol{\Theta}}$ during the entire task phases. Color hue indicates stimulus orientation, while color saturation denotes the passage of time. Arrows capture the direction of state changes for different stimulus orientations. Dotted circles correspond to the ellipses fitted to the starting points for all stimuli. **a**,**b**, Rotational dynamics in the "winning" ($\boldsymbol{r}_{\boldsymbol{R}}^{\mathrm{CW}}$; **a**) and the "losing" ($\boldsymbol{r}_{\boldsymbol{R}}^{\mathrm{CCW}}$; **b**) subpopulations of the reference-receiving population ($\boldsymbol{r}_{\boldsymbol{R}}$) for the relative reference of $-7.5°$. Note that $\boldsymbol{r}_{\boldsymbol{R}}^{\mathrm{CW}}$ and $\boldsymbol{r}_{\boldsymbol{R}}^{\mathrm{CCW}}$ start from different stable states, being rotated by $45°$ in the opposite directions, as indicated by the dots with low saturation around the dotted circles. **c**, Rotational dynamics of the stimulus-receiving, estimation-related population $\boldsymbol{r}_{\boldsymbol{\Theta}}$. Note that the rotation direction is the same as the "winning" direction. **d-f**, Projection of the population dynamics of heterogeneous (with drift) RNNs during the DM epoch on the common PC space used in **a-c**. Same as **a-c**, but for the population activities of heterogeneous RNNs. Note that the overall dynamics are warped by the stimulus-specific bias in the shape of an ellipse, with the major axes corresponding to oblique orientations.

RNN mechanism for WM and DM interactions

Next, we performed state space analyses to learn how DM is formed and influences WM in RNNs. To understand such processes without any influence from drift dynamics, we began by studying 'homogeneous RNNs,' which were independently trained with input lacking orientation-specific variability and thus do not exhibit drift dynamics or stimulus-specific bias. These RNNs have three subpopulations of neurons: those receiving stimuli ($\boldsymbol{r}_{\Theta}$), those receiving references and voting for clockwise decisions ($\boldsymbol{r}_{R}^{\mathrm{CW}}$), and those receiving references and voting for counter-clockwise decisions ($\boldsymbol{r}_{R}^{\mathrm{CCW}}$). We projected the noise-free activities of homogeneous RNNs using a low-dimensional subspace which was built from the population activity of $\boldsymbol{r}_{\Theta}$ (**Supplementary Information**).

When projected onto this subspace, $\boldsymbol{r}_{R}^{\mathrm{CW}}$ and $\boldsymbol{r}_{R}^{\mathrm{CCW}}$ reached their own stable states, each rotated around 45° in the opposite directions before the DM epoch (transparent color dots around the circle in **Fig. 8a,b**). During the DM epoch, upon the presentation of the reference tilted counter-clockwise (-7.5°), the winning ($\boldsymbol{r}_{R}^{\mathrm{CW}}$) and losing ($\boldsymbol{r}_{R}^{\mathrm{CCW}}$) sub-populations of the reference-receiving neurons expanded and contracted with the addition of the reference signal, respectively, while both were rotating in their preferred directions (saturating color dots along the arrows in **Fig. 8a,b**). These rotation-addition dynamics were also confirmed by the analysis of the connectivity weights between neurons (**Supplementary Fig. 8a-c**). Furthermore, with the loser's nonlinear dynamics, the winner's dynamics became dominant in the reference-receiving neurons (**Fig. 8a**,**b**). This led to the stimulus-receiving neurons repulsed away from the reference, rotating in the same direction as the winning reference-receiving neurons (saturating color dots along the arrows in **Fig. 8c**). Notably, the reference-receiving neurons quickly relaxed back to the pre-DM states (**Supplementary Fig. 8d**) and did not carry any reference-related influence during the post-DM delay period. Thus, as we assumed in the model that captured human behavior at a phenomenal level (**Fig. 3**), DM induced the immediate update of WM representation, whose effect remained only in the WM population in the homogeneous RNNs.

Next, to understand the impact of drift dynamics on the interplay between the stimulus-receiving and reference-receiving neurons, we examined the original RNNs ('heterogeneous RNNs'), which display stimulus-specific drifts. For fair comparisons, we projected the neural activities of the heterogeneous RNNs on the subspace built from the homogeneous RNNs. In heterogeneous RNNs, the activities of the reference-receiving neurons were warped such that the rotational dynamics were biased in an elliptic shape, while the expansion and contraction dynamics were quenched around the oblique orientations (**Fig. 8d-e**). Still, repulsion away from the reference was consistent with that in the homogeneous RNNs (**Fig. 8f**). In sum, our study of RNNs proposes the exertion of the rotation-addition dynamics by the DM-related neurons onto the WM-related neurons as a potential neural mechanism for the decision-*induced* bias in WM while suggesting that such exertion would vary in magnitude depending on stimulus orientation.

Discussion

The current study examined how humans form WM that biases and interplays with DM. Using a task paradigm where DM occurs at different time points in the time course of WM, we could access the drifting dynamics of WM and gain insights into how WM interacts with DM. Two novel aspects of our paradigm allowed for such accessibility and insights. First, the unprecedentedly long delay enabled us to probe WM states at moments sufficiently apart behaviorally and track their ongoing evolution via BOLD measurements. Second, the DM task based on mnemonic representations allowed us to preclude the stimulus evidence accumulation processes from interfering with WM representations, unlike previous studies where the stimulus was accompanied by the DM task[2,3,11,28]. Having confirmed the stimulus-specific and decision-consistent biases in our paradigm, we verified the predictions on the growth and decline of those two biases that are distinct between two different scenarios of WM dynamics. The behavioral, BOLD, and modeling results jointly supported the scenario with drift. Our findings offer a novel account of how WM representations continuously develop in a decision-consistent manner through stimulus-specific drift dynamics: a drifting WM representation leads to a bias in DM and then continues to drift in the same direction as that bias.

Our findings provide empirical support for drifting WM's contribution to stimulus-specific behavioral biases. A theoretical account based on efficient encoding points to the stimulus-specific sensory likelihoods constrained by neural discriminability as the origin of stimulus-specific biases[20,29–34]. Our approach also endorses the heterogeneity in sensory encoding. However, it goes beyond that by incorporating stimulus-specific drifts in WM evolution and validating the presence of such drifts in both behavioral and neural data. For that matter, our decoding of WM from the BOLD activity provides the first neural evidence for the gradual formation of stimulus-specific biases in WM over a span of tens of seconds. Given WM's contributions to stimulus-specific biases suggested in various domains, including orientation (*cardinal repulsion*)[21,30,35], color[36], and position[35], it seems worth investigating the unifying WM mechanism that relates drift dynamics to the stimulus-specific biases.

Our work provides novel insights into the mechanics of how the neural process of DM *induces* a bias in stimulus representation, namely the *decision-induced bias*. Previous accounts[2,11] have proposed a non-uniform weighting strategy as the source of decision-induced bias, but how this bias unfolds over time remains unknown. Our study on RNNs bridges this gap by suggesting that a subpopulation of neurons related to DM transiently exerts its rotational and expanding dynamics on WM-related neurons only at the moment of DM without any prolonged impact. Further, when trained without penalty for DM errors, RNNs stop displaying the increase of estimation errors in the near-reference conditions, a diagnostic signature of the decision-induced bias[3]. Based on these results, we propose transmitting the transient dynamics of rotation and addition from the DM-related to WM-related neural populations during the DM epoch as one possible neural mechanic responsible for the decision-induced bias. This proposal is not only distinct from previous proposals that claimed the decision-induced bias occurs in the late estimation epoch[3,28] but also readily testable by probing the DM-related and WM-related neural population responses with high temporal fidelity using the task paradigm developed in our study.

Our work refines the current understanding of the decision-*consistent* bias by stressing that it should be distinguished from the decision-*induced* bias and taken as a complex phenomenon with multiple origins. While the decision-induced bias refers to the estimation bias directly caused by the cognitive act of DM, the decision-consistent bias refers to any deviation of the mean of the error distribution conditioned on the choices of DM. As previously pointed out[3,10,11], the decision-consistent bias can be observed simply due to the stochastic noise in sensory encoding, even without decision-induced bias. Our work points to the drift dynamics of WM as another, previously unidentified, origin of the decision-consistent bias. Specifically, we demonstrated that the stimulus-specific drift initially biases WM, which in turn biases DM, and continues to bias WM after DM in the direction consistent with the bias in DM. Thus, to accurately claim the existence of decision-induced bias, it is necessary to take into account the drift dynamics in addition to the stochastic noise (diffusion dynamics). Otherwise, the post-decision drift in WM can be incorrectly ascribed to the decision-induced bias.

Our approach has intriguing implications for research on confirmation bias[37–39], a tendency to form evidence to support one's own decisions. In our view, the temporal evolution of the post-decision bias is attributable to the drift dynamics in WM, without resorting to attention[40,41] or hierarchical inference[42]. Meanwhile, our view remains agnostic to the role of subjective uncertainty or confidence in confirmation bias, which has been suggested previously[43–45]. Future research may investigate how uncertainty or confidence interacts with the drifting WM to contribute to the decision-consistent bias. Especially given recent suggestions of the shared representation of visual WM and its uncertainty[46,47], incorporating confidence reports in our paradigm might reveal the relationship between the trial-to-trial variability in confidence and the representational dynamics underlying the decision-consistent bias.

Cognitive computations are not carried out in isolation. Our approach of putting the interplay of DM and WM in a dynamic context conferred a new perspective on how the brain manages its internal representations before and after committing to a categorical decision. We have yet to fully understand how drift dynamics work across different tasks and timescales (e.g., representational drift[48,49]). Still, our work draws an initial sketch for comprehending the interacting origins of the biases accompanying humans' consecutive cognitive acts in everyday life.

Methods

Experiments

Participants This study was conducted in accordance with the guidelines of and under the approval of the Institutional Review Board of Seoul National University. 50 healthy individuals (30 females, $19-32$ years old; normal or corrected-to-normal vision) completed at least two or all of the three-day sessions of the main fMRI experiment across three days. Each participant provided written informed consent prior to the experiment and was naïve to the purpose of the study.

Experiment stimuli and procedure Stimuli were generated using MGL[50] and presented by an LCD projector (60Hz). Participants viewed the stimuli in the projection area at a visual angle of $22°$ (width) $\times$ $17°$ (height). Throughout the whole experiment, all the stimuli were presented within a whole field (eccentricity, $8.5°$) gray circular Gaussian envelope aperture on a black background. During the experiment, a black fixation dot (eccentricity, $0.07°$) and a surrounding black fixation ring (inner eccentricity, $0.83°$; outer eccentricity, $0.9°$) were constantly present at the center of the screen during the experiment. There were 20 runs across the three scanning sessions, each run consisting of 12 trials, although several runs from some of the participants were excluded due to excessive motion (more than two runs whose maximum motion across axes in rotation and translation surpassed T2*-weighted voxel size). Before the experiment scanning sessions, all participants took part in a practice session of approximately one hour, a few days before the main fMRI experiment.

Inside the scanner, the participants were asked to maintain central fixation throughout a single run. They were asked to give responses using a button box with linearly aligned keys labeled Key1 to Key4. Before starting each trial, the fixation dot dilated ($0.14°$) for 0.5s to cue the stimulus onset. Each trial started with the 1.5s presentation of an alternating (frequency, 8/3 Hz) donut-shaped oriented grating (spatial frequency, one cycle per degree) spanning the peripheral visual field (aperture radii: inner, $2°$; outer, $8.5°$). The orientations of the grating were uniformly sampled from $0°$ to $172.5°$ with a step size of $7.5°$. Following the offset of the target stimulus presentation, there was a first-epoch delay, a discrimination task, a second-epoch delay, and an estimation task. There were two possible combinations for the delays: early DM trials involved 4.5s first-epoch and 10.5s second-epoch delays, whereas late DM trials involved 10.5s first-epoch and 4.5s second-epoch delays. These combinations varied on a trial-by-trial basis.

In the discrimination task, an oriented reference frame, a virtual line connecting the two yellow nonius dots (mark size, $0.1°$) presented on the fixation ring, was presented. To cue the task onset, the fixation dot was turned yellow and transiently dilated to $0.14°$, 0.5s before the onset of the reference. The participants were asked to choose whether the target was tilted clockwise or counter-clockwise against the reference by pressing the Key2 (CCW) or Key3 (CW) button using their left or right thumb, respectively. The relative orientations of the reference to the target stimulus were uniformly selected from $[-21°, -4°, 0°, 4°, 21°]$, whose range approximately matches those of the previous studies[2,3]. The participants were asked to respond within the 1.5s time limit, but their responses were recorded without their knowledge, with a buffer time of 0.5s. To signal responses being made, the fixation dot dilated to $0.14°$ and turned blue for 0.75s. If the response was not made within 1.5s, the fixation dot dilated to $0.14°$ and turned red for additional 0.75s. The reference frame disappeared upon the button press.

In the estimation task, the participants were asked to reproduce the target stimulus from their memory by rotating the estimation frame of the two green nonius dots with Key2 (CCW rotation) and Key3 (CW rotation) buttons within the 4.5s. To cue the task onset, the fixation dot was turned green and transiently dilated to eccentricity $0.14°$, 0.5s prior to the onset of the estimation frame. The participants were asked to confirm their adjusted report by pressing the Key1 button using their thumbs. To signal their response being made, the fixation dot dilated to

$0.14°$ and turned blue for 0.75s. If the response was not made within 4.5s, the fixation dot dilated to $0.14°$ and turned red for an additional 0.75s. The starting orientation of the estimation nonius dots was randomly chosen from $0°$ to $180°$. The estimation task was followed by the 5.5s inter-trial interval (ITI). Each trial lasted 28 seconds, making a total run of 336 seconds. After one run was completed, the participants were prompted with a summary of their performance for the run.

MRI data acquisition and preprocessing The MR data were collected using a Siemens 3 Tesla Tim Trio with a 32-channel head matrix coil at the Seoul National University Brain Imaging Center. All of the participants participated in the T1-weighted, high-resolution ($0.8 \times 0.8 \times 0.8\text{mm}^3$) anatomical scans (repetition time (TR), $2.4s$; inversion time (TI), 1s; time to echo (TE), 2.19ms; flip angle (FA), $8°$). In the following three separate days, they participated in the main T2*-weight fMRI scanning sessions: the first day of the fMRI scanning included a retinotopy-mapping run (scan time: 96s), a hemodynamic impulse response function (HIRF) estimation run (scan time: 96s), and task runs (6 runs, scan time: 336s); for the second day, the fMRI scanning included the task runs (8 runs); in the third day, the fMRI scanning included the task runs (6 runs). Some participants did not exactly follow the above sessions due to technical glitches. For the retinotopy-mapping, HIRF, and task runs, the scan parameters were set as follows: voxel size, $2.3 \times 2.3 \times 2.3\text{mm}^3$; TR, 2.0s; TE, 30ms; FA, $77°$.

After the acquisition of fMRI scanner data, the initial preprocessing steps for the anatomical and functional images were done following the fMRIprep workflow[51], version 20.2.0. Along with default settings, we used the field map-free distortion correction option in fMRIprep (–use-syn-sdc).

ROI definition and voxel selection The ROIs V1, V2, and V3 were defined using standard traveling wave methods[52]. Two $15°$-wide wedge bowties on the vertical and horizontal meridians were used as stimuli as in our previous study[53]. To characterize the voxel-wise signal-to-noise ratio (SNR), we used a checkerboard whole-field impulse (eccentricity, $8°$) with 1/24Hz presentation frequency in the HIRF scan. Using the retinotopy scan, subjects' V1, V2, and V3 ROIs across the left and right hemispheres and across dorsal and ventral areas were defined and combined for the BOLD analysis. The voxel-wise SNR was defined for each voxel as the stimulus frequency (1/24Hz) amplitude in the HIRF scan divided by the average amplitude of the frequencies higher than the third harmonics. Voxels with SNRs under two were discarded, following our previous research[54].

After identifying the ROIs, each voxel's time series was then converted into percent signal change by dividing by its average over the entire time series. To minimize the artifacts, the fMRIprep-derived confounding variables were regressed out, consisting of white matter, CSF, and six additional three-dimensional motion regressors, along with the discrete cosine transform bases with frequencies lower than 0.008Hz to reduce low-frequency components. No additional spatial smoothing was applied to the data. The confounders were regressed out simultaneously to minimize the potential artifacts from the stepwise regression[55]. For all further analyses, the resulting time series were z-scored on a voxel-by-voxel and run-by-run basis.

Analysis of data

Stimulus-specific bias function. We defined stimulus-specific bias as a conditional mean of the estimation errors given the stimulus θ and characterized it as a smooth, periodic function $\kappa = \kappa(\theta)$ over the orientation domain $[0, \pi]$. To parametrize its idiosyncratic form for each participant, we used the least squares method to find $\hat{\kappa}(\theta) = \mathbf{v}(\theta)^\top \boldsymbol{\omega}^*$ minimizing the estimation error $\varepsilon(\theta) = \hat{\theta} - \theta$ as

$$\boldsymbol{\omega}^* = \mathrm{argmin}_{\boldsymbol{\omega}} \sum_{j=1}^{N_{\mathrm{trial}}} \left\| \mathbf{v}(\theta_j)^\top \boldsymbol{\omega} - \varepsilon(\theta_j) \right\|^2 \tag{1}$$

where $\mathbf{v}(\theta) = \left[1, \phi_1'(\theta), \cdots, \phi_{N_{\mathrm{basis}}}'(\theta)\right]^\top$ with von Mises probability density function ϕ_j with the center $2j\pi/N_{\mathrm{basis}}$ and the precision as $N_{\mathrm{basis}}/2$, with $N_{\mathrm{basis}} = 12$.

Stimulus-specific bias weight. To assess how stimulus-specific bias depends on the timing of DM, we estimated the weight of participant-specific $\hat{\kappa}(\theta)$ in explaining the stimulus-specific point of subjective equality in DM psychometric functions. We assessed the contribution of stimulus-specific biases in the early and late DM conditions by maximizing the following likelihood function

$$L^{\mathrm{DM}} = L^{\mathrm{DM}}\left(\hat{c}_1, \cdots \hat{c}_{N_{\mathrm{trial}}}\right) = \prod_{j=1}^{N_{\mathrm{trial}}} \psi\left(\theta_j, r_j \vartheta_j,\right)^{\hat{c}_j} \left(1 - \psi\left(\theta_j, r_j, \vartheta_j\right)\right)^{1-\hat{c}_j} \tag{2}$$

where the psychometric function Ψ is given as

$$\Psi = \Psi\left(\theta_j, r_j, \vartheta_j\right) = \lambda + (1 - 2\lambda) \cdot \Phi\left(r_j^*; \mu_j, \sigma_{\vartheta_j}^2\right) \tag{3}$$

with $\theta_j, r_j, \vartheta_j$ specify j^{th} trial stimulus orientation, absolute reference orientation, and decision timing ($\vartheta_j \in E, L$) corresponding to early and late DM conditions, respectively. $r_j^* = r - \theta$ denotes the relative reference orientation, and Φ is a Gaussian cumulative density function. To parameterize the stimulus-specific modulation depending on decision timing, we used $\mu_j = w_{\vartheta_j} \cdot \hat{\kappa}\left(\theta_j\right)$, where $w_{\mathrm{E}}, w_{\mathrm{L}}$ denote the bias weight. We fitted the maximum-likelihood parameters $\rho = \{w_{\mathrm{E}}, w_{\mathrm{L}}, \sigma_{\mathrm{E}}, \sigma_{\mathrm{L}}, \lambda\}$, where $\sigma_{\mathrm{E}} \leq \sigma_{\mathrm{L}}$ denote the slopes of psychometric functions, and λ is the lapse rate.

Decision-consistent bias function. We defined the decision-consistent bias as a conditional mean of the estimation task errors ϵ given binary choice $\hat{c} \in \{CW, CCW\}$, and denote it as $b = b(\hat{c}) = \mathbb{E}[\epsilon|\hat{c}]$. We decomposed b into the contributions prior to DM (b_{pre}) and after DM (b_{post}) as follows.

$$b = b_{\mathrm{pre}} + b_{\mathrm{post}} \tag{4}$$

Previous studies[2,3] revealed that decision-consistent bias was prominent only when the relative references were near the orientation. Thus, we constrained our computation of b to the near-reference conditions ($\theta - r \in -4°, 0°, 4°$) for further analyses.

Estimating b_{post} from behavior data. To estimate post-decision bias b_{post} from the behavior data in a "model-free" manner (*i.e.*, not relying on our models proposed elsewhere in this paper), we took the following steps. First, the decision-consistent bias b was computed straightforwardly as a separation in the estimation errors ϵ as

$$2 \cdot \hat{b} = \mathbb{E}[\epsilon | \hat{c} = CW] - \mathbb{E}[\epsilon | \hat{c} = CCW] \tag{5}$$

where the expectation is taken across the near-reference trials. Next, we estimated b_{pre} by leveraging the discrimination task responses to calculate the summary statistics of the underlying across-trial distribution of $\hat{\epsilon}_{t_{\text{DM}}}$, the estimated memory error at t_{DM}. Specifically, reusing psychometric functions Ψ in Eq. 3 for the data averaged over different stimuli, we fitted another set of maximum-likelihood parameters $\rho' = \mu, \varsigma_{\text{E}}, \varsigma_{\text{L}}, \lambda'$, where μ and ς denote the center and standard deviation of $\hat{\epsilon}_{t_{\text{DM}}}$ for early and late DM timing with $\varsigma_{\text{E}} \leq \varsigma_{\text{L}}$, and λ' is a lapse rate. Using these parameters, we performed a Monte Carlo simulation of $\hat{\epsilon}_{t_{\text{DM}}} \sim \mathcal{N}(\hat{\mu}, \hat{\varsigma}^2)$ with the number of simulations $N_{\text{MC}} = 10^5$ for each of the early and late DM conditions. Then, we defined the model's choice $\hat{c}^* = \hat{c}^*(r^*, \hat{\epsilon}_{t_{\text{DM}}})$ for each simulation for a randomly selected $r^* \in -4°, 0°, 4°$ as

$$\hat{c}^* \sim Bernoulli(1 - \hat{\lambda}_1)\mathbf{1}_{\hat{\epsilon}_{t_{\text{DM}}} > r^*} + Bernoulli(\hat{\lambda}_1)\mathbf{1}_{\hat{\epsilon}_{t_{\text{DM}}} < r^*} \tag{6}$$

Next, to correct the effect of reference attraction[56], we obtained linear coefficient $\hat{\beta}_r$ by regressing ϵ against r^*, for each of the early and late conditions. Then, we corrected the Monte Carlo estimates as $\hat{\epsilon}^*_{t_{\text{DM}}} = (1 - \hat{\beta}_r) \cdot \hat{\epsilon}_{t_{\text{DM}}} + \hat{\beta}_r \cdot r^*$. With $\hat{\epsilon}^*_{t_{\text{DM}}}$ and $\hat{c}^*$, we estimated b_{pre} as

$$2 \cdot \hat{b}_{\text{pre}} = \mathbb{E}[\hat{\epsilon}^*_{t_{\text{DM}}} | \hat{c}^* = CW] - \mathbb{E}[\hat{\epsilon}^*_{t_{\text{DM}}} | \hat{c}^* = CCW] \tag{7}$$

Finally, we estimated the post-decision bias, b_{post}, by $\hat{b}_{\text{post}} = \hat{b} - \hat{b}_{\text{pre}}$.

BOLD decoding To decode the stimulus information encoded by the population of visual cortex voxels at a given time, we fitted a linear model[23,57] on the voxels combined across V1, V2, and V3. Each trial's BOLD response at a given time point (TR) was modeled as the linear combination of the channel responses assumed to be the unimodal, symmetric bases $y_\varphi(\theta) = |\cos(\theta - \varphi)|^8$ for the trial's stimulus θ and the channel center at φ.

$$\mathbf{X} = \mathbf{W}\mathbf{Y} \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{N_{\text{voxels}} \times N_{\text{trials}}}$ denotes the BOLD responses, and $\mathbf{Y} \in \mathbb{R}^{N_{\text{channels}} \times N_{\text{trials}}}$ denotes the design matrix whose each column corresponds to the channel response vector $\mathbf{y}(\theta_j) = [y_{\varphi_1}(\theta_j), \cdots, y_{\varphi_8}(\theta_j)]^\top$ for the trial's stimulus θ_j, in which the basis functions were selected as

uniformly paced, each centered on $\varphi_1, \cdots, \varphi_8$ tiling the orientation space $[0, \pi]$. $\mathbf{W} \in \mathbb{R}^{N_{\text{voxels}} \times N_{\text{channels}}}$ was the matrix of linear weights.

To fit the linear weights for the time courses associated with different DM timing conditions, we followed a leave-one-run-out cross-validation procedure for each DM timing. For a given run and a DM timing, we split the trials into the "train" trials (all the trials of the DM timing but the run) and "validation" trials (all the trials of the DM timing in the run), resulting in the train data pair $(\mathbf{X}_T, \mathbf{Y}_T)$ and the held-out validation data $\mathbf{X}_V$. Given $\mathbf{X}_T$ and $\mathbf{Y}_T$, we fit the weight by the least-squares method as

$$\widehat{\mathbf{W}} = \mathbf{X}_T {\mathbf{Y}_T}^\top \left(\mathbf{Y}_T {\mathbf{Y}_T}^\top\right)^{-1} \tag{9}$$

To reconstruct the channel responses for the held-out validation data $\mathbf{X}_V$, we used

$$\widehat{\mathbf{Y}}_V = \left(\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}}\right)^{-1} \widehat{\mathbf{W}}^\top \mathbf{X}_V \tag{10}$$

To reconstruct the channel responses in a near-continuous space, we repeated the process of estimating weights and channel responses 15 times, in which the centers of the 8 channel basis functions were equally shifted by 1.5° with the resulting 120-dimensional reconstructed channel responses in total, with the corresponding channel centers $\boldsymbol{\varphi} = [0°, 1.5°, \cdots, 178.5°]^\top$ uniformly covering the orientation space $[0, \pi]$. We repeated the whole process for each run, the DM timing, and the time point, resulting in the reconstruction of channel response vectors $\widehat{\mathbf{Y}}(t)$, with the columns corresponding to all the trials for a given time point t of interest among 3-14 TRs.

To decode the stimulus information in trial j from the reconstructed channel responses, we mapped each of the reconstructed channel responses to a point readout on the circular stimulus space.

$$\hat{\theta}^{\text{BOLD}}{}_j(t) = arctan\left(\sin(2\boldsymbol{\varphi})^\top \widehat{\mathbf{Y}}_j(t) / \cos(2\boldsymbol{\varphi})^\top \widehat{\mathbf{Y}}_j(t)\right) \tag{11}$$

where $\widehat{\mathbf{Y}}_j(t)$ denotes the estimated channel response functions of the trial j and TR t.

Models

Drift-diffusion models To assess the WM process underlying stimulus-specific and decision-consistent biases in participants' behavior and cortical responses, we modeled each trial's memory dynamics $m(t)$ as a one-dimensional stochastic integral equation over the domain of orientation in $[0, \pi]$, described by

$$m_t = m_0 + \int_0^t K\,(m_s) ds + \int_0^t D\,(m_s) dB_s + \beta\left(m_{t_{\text{DM}}}, r\right) \cdot \mathbf{1}_{t \geq t_{\text{DM}}} \tag{12}$$

where $K(m_s)$ and $D(m_s)$ denote the drift and diffusion functions of the memory process m_s, respectively, and B_s is the Brownian motion. For the initial condition m_0, we leveraged a model of perceptual encoding based on efficient coding[20].

DM is computed by comparing the absolute reference orientation r to m, namely, $\hat{c} = \hat{c}(m,r) = sign(m-r)$. WM updating at DM timing, induced by the binary choice, was modeled as impulse weighted $\beta(m,r)$ given as

$$\beta(m,r) = \begin{cases} w_\beta \cdot \hat{c}(m,r) + w_r \cdot r^*, & r^* \in -4°, 0°, 4° \\ w_r \cdot r^*, & r^* \in -21°, 21° \end{cases} \tag{13}$$

where r^* denotes the relative reference orientation, and w_β is the strength of decision-induced bias. Respecting the previous literature on decision-induced bias[2,3], we restricted the impact of DM to the proximal relative reference conditions, $r^* \in -4°, 0°, 4°$, theorized to underlie the 'near-reference variability.' w_r is the strength of reference attraction, based on the previous literature on towards-distractor biases.

We considered the drift-diffusion model (full model) and the drift-only model (reduced model) as two parametric subsets of the integral equation. While both models posit the same initial encoding m_0, the former assumes diffusion dynamics only, while the latter assumes both drift and diffusion. For both models, we assumed a constant diffusion coefficient with $D(m) = w_D^2$. For the full model, we defined the drift function $K(m)$ as

$$K(m) = w_K \cdot \hat{\kappa}^\dagger(m) \tag{14}$$

where $\hat{\kappa}^\dagger \equiv \hat{\kappa}/\max|\hat{\kappa}|$ with $\hat{\kappa}$ from Eq. (1), and w_K is a scaling factor of the drift rate. For the reduced model, we used $w_K = 0$. We considered the Fokker-Planck equation equivalent to Eq. 12 to fit these drift-diffusion models to the behavior. See Supplementary Information for further details on the encoding model and fitting procedure.

Matching drift-diffusion models with BOLD responses. To evaluate the model prediction of continuous dynamics, we compared trial-to-trial BOLD response trajectories to the drift-diffusion models fitted to each participant's behavior data. To construct the time-dependent hemodynamic response $\nu(t)$ from the fitted drift-diffusion model $m(t)$, we use the following complex exponential notation for convolution on a circular space,

$$e^{2i\nu(t)} = h(t) * \left(e^{2im(t)} + \rho_\theta \cdot e^{2i\theta} \cdot \mathbf{1}_{t\in\mathcal{T}_\theta} + \rho_r \cdot e^{2ir} \cdot \mathbf{1}_{t\in\mathcal{T}_r}\right) \tag{15}$$

where $h(t)$ denotes the canonical double-gamma hemodynamic response function[58]. ρ_θ *and* ρ_r denote the strengths of visual events driven by the presentation of the stimulus θ and the absolute reference r, respectively. $\mathcal{T}_\theta$ and $\mathcal{T}_r$ define the time windows of the boxcar visual events. We used $\mathcal{T}_\theta = [0,1.5]$ for stimulus presentation time window, and $\mathcal{T}_r = [6,6+\tau_D]$ and $\mathcal{T}_r = [12,12+\tau_D]$ for reference presentation time windows with median response time τ_D inferred from the data in early and late DM conditions, respectively. To estimate ρ_θ *and* ρ_r, we used the population data under far-reference conditions, which were held-out from the BOLD data analysis elsewhere. We

minimized the L2 distance between the far-reference condition decoding trajectories to the ideal trajectory under the boxcar-like visual drives, namely, $h(t) * \left(1 + \rho_\theta \cdot \mathbf{1}_{t\in\mathcal{T}_\theta} + \rho_r \cdot e^{2ir^*} \cdot \mathbf{1}_{t\in\mathcal{T}_r}\right)$.

Estimating b_{post} from BOLD data. To estimate the contribution of post-decision bias b_{post} from the BOLD data in a model-free manner, we designed a piece-wise linear function that allows for a change in slope or intercept at t_{DM}, given by

$$g(t) = (\alpha_0 + \beta_0 \cdot t)\mathbf{1}_{t\le t_{\text{DM}}} + \left(\alpha_1^* + \beta_1 \cdot (t - t_{\text{DM}})\right)\mathbf{1}_{t>t_{\text{DM}}} \tag{16}$$

where $\alpha_1^* = \alpha_0 + \alpha_1 + \beta_0 \cdot t_{\text{DM}}$. Note there exists a discontinuous jump at t_{DM} by α_1, corresponding to the decision-induced bias. To estimate b_{post}, we used each participant's decision-conditioned near-reference error trajectory for each early and late DM condition. We minimized the L_2 discrepancy between this trajectory and the boxcar trajectory $h(t) * \left(e^{2ig(t)} + \rho_\theta \cdot \mathbf{1}_{t\in\mathcal{T}_\theta} + \rho_r \cdot \mathbf{1}_{t\in\mathcal{T}_r}\right)$ to find $\hat{\alpha}_0, \hat{\beta}_0 \hat{\alpha}_1, \hat{\beta}_1$ for each participant. Using these quantities, we computed $\hat{b}_{\text{pre}} = \hat{\alpha}_0 + \hat{\beta}_0 \cdot t_{\text{DM}}$ and $\hat{b}_{\text{post}} = \hat{\alpha}_1 + \hat{\beta}_1 \cdot (t_{\text{EM}} - t_{\text{DM}})$. Note that since we did not restrict the signs of $\alpha_0, \alpha_1, \beta_0, and\ \beta_1$, this method serves as an unbiased way of characterizing Δb_{pre} and Δb_{post}.

Recurrent neural network models. The following equations describe the dynamics of RNN

$$\tau \frac{d\mathbf{r}}{dt} = -\mathbf{r} + f(\mathbf{J}\mathbf{r} + \mathbf{J}_\Theta \mathbf{u}_\Theta + \mathbf{J}_R \mathbf{u}_R + \boldsymbol{\eta}) \tag{17}$$

$$\tau' \frac{d\boldsymbol{\eta}}{dt} = -\boldsymbol{\eta} + \sqrt{2\tau'}\xi \tag{18}$$

where $\mathbf{r}$ and $\boldsymbol{\eta}$ are N_{rec}-dimensional unit activity and noise modeled as Ornstein–Uhlenbeck process. τ and τ' denote the time constants of recurrent units and noise, which were 100 ms and 200 ms, respectively. $\mathbf{u}_\Theta\ and\ \mathbf{u}_R$ are N_{in}-dimensional stimulus and reference inputs, whose units are tuned according to a von Mises distribution. We used $N_{\text{rec}} = 96$ and $N_{\text{in}} = 24$. $f = 1/\left(1 + exp(-x)\right)$ is the sigmoid activation function, and $\mathbf{J}, \mathbf{J}_\Theta\ and\ \mathbf{J}_R$ are recurrent, stimulus input, and reference input weights. ξ is independent Gaussian noise with a standard deviation of 0.05. The initial conditions for $\mathbf{r}$ and $\boldsymbol{\eta}$ were set to be zero.

As stimulus and reference inputs occupied different parts of the visual fields in the human experiment, $\mathbf{u}_\Theta\ and\ \mathbf{u}_R$ projected to two separate populations of the recurrent units $\mathbf{r}$, namely $\mathbf{r}_\Theta, \mathbf{r}_R$, each of which was 48-dimensional. We used the forward Euler approximation of the above equation with the discretization time step $\Delta t = 20$ms. For the discrimination and estimation outputs, $\mathbf{z}^{\text{DM}}\ and\ \mathbf{z}^{\text{EM}}$, we used the following linear mapping,

$$\mathbf{z}^{\text{DM}} = \mathbf{J}_{\text{DM}}\mathbf{r},\ \ \mathbf{z}^{\text{EM}} = \mathbf{J}_{\text{EM}}\mathbf{r} \tag{19}$$

where $\mathbf{z}^{\mathrm{DM}} = (z_1^{\mathrm{DM}}, z_2^{\mathrm{DM}})$ with z_1^{DM} *and* z_2^{DM} corresponding to CW and CCW decisions, respectively, and $\mathbf{z}^{\mathrm{EM}}$ is a 24-dimensional 'labeled line' response vector, each corresponding to the equally discretized points on the periodic orientation domain $[0, \pi]$.

The task we used for training and generalizing RNNs was similar to the paradigm used for human participants. We first trained the RNNs to a short task timescale ("train episode") and generalized them to an extended task timescale ("generalization episode"). To model the orientation-specific variability respecting the natural scene statistics, we added variability to stimulus input $\mathbf{u}_\Theta$ as

$$p(\Theta|\theta) \sim \mathcal{N}(\theta, \gamma_D \cdot |\sin(2\theta)|^2) \tag{20}$$

with $\gamma_D = 10^2$. For the "ground truth" of the supervised learning given the input pairs $(\mathbf{u}_\Theta, \mathbf{u}_R)$, we defined the desired outputs $\mathbf{q}^{\mathrm{DM}}$ *and* $\mathbf{q}^{\mathrm{EM}}$ as

$$\mathbf{q}^{\mathrm{DM}} = [\mathbf{1}_{\theta > r}, \mathbf{1}_{\theta \leq r}]^\top \tag{21}$$

$$(\mathbf{q}^{\mathrm{EM}})_i = \exp\big(\kappa_\Theta(\cos(\theta - \theta_i) - 1)\big) \tag{22}$$

such that $\mathbf{q}^{\mathrm{DM}}, \mathbf{q}^{\mathrm{EM}}$ were 2-dimensional and 24-dimensional vectors, respectively. We trained recurrent weight $\mathbf{J}$ while maintaining other weights fixed. The joint loss $\mathcal{L}$ was computed by time-averaging the cross entropies between the network output and the ground-truth target.

To dissociate the contributions of drift dynamics and decision-induced bias in RNNs, we independently trained 50 "homogeneous" RNNs, assuming the veridical encoding $\Theta|\theta = \theta$, along with the original "heterogeneous" RNNs. To inspect the properties of the sub-populations, we separated CW-projecting and CCW-projecting populations $\mathbf{r}^{\mathrm{CW}}$ and $\mathbf{r}^{\mathrm{CCW}}$, along with stimulus-receiving population $\mathbf{r}_\Theta$. To compare the RNN dynamics on the common ground, we projected the population activities of homogeneous and heterogeneous RNNs onto the shared subspace spanned by $\mathbf{r}_\Theta$ of homogeneous RNNs. See Supplementary Information for further details on the task structure, RNN training, and state-space analysis.

Data Availability

Behavior and fMRI Data used here will be available upon acceptance for publication.

Code Availability

Custom code scripts, written in Python, for analyzing behavior and fMRI data, training RNN models, and generating the main figures will be available upon publication in the first author's GitHub repository.

References

1. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).

2. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).

3. Luu, L. & Stocker, A. A. Post-decision biases reveal a self-consistency principle in perceptual inference. *Elife* **7**, 1–24 (2018).

4. Gold, J. I. & Shadlen, M. N. The Neural Basis of Decision Making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).

5. Mante, V., Sussillo, D., Shenoy, K. V & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

6. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).

7. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).

8. Machens, C. K., Romo, R. & Brody, C. D. Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. *Science (80-.).* **307**, 1121–1124 (2005).

9. Stocker, A. A. & Simoncelli, E. P. A Bayesian model of conditioned perception. *Adv. Neural Inf. Process. Syst. 20 - Proc. 2007 Conf.* (2009).

10. Fritsche, M. & de Lange, F. P. Reference repulsion is not a perceptual illusion. *Cognition* **184**, 107–118 (2019).

11. Zamboni, E., Ledgeway, T., McGraw, P. V & Schluppeck, D. Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proc. R. Soc. B Biol. Sci.* **283**, 20160263 (2016).

12. Meister, M. L. R., Hennig, J. A. & Huk, A. C. Signal Multiplexing and Single-Neuron Computations in Lateral Intraparietal Area During Decision-Making. *J. Neurosci.* **33**, 2254–2267 (2013).

13. Murray, J. D., Jaramillo, J. & Wang, X. J. Working memory and decision-making in a frontoparietal circuit model. *J. Neurosci.* **37**, 12167–12186 (2017).

14. Wang, X. J. Decision Making in Recurrent Neuronal Circuits. *Neuron* **60**, 215–234 (2008).

15. Rademaker, R. L., Park, Y. E., Sack, A. T. & Tong, F. Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **44**, 925–940 (2018).

16. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).

17. Compte, A. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cereb. Cortex* **10**, 910–923 (2000).

18. de Gardelle, V., Kouider, S. & Sackur, J. An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. *J. Vis.* **10**, 6–6 (2010).

19. Tomassini, A., Morgan, M. J. & Solomon, J. A. Orientation uncertainty reduces perceived obliquity. *Vision Res.* **50**, 541–547 (2010).

20. Wei, X. X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).

21. Yu, Q., Panichello, M. F., Cai, Y., Postle, B. R. & Buschman, T. J. Delay-period activity in frontal, parietal, and occipital cortex tracks noise and biases in visual working memory. *PLOS Biol.* **18**, e3000854 (2020).

22. Serences, J. T., Ester, E. F., Vogel, E. K. & Awh, E. Stimulus-specific delay activity in human primary visual cortex. *Psychol. Sci.* **20**, 207–214 (2009).

23. Rademaker, R. L., Chunharas, C. & Serences, J. T. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat. Neurosci.* **22**, 1336–1344 (2019).

24. Friston, K. J. *et al.* Event-related fMRI: Characterizing differential responses. *Neuroimage* **7**, 30–40 (1998).

25. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X. J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).

26. Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat. Neurosci.* **25**, 783–794 (2022).

27. Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).

28. Luu, L. & Stocker, A. A. Categorical judgments do not modify sensory representations in working memory. *PLoS Comput. Biol.* **17**, 1–28 (2021).

29. Wei, X. X. & Stocker, A. A. Lawful relation between perceptual bias and discriminability. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10244–10249 (2017).

30. Taylor, R. & Bays, P. M. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *J. Neurosci.* **38**, 7132–7142 (2018).

31. Mao, J. & Stocker, A. A. Holistic inference explains human perception of stimulus

orientation. *bioRxiv* (2022).

32. Ganguli, D. & Simoncelli, E. P. Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Comput.* **26**, 2103–2134 (2014).

33. Hahn, M. & Wei, X.-X. A unifying theory explains seemingly contradicting biases in perceptual estimation. *bioRxiv* (2022).

34. Benjamin, A. S., Zhang, L. Q., Qiu, C., Stocker, A. A. & Kording, K. P. Efficient neural codes naturally emerge through gradient descent learning. *Nat. Commun.* **13**, 1–12 (2022).

35. Bae, G. NeuroImage Neural evidence for categorical biases in location and orientation representations in a working memory task. *Neuroimage* **240**, 118366 (2021).

36. Panichello, M. F., DePasquale, B., Pillow, J. W. & Buschman, T. J. Error-correcting dynamics in visual working memory. *Nat. Commun.* **10**, 1–11 (2019).

37. Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S. J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Comput. Biol.* **13**, e1005684 (2017).

38. Lord, C. G., Ross, L. & Lepper, M. R. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098–2109 (1979).

39. Nickerson, R. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2(2):175–2**, (1998).

40. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Curr. Biol.* **28**, 3128--3135.e8 (2018).

41. Talluri, B. C. *et al.* Choices change the temporal weighting of decision evidence. *J. Neurophysiol.* **125**, 1468–1481 (2021).

42. Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L. & Haefner, R. M. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *PLOS Comput. Biol.* **17**, e1009517 (2021).

43. Rollwage, M. *et al.* Confidence drives a neural confirmation bias. *Nat. Commun.* **11**, (2020).

44. Glickman, M., Moran, R. & Usher, M. Evidence integration and decision confidence are modulated by stimulus consistency. *Nat. Hum. Behav.* **6**, 988–999 (2022).

45. Peters, M. A. K. *et al.* Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1**, 1–8 (2017).

46. Li, H., Sprague, T. C., Yoo, A. H., Ma, W. J. & Curtis, C. E. Joint representation of working memory and uncertainty in human cortex. *Neuron* **109**, 3699-3712.e6 (2021).

47. Geurts, L. S., Cooke, J. R. H., van Bergen, R. S. & Jehee, J. F. M. Subjective confidence reflects representation of Bayesian probability in cortex. *Nat. Hum. Behav.* **6**, 294–305 (2022).

48. Qin, S. *et al.* Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nat. Neurosci.* **26**, 339–349 (2023).

49. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986-999.e16 (2017).

50. Gardner, J. L., Merriam, E. P., Schluppeck, D. & Larsson, J. MGL: Visual psychophysics stimuli and experimental design package, version 2.0. (2018). doi:10.5281/zenodo.1299497

51. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).

52. Engel, S. A. *et al.* fMRI of human visual cortex. *Nature* **369**, 525–525 (1994).

53. Choe, K. W., Blake, R. & Lee, S.-H. Dissociation between Neural Signatures of Stimulus and Choice in Population Activity of Human V1 during Perceptual Decision-Making. *J. Neurosci.* **34**, 2725–2743 (2014).

54. Ryu, J. & Lee, S.-H. Stimulus-Tuned Structure of Correlated fMRI Activity in Human Visual Cortex. *Cereb. Cortex* **28**, 693–712 (2017).

55. Lindquist, M. A., Geuter, S., Wager, T. D. & Caffo, B. S. Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *Hum. Brain Mapp.* **40**, 2358–2376 (2019).

56. Rademaker, R. L., Bloem, I. M., De Weerd, P. & Sack, A. T. The impact of interference on short-term memory for visual orientation. *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 1650–1665 (2015).

57. Brouwer, G. J. & Heeger, D. J. Decoding and Reconstructing Color from Responses in Human Visual Cortex. *J. Neurosci.* **29**, 13992–14003 (2009).

58. Glover, G. H. Deconvolution of Impulse Response in Event-Related BOLD fMRI. *Neuroimage* **9**, 416–429 (1999).

Supplementary Information

Supplementary Figures

1 Near-reference variability across different datasets . . . 2
2 Idiosyncrasy of stimulus-specific biases across participants . . . 3
3 Goodness-of-fit of the drift-diffusion models and the fitted parameters . . . 4
4 The shape of decision-consistent biases in behavior is captured by the full model . . . 5
5 Near-reference variability reproduced by the full model with nonzero bias parameter . . . 6
6 Quantification of the match between BOLD neural dynamics and drift-and-diffusion models . 7
7 Reduced models cannot reproduce gradual changes in working memory biases . . . 8
8 Rotation-addition mechanism and immediate impact of DM on WM in trained RNNs . . . 9

Contents

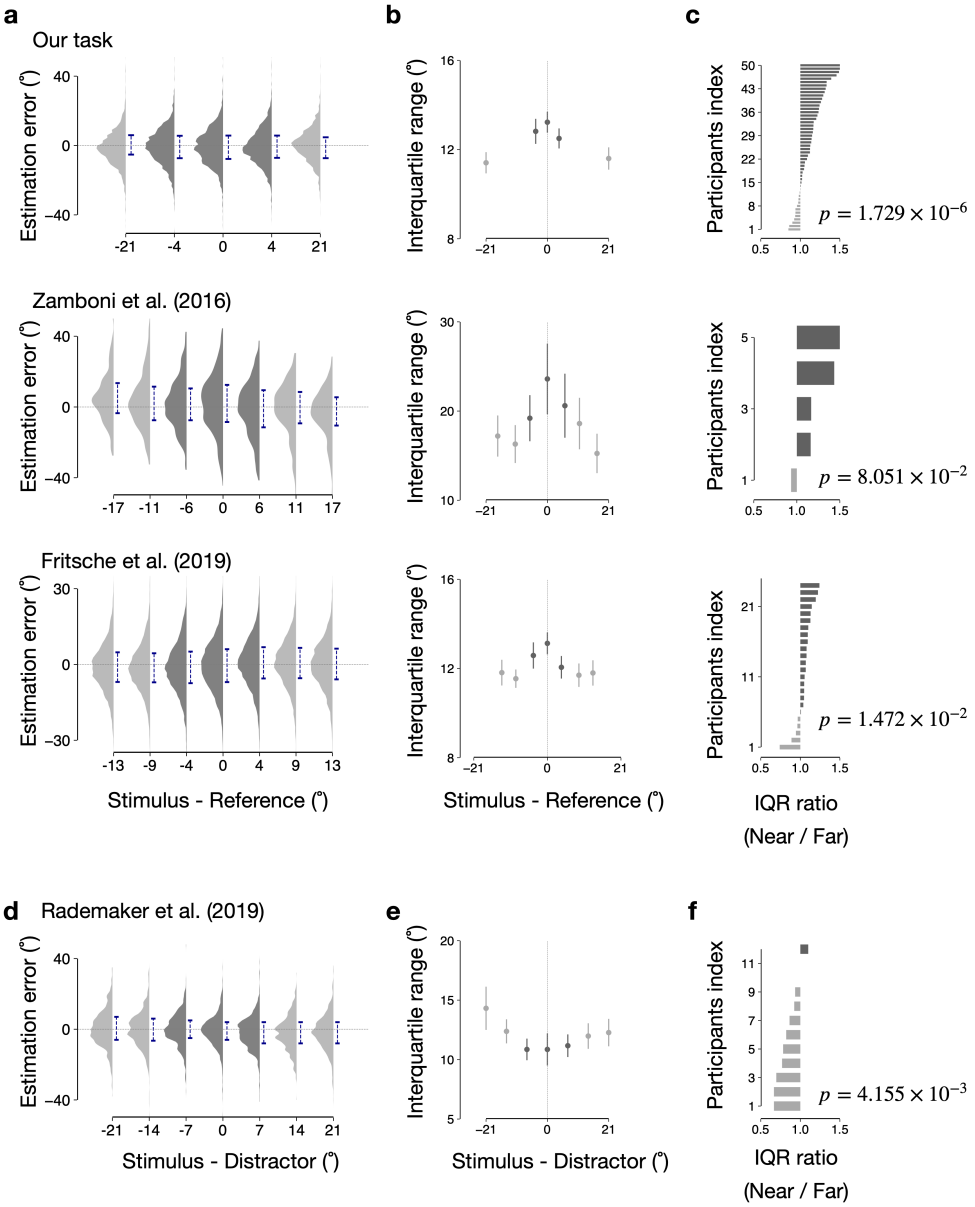
1 Analysis of BOLD responses **11**
1.1 Decoding of BOLD dynamics . . . 11
1.2 Estimating stimulus-specific bias weight from the BOLD data . . . 11
1.3 Matching dynamical models with the visual cortex BOLD responses . . . 11

2 Drift-diffusion model **12**
2.1 Task structure . . . 12
2.2 Efficient sensory encoding as an initial condition . . . 12
2.3 Model fitting . . . 13
2.4 Model parameters . . . 14

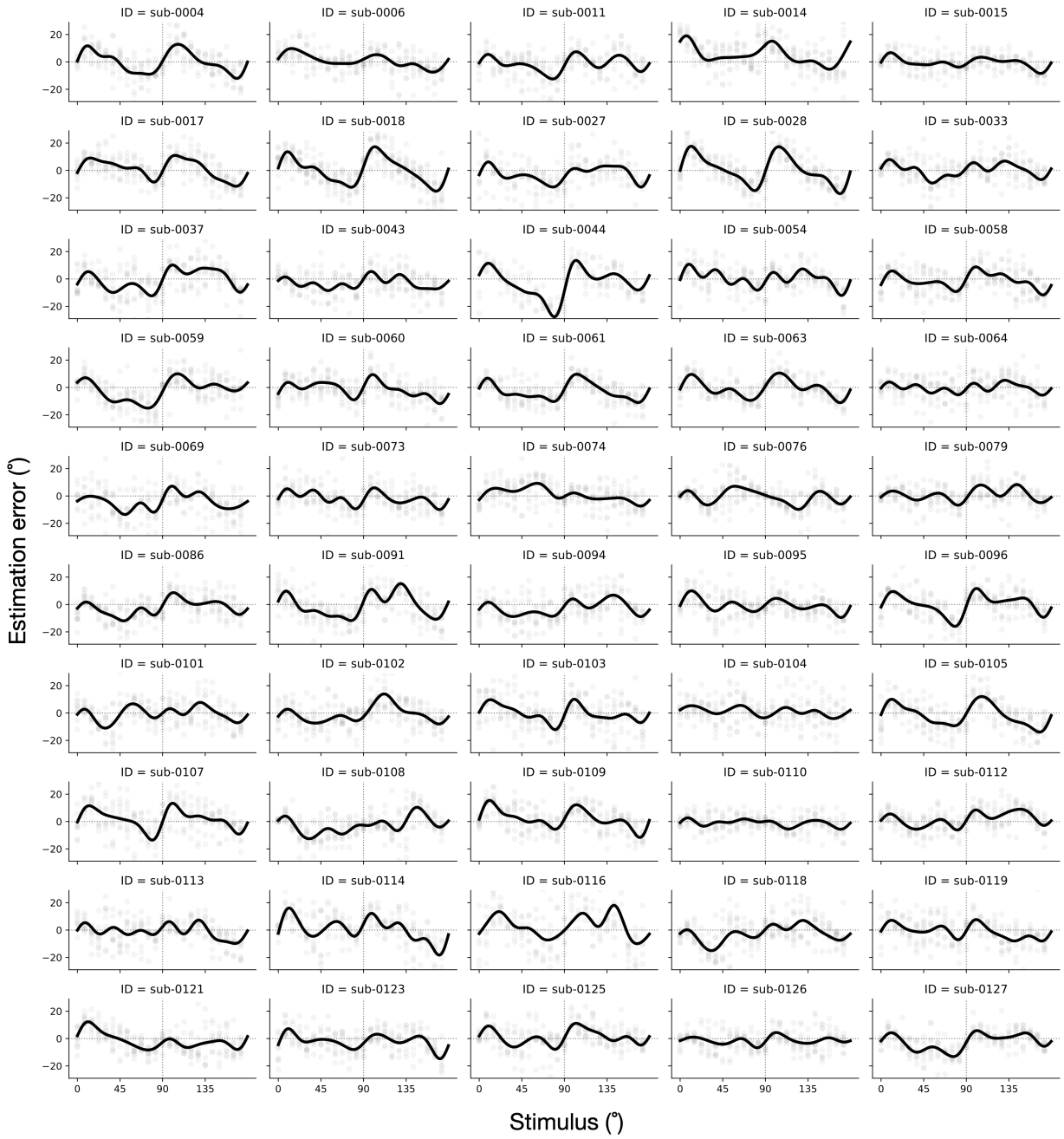
3 Recurrent neural network model **14**
3.1 Task structure . . . 14
3.2 Training procedure . . . 15
3.3 State space analysis . . . 16

4 Analysis of near-reference variability **16**
4.1 Definition of near-reference variability . . . 16
4.2 Analysis of different datasets . . . 17

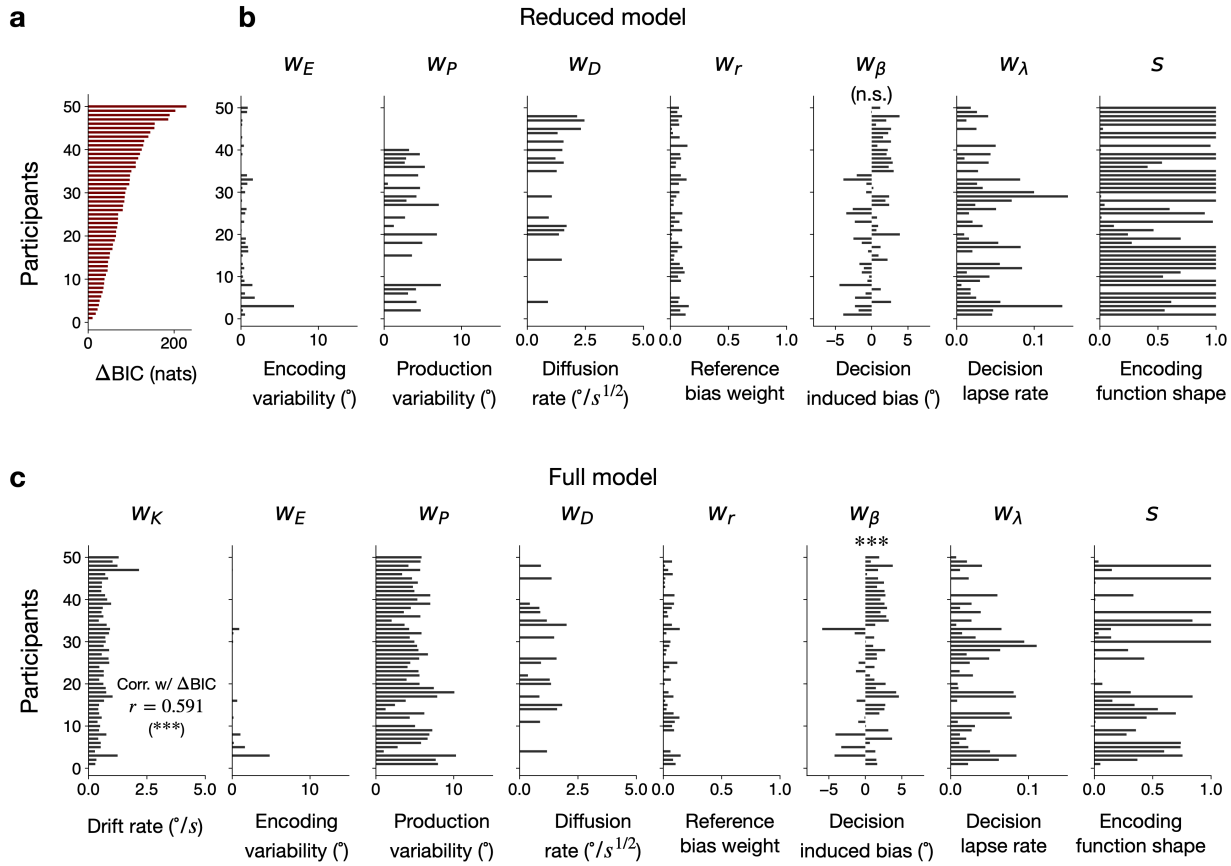
Supplementary References **17**



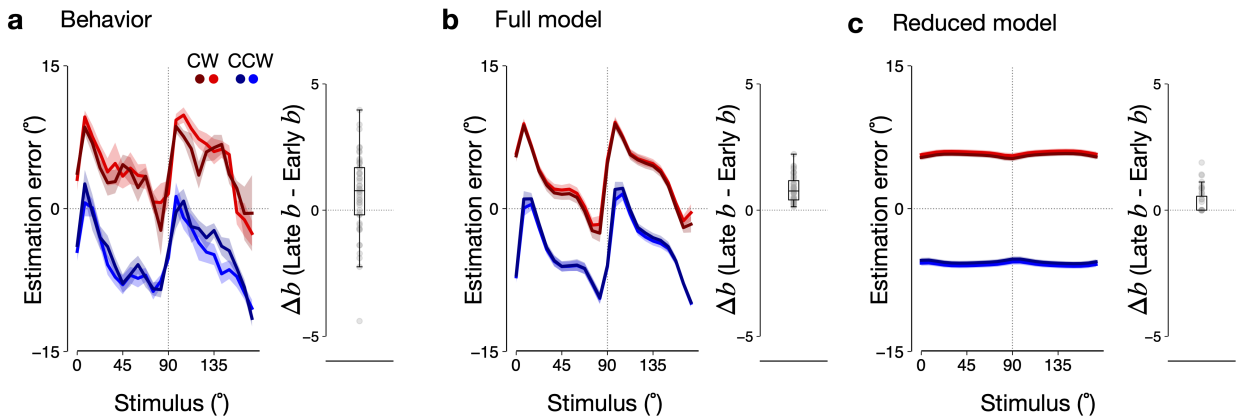
Supplementary Fig. 1 | Near-reference variability across different datasets. **a-c**, Presence of near-reference variability when the intervening stimulus serves as a boundary for decision (*i.e.*, reference), across the datasets of our own (top), Zamboni et al. (middle), and Fritsche and de Lange (bottom). **a**, Marginal distribution of estimation error as a function of stimulus orientation relative to the reference orientation. Two horizontal lines denote the 75% quantile (top) and 25% quantile (bottom). **b**, Interquartile range (IQR) as a function of stimulus orientation, relative to the reference orientation. Error bars denote ±s.e.m.s across participants. **c**, Ratio of IQR in the near-reference condition to that in far-reference conditions, for each of the participants. IQR ratio larger than 1 denotes evidence for the wider error distribution in the near-reference condition. Relative reference was considered "near" when (stimulus - reference) is within $[-8°, 8°]$ across all the data sets. IQR ratio was sorted in descending order across participants. One-sample t-tests against 1, $t\,(49) = 5.433$, 95% CI=$[1.087, 1.189]$ (top), $t\,(4) = 2.327$, 95% CI=$[0.954, 1.524]$ (middle), $t\,(23) = 2.637$, 95% CI=$[1.013, 1.103]$ (bottom). **d-f**, Absence of near-reference variability when the intervening stimulus serves as a distractor in the dataset of Rademaker et al. **f**, One-sample t-test against 1, $t\,(11) = -3.602$, 95% CI=$[0.762, 0.942]$.



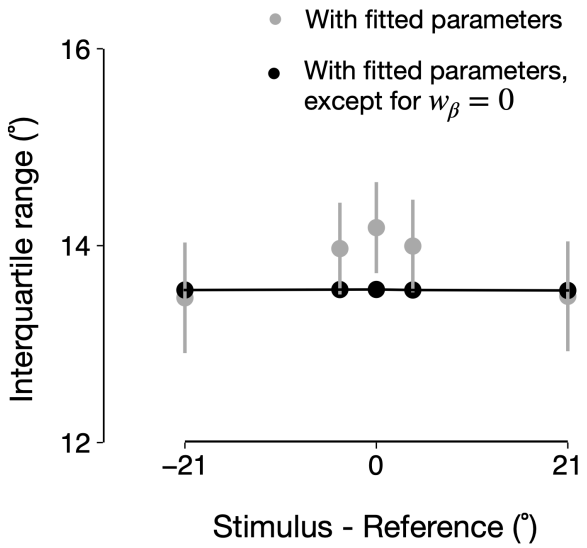
Supplementary Fig. 2 | Idiosyncrasy of stimulus-specific biases across participants. Distribution of estimation errors combined over the early and late DM conditions in each participant. Black lines show the fitted stimulus-specific bias function $\hat{\kappa}$ for each participant. Stimulus-specific biases were idiosyncratic: while cardinal repulsion is prevalent in most of the participants, sub-0076 (5th row and 4th column) shows a reversed pattern with a negative-slope bias pattern around 90°, which is the attraction towards the cardinal axis.



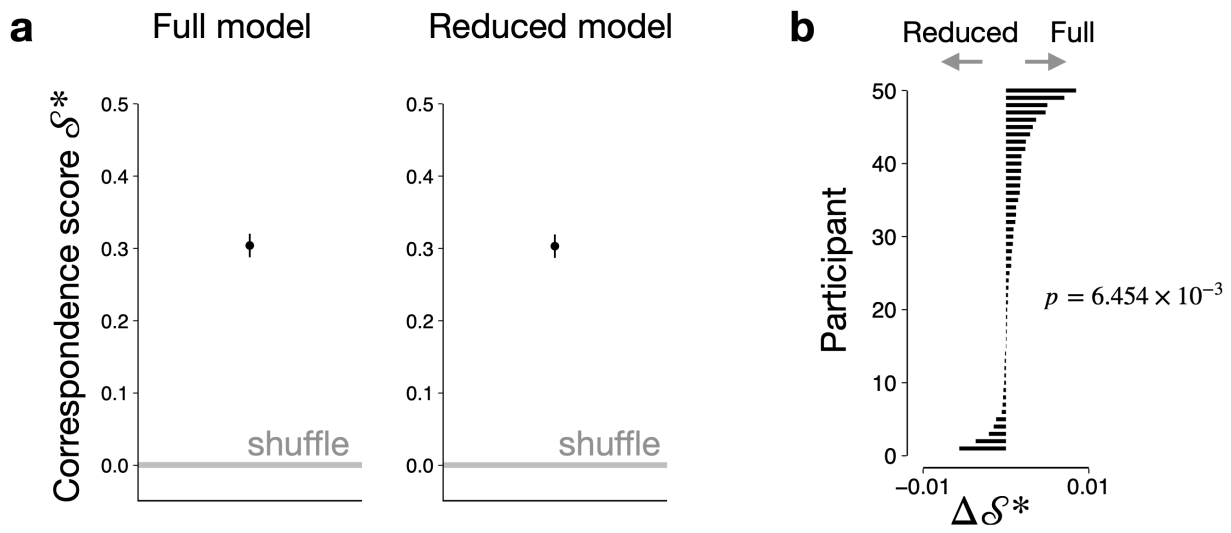
Supplementary Fig. 3 | Goodness-of-fit of the drift-diffusion models and the fitted parameters. **a**, Model comparison based on Bayesian information criterion (BIC) (in nats) across participants. Positive ΔBIC indicates the evidence in favor of the full model. Across all the participants, data suggested a better fit for the full model. **b,c**, Best-fitting parameters of the reduced model (**b**) and the full model (**c**). Participants were sorted according to their values of ΔBIC. We note that fitted decision-induced bias, w_β, in the reduced model was not significantly different from 0 (Paired samples t-test, $t\,(49) = 0.852$, $p = 0.398$, 95% CI=$[-0.359°, 0.888°]$), while w_β in the full model was significantly different from 0 (Paired samples t-test, $t\,(49) = 3.715$, $p = 5.218 \times 10^{-4}$, 95% CI=$[0.524°, 1.758°]$). Also, in the full model, the fitted memory drift rate, w_K, was significantly correlated with the gain in BIC (ΔBIC) (Pearson correlation, $p = 6.287 \times 10^{-6}$, 95% CI=$[0.374, 0.746]$). ***, $p < 0.001$, n.s., $p > 0.05$.



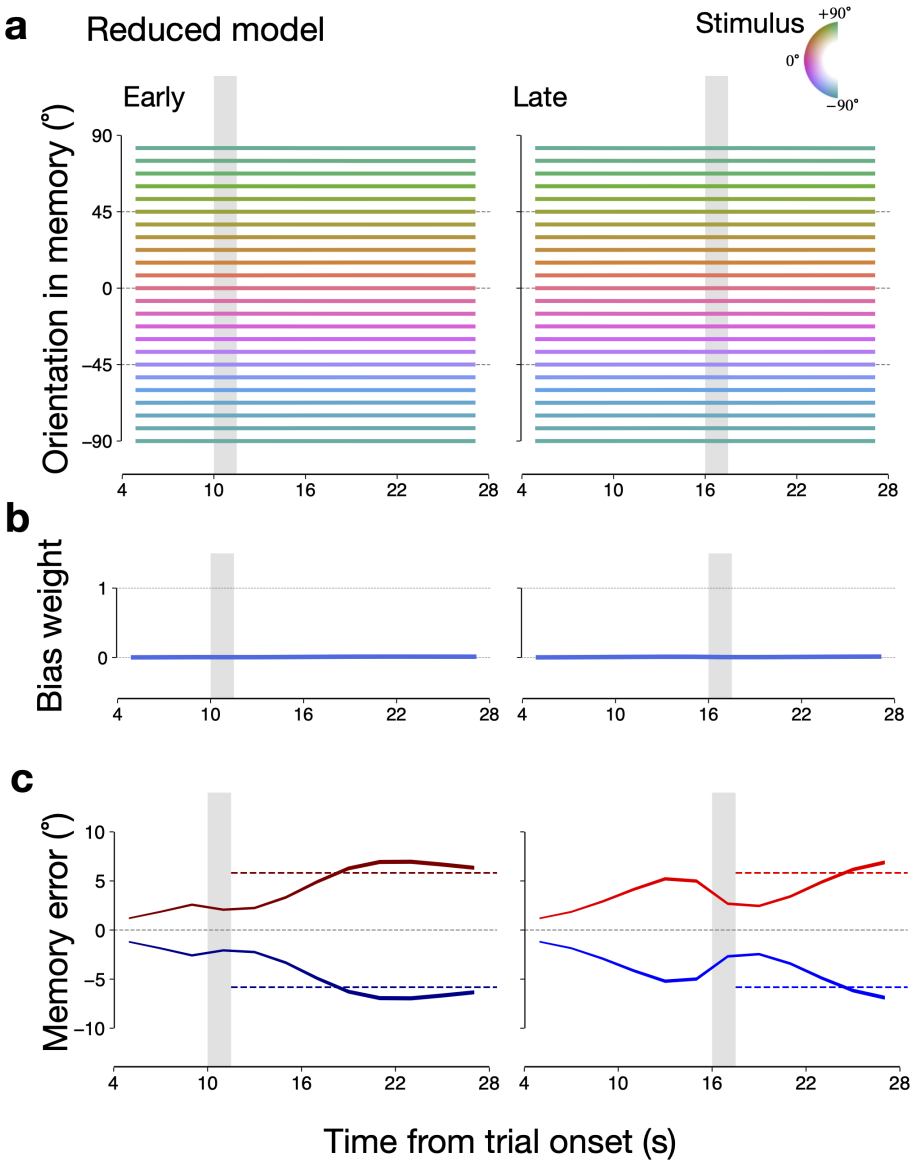
Supplementary Fig. 4 | The shape of decision-consistent biases in behavior is captured by the full model. **a**, (left) Estimation errors, conditioned on each choice and stimulus, in the early (darker) and late (brighter) DM conditions. Colored shades denote mean ± s.e.m.s across participants. (right) Box plot of the difference in the decision-consistent bias b between the early and late DM conditions, Δb. Each dot denotes a participant. **b-c**, (left) Model predictions of estimation errors with the best-fitting parameters for each participant, in the full model (**b**) and the reduced model (**c**). (right) Box plot of Δb with the best-fitting parameters for each participant in the full model (**b**) and the reduced model (**c**). For the reduced model, the population median was 0.



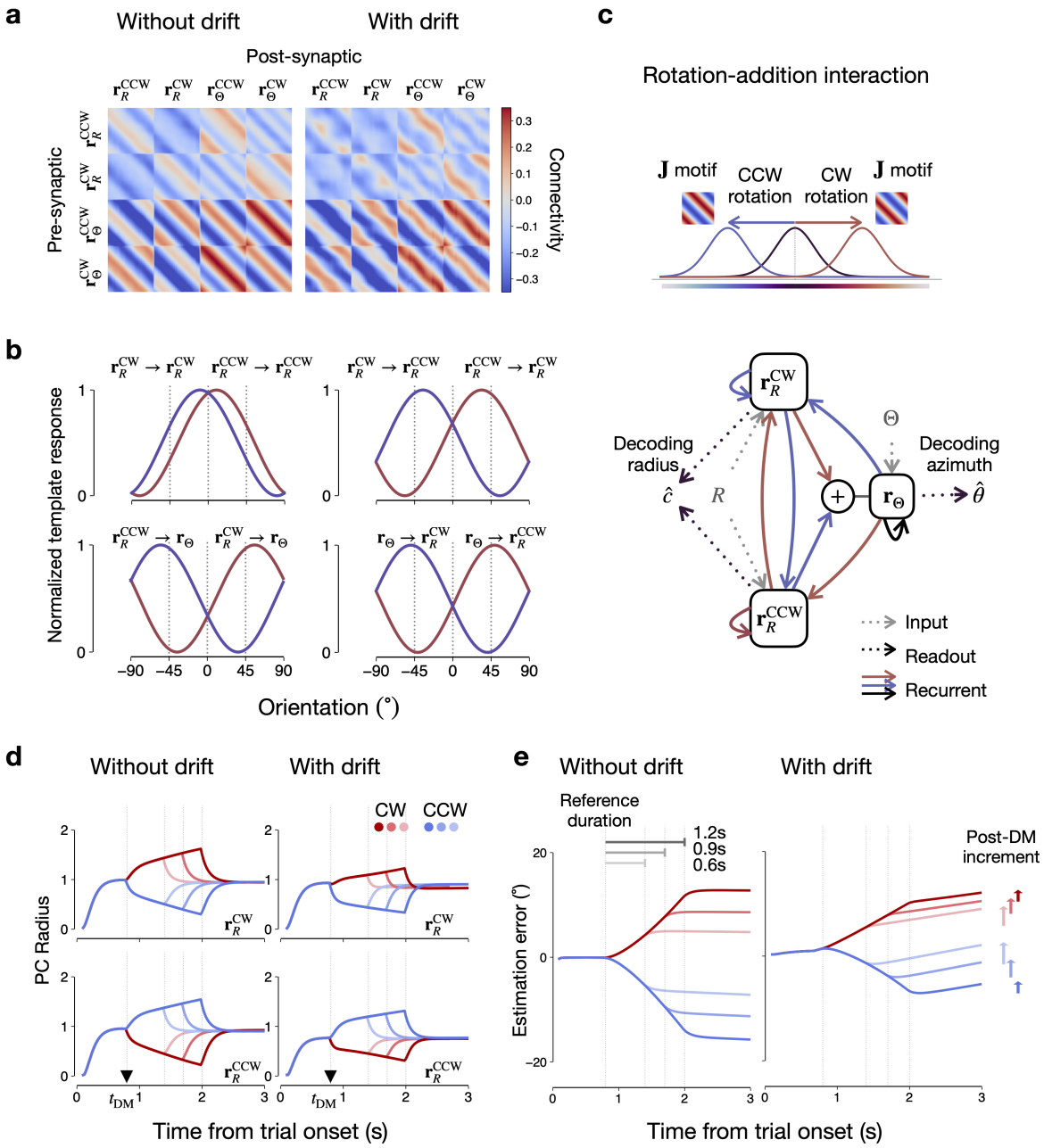
Supplementary Fig. 5 | Near-reference variability reproduced by the full model with nonzero bias parameter. Interquartile ranges in the marginal distribution of the estimation errors across relative reference conditions. The original full model with the best-fitting parameters for each participant (gray dots) reproduced the broader IQR observed in behavior data, while the full model with the best-fitting parameters except for $w_\beta = 0$, failed (black dots). Error bars denote ± s.e.m.s across participants.



Supplementary Fig. 6 | Quantification of the match between BOLD neural dynamics and drift-and-diffusion models. **a**, Cosine score between the BOLD decoding trajectory and the full dynamical model prediction (left) and the reduced dynamical model prediction (right), versus the shuffled model prediction (gray). Error bars denote ± s.e.m.s across participants. **b**, Comparison of the scores between the full and reduced models across participants. Paired samples t-test, $t\,(49) = 2.846$, 95% CI$=\left[2.692 \times 10^{-4}, 1.563 \times 10^{-3}\right]$.



Supplementary Fig. 7 | Reduced models cannot reproduce gradual changes in working memory biases. **a-c**, Evolution of stimulus estimation (**a**), bias weight (**b**) and decision-consistent bias (**c**) in early (left) and late (right) DM conditions in the reduced model without drift. Stimulus-conditioned errors do not exhibit bias towards the oblique orientations (**a**), nor the growth of stimulus-specific bias weights, *i.e.* the amplitude of stimulus-specific bias at each point relative to the behavior-derived stimulus-specific bias κ (**b**). On the other hand, decision-consistent bias grows, which is attributed to the recovery of the input-level and pre-decision separation of decision-conditioned errors (b_{pre}), as the impact of stimulus visual drive fades away. Colored dashed lines denote the decision-consistent bias at the onset of the post-decision epoch.



Supplementary Fig. 8 | (Continued on the following page.)

Supplementary Fig. 8 | Rotation-addition mechanism and immediate impact of DM on WM in trained RNNs. **a**, Average of the trained connectivity weight $\mathbf{J}$ of homogeneous RNNs (left) and heterogeneous RNNs (right). **b**, Template matching of the trained . Apparent clusters emerged from $\mathbf{J}$, each of which corresponds to the subpopulations $\mathbf{r}_R^{\mathrm{CW}}$, $\mathbf{r}_R^{\mathrm{CCW}}$, $\mathbf{r}_\Theta^{\mathrm{CW}}$, $\mathbf{r}_\Theta^{\mathrm{CCW}}$, based on their input-output mapping profiles. We multiplied each subpopulation with idealized block circulant matrices, each of which consisted of von Mises density functions with a precision of 0.1, with the centers tiling the orientation space $[0, \pi]$. The template response was normalized between zero and one for each case of interaction. **c**, Schematics of rotation-addition interaction between subpopulations. (top) Based on the template response profiles in **b**, we characterized the direction of rotational functions of each block in the connectivity matrix. (bottom) Input and readout mappings (dashed arrows) have no rotational functions as well as the self-recurrent loop of $\mathbf{r}_\Theta$. However, interactions between $\mathbf{r}_R^{\mathrm{CW}}$, $\mathbf{r}_R^{\mathrm{CCW}}$ and $\mathbf{r}_\Theta$ are rotational where blue is in the CCW direction and red is in the CW direction. The projections from $\mathbf{r}_R^{\mathrm{CW}}$ and $\mathbf{r}_R^{\mathrm{CCW}}$ onto $\mathbf{r}_\Theta$ are first rotated in the opposite direction and summed. **d**, Temporary decision-related changes in $\mathbf{r}_R^{\mathrm{CW}}$ and $\mathbf{r}_R^{\mathrm{CCW}}$. We varied the durations of the DM epoch as 0.6s, 0.9s, and 1.2s. (left) For homogeneous RNNs, the reference input during the DM epoch induces the expansion of the to-be-selected (winner) subpopulation, and contraction of the not-to-be-selected (loser) subpopulation, with almost symmetric amount of radius deviations. After the DM, the expansion or contraction quickly relaxes back to the original level. (right) For heterogeneous RNNs, similar dynamics were observed, but with an asymmetric amount of radius deviations. Here, we used the stimulus near (*i.e.*, $+15°$) to the cardinal orientation $\theta = 90°$. **e**, Impact of the decision-induced bias in memory states for estimation. Even though the decision-related changes were short-lived in $\mathbf{r}_R^{\mathrm{CW}}$ and $\mathbf{r}_R^{\mathrm{CCW}}$, the immediately updated memory representation in $\mathbf{r}_\Theta$ persists after the disappearance of the reference. (left) For the homogeneous RNNs, the amount of decision-consistent bias is modulated by the duration of reference inputs. Yet, without drift dynamics, there is no post-DM increment, *i.e.*,b_{post} minus the impact of decision-induced bias. (right) For the heterogeneous RNNs, the amount of decision-consistent bias is modulated not only by the duration of reference inputs but also by drift. As the decision impact moves the memory to the states with different drift rates, the post-DM increment becomes pronounced in the CW choice case.

1 Analysis of BOLD responses

See the **BOLD decoding** and **Matching drift-diffusion models with BOLD responses** sections in the **Methods** for notation.

1.1 Decoding of BOLD dynamics

To inspect the dynamics of the stimulus-specific bias carried out by the visual cortex BOLD responses, we conditioned estimated channel responses $\widehat{\mathbf{Y}}$ on each stimulus $\theta_k^* \in \boldsymbol{\theta}^* = \{0°, 7.5°, \cdots, 172.5°\}$ used in the experiment for each time step t.

$$\hat{\theta}_k^*(t) = \arctan\left(\sin(2\boldsymbol{\varphi})^\top \sum_{j:\theta_j=\theta_k^*} \widehat{\mathbf{Y}}_j(t) \Big/ \cos(2\boldsymbol{\varphi})^\top \sum_{j:\theta_j=\theta_k^*} \widehat{\mathbf{Y}}_j(t)\right) \tag{S1}$$

Next, to inspect the dynamics of the decision-consistent bias from the BOLD responses, we conditioned the estimated channel responses on behavioral choices $\hat{c} \in \{\mathrm{CW}, \mathrm{CCW}\}$ made on the trial. To inspect the near-reference effects, we restricted our analyses to the near-reference trials. For CW case:

$$\hat{\theta}_{\mathrm{CW}}(t) = \arctan\left(\sin(2\boldsymbol{\varphi})^\top \sum_{j:\hat{c}_j=\mathrm{CW}} \widehat{\mathbf{Y}}_j(t) \Big/ \cos(2\boldsymbol{\varphi})^\top \sum_{j:\hat{c}_j=\mathrm{CW}} \widehat{\mathbf{Y}}_j(t)\right) \tag{S2}$$

For the visualization of the above analyses, we relied on the population data to achieve a sufficient signal-to-noise ratio. To estimate the reliability of $\hat{\theta}_k^*(t)$, we repeatedly computed the above equation for $10,000$ times by randomly sampling the trials from those corresponding to each stimulus θ_k^*.

1.2 Estimating stimulus-specific bias weight from the BOLD data

To evaluate how the stimulus-specific bias carried out by the BOLD signals depends on time, we computed the stimulus-specific bias, $\hat{\theta}_k^*(t) - \theta$, using each participant's BOLD data. Using the idiosyncratic stimulus-specific bias function $\hat{\kappa}(\theta)$ estimated from each participant's behavioral estimation task errors, we calculated the bias weight as the coefficient of linear regression of $\hat{\kappa}(\theta)$ without intercept against $\hat{\theta}_k^*(t) - \theta$ for each time point.

1.3 Matching dynamical models with the visual cortex BOLD responses

We evaluated the model correspondence for held-out BOLD data under the near references. To evaluate the correspondence between the model-based prediction $\nu_j(t)$ defined in the main text equation (15), and the visual cortex BOLD decoding $\hat{\theta}_j^{\mathrm{BOLD}}(t)$ for each trial j, we formulated the following "correspondence score" $\mathcal{S}$ using the average of the positive parts of cosine distance between the time series of $\nu_j(t)$ and $\hat{\theta}_j^{\mathrm{BOLD}}(t)$:

$$\mathcal{S}(t,j) = \mathbb{E}_{m_j(t)}\left(\cos\left[2\left(\hat{\theta}_j^{\mathrm{BOLD}}(t) - \nu_j(t)\right)\right]_+\right) \tag{S3}$$

whose range lies between 0 and 1. Rather than to directly calculate $\mathcal{S}$, we derived a "surrogate" correspondence score $\mathcal{S}^*$ such that it serves as a lower bound of the original score, $\mathcal{S}^* \leq \mathcal{S}$:

$$\begin{aligned}
\mathcal{S}(t,j) &= \mathbb{E}_{m_j(t)}\left[\cos\left[2\cdot\left(\hat{\theta}_j^{\mathrm{BOLD}}(t) - \nu_j(t)\right)\right]_+\right] && \text{(S4)}\\
&\geq \mathbb{E}_{m_j(t)}\left[z\left(2\cdot\hat{\theta}_j^{\mathrm{BOLD}}(t)\right)^\top z\left(2\cdot\nu_j(t)\right)/\zeta_j(t)\right] && \text{(S5)}\\
&= \frac{1}{\zeta_j(t)}\cdot\left[\sum_{u=1}^{U_1} h_{tu}\cdot\cos\left(2\cdot\left(\hat{\theta}_j^{\mathrm{BOLD}}(t) - \mathbb{E}\left(m_j(u)\right)\right)\right) + |e^{2ib_j(t)}|\cdot\cos\left(2\cdot\left(\hat{\theta}_j^{\mathrm{BOLD}}(t) - b_j(t)\right)\right)\right] && \text{(S6)}\\
&= \mathcal{S}^*(t,j) && \text{(S7)}
\end{aligned}$$

where $\mathbf{H} = (h_{ij})$ denotes the convolution matrix given by $h(t)$, the normalization factor $\zeta^j(t) = \left(\sum_{u=1}^{U_1} |h_{tu}|\right) + |e^{2ib_j(t)}|$, and $z(2 \cdot b_j(t)) = h(t) * \left(\rho_\theta \cdot e^{2i\theta} \cdot \mathbf{1}_{\mathrm{t}\in\mathcal{T}_\theta} + \rho_r \cdot e^{2ir} \cdot \mathbf{1}_{\mathrm{t}\in\mathcal{T}_\mathrm{r}}\right)$ with $z(\theta) = (\cos\theta, \sin\theta)^\top$. By virtue of using the surrogate score, we can directly use Fokker-Planck density without the need for Monte-Carlo estimation. Finally, we averaged $\mathcal{S}(t, j)$ across time and the near reference trials as

$$\mathcal{S}^* = \frac{1}{N_{\mathrm{near}}} \frac{1}{N_{\mathrm{TR}}} \cdot \sum_{j=1}^{N_{\mathrm{near}}} \sum_{t=T_s}^{T_e} \mathcal{S}^*(t, j) \tag{S8}$$

2 Drift-diffusion model

See the **Drift-diffusion models** section in the **Methods** for notation.

2.1 Task structure

Based on the temporal structure of the experiment (Fig. 1a), we parsimoniously modeled each task epoch as an instantaneous event point corresponding to the onset of the event. Specifically, starting from the stimulus onset at $t = 0s$, early discrimination task onset was at $t_{\mathrm{DM}} = 6s$, late discrimination task onset was at $t_{\mathrm{DM}} = 12s$, and estimation task onset was at $t_{\mathrm{EM}} = 18s$.

2.2 Efficient sensory encoding as an initial condition

A previously established encoding model[1,2] constrains our drift-diffusion models as an initial condition of the memory process. To apply the model to our formulation, we leveraged the following relationship between the stimulus-specific bias $\kappa(\theta)$ and the stimulus-to-sensory encoding function $\mathcal{F}(\theta)$[2]:

$$\mathcal{F}' \propto \left(\int \kappa d\theta\right)^{-1/2} \tag{S9}$$

under the constraint $\int \mathcal{F}'(\theta')d\theta' = 1$. To estimate $\mathcal{F}$, we approximated $\mathcal{K} = \int \kappa d\theta$ as follows:

$$\hat{\mathcal{K}}(\theta) = \mathbf{V}(\theta)^\top \boldsymbol{\omega}^* + C \tag{S10}$$

where C: constant of integration, $\boldsymbol{\omega}^*$: previously estimated weights for $\hat{\kappa}$, and $\mathbf{V}(\theta) = [\phi_1(\theta), \cdots, \phi_{N_{\mathrm{basis}}}(\theta)]^\top$ with ϕ_j: previously used von Mises probability density function with the center $2j\pi/N_{\mathrm{basis}}$ and the precision value as $N_{\mathrm{basis}}/2$, with $N_{\mathrm{basis}} = 12$. Thus, we estimated $\hat{\mathcal{F}}' \propto \left(\hat{\mathcal{K}}^\dagger\right)^{-1/2}$ such that

$$\hat{\mathcal{K}}^\dagger = s \cdot \frac{\hat{\mathcal{K}} - \min(\hat{\mathcal{K}})}{\max(\hat{\mathcal{K}}) - \min(\hat{\mathcal{K}})} + (1 - s) \tag{S11}$$

with the shape-controlling parameter s ranged in $[0, 1]$.

Next, we computed the encoding density $p(m_0|\theta)$ using the relation $m_0 = \mathcal{F}(\theta)$. With the estimated encoding function $\hat{\mathcal{F}}$, the density was calculated using the change of variables as follows:

$$p(m_0|\theta) = \frac{\hat{\mathcal{F}}'(m_0)}{2\pi \cdot I_0(k)} \cdot \exp\left(k \cdot \cos\left(\hat{\mathcal{F}}(m_0) - \hat{\mathcal{F}}(\theta)\right)\right) \tag{S12}$$

with I_0: the modified Bessel function of order 0, $k = 1/\sqrt{w_E}$ with w_E: encoding variability. As such, we fully constrained the distribution of the encoding process, $p(m_0|\theta)$, with $\hat{\kappa}$ and a set of two free parameters, s and w_E.

2.3 Model fitting

To fit these dynamical models to the behavioral reports of discrimination and estimation, we considered the following Fokker-Planck equation equivalent to the stochastic integral equation (12) in the main text:

$$\frac{\partial}{\partial t}p(m,t) = -\frac{\partial}{\partial m}K^*(m,t)p(m,t) + \frac{1}{2}\frac{\partial^2}{\partial m^2}D(m)p(m,t) \tag{S13}$$

Here, we let $K^*(m,t) = K(m)+\beta(m,r)\delta(t-t_{\mathrm{DM}})$, with $\delta(\cdot)$ denoting the Dirac delta function. To numerically solve the equation, we used the finite difference method with $N_{\mathrm{disc}} = 96$ discretization points with $\Delta m = \pi/N_{\mathrm{disc}}$, tiling the orientation space.

$$\frac{\partial \mathbf{p}(t)}{\partial t} = \mathbf{L}^*(t)\cdot\mathbf{p}(t) \tag{S14}$$

where $\mathbf{p}(t)$ denotes a vector representing the densities $p(m,t)$ evaluated at the discretization points of m, and the transition matrix $\mathbf{L}^*(t)$ denotes the following:

$$\mathbf{L}^*(t) = -\mathbf{H}_{\mathrm{Drift}}\cdot\mathbf{K}^*(t) + \frac{1}{2}\cdot\mathbf{H}_{\mathrm{Diffusion}}\cdot\mathbf{D} \tag{S15}$$

where $\mathbf{K}^*(t)$ and $\mathbf{D}$ denote the diagonal matrices with the diagonal elements corresponding to the evaluation of $K^*(m,t)$ and $D(m)$ at the discretization points of m, and $\mathbf{H}_{\mathrm{Drift}}$ and $\mathbf{H}_{\mathrm{Diffusion}}$ denote the $N_{\mathrm{disc}}\times N_{\mathrm{disc}}$ dimensional finite difference matrices, anti-symmetric and symmetric, respectively, with the elements equal to zero except for the following:

$$(\mathbf{H}_{\mathrm{Drift}})_{j,j+1} = -(\mathbf{H}_{\mathrm{Drift}})_{j,j-1} = (2\Delta m)^{-1} \tag{S16}$$

$$(\mathbf{H}_{\mathrm{Diffusion}})_{j,j+1} = (\mathbf{H}_{\mathrm{Diffusion}})_{j,j-1} = -\frac{1}{2}(\mathbf{H}_{\mathrm{Diffusion}})_{j,j} = (\Delta m)^{-2} \tag{S17}$$

with the N_{disc}-periodic subscripts for the wrapping-around orientations. For convenience, we define the time-independent transition matrix $\mathbf{L}$ that accounts for every time point but t_{DM}:

$$\mathbf{L} = \lim_{t\to t_{\mathrm{DM}}}\mathbf{L}^*(t) \tag{S18}$$

To compute the full density of the memory process, using the efficient sensory encoding constrain, we first computed post-encoding distribution $\mathbf{p}_0$, which is a vector representing the densities $p(m_0|\theta)$ for the given θ, evaluated at the discretization points of m. Next, to evaluate the density of working memory at t_{DM}, we computed $\mathbf{p}_{\mathrm{DM}}$, a vector representing the densities $\lim_{t\uparrow t_{\mathrm{DM}}} p(m,t)$, as follows.

$$\mathbf{p}_{\mathrm{DM}} = e^{\mathbf{L}\cdot t_{\mathrm{DM}}}\cdot\mathbf{p}_0 \tag{S19}$$

To compute the conditional distributions after the decision-making, we defined a set of mask functions:

$$M^{\mathrm{CW}}(m,r) = \begin{cases}\Phi(m;r,\varsigma^2), & m\in[r-\pi/4, r+\pi/4)\\ 1-\Phi(m;r,\varsigma^2), & m\in[r+\pi/4, r+3\pi/4)\end{cases} \tag{S20}$$

$$M^{\mathrm{CCW}}(m,r) = 1 - M^{\mathrm{CW}}(m,r) \tag{S21}$$

where Φ denotes the Gaussian cumulative density function, and $\varsigma = 0.2$ captures the amount of the noise associated with the reference, which results in the smoothness of decision conditioning. Next, we applied the masks and decision-induced bias with strength w_β to $\mathbf{p}_{\mathrm{DM}}$:

$$\mathbf{p}^{\mathrm{CW}} = -w_\beta\cdot\mathbf{H}_{\mathrm{Drift}}\cdot\mathbf{M}_r^{\mathrm{CW}}\cdot\mathbf{p}_{\mathrm{DM}} \tag{S22}$$

$$\mathbf{p}^{\mathrm{CCW}} = w_\beta\cdot\mathbf{H}_{\mathrm{Drift}}\cdot\mathbf{M}_r^{\mathrm{CCW}}\cdot\mathbf{p}_{\mathrm{DM}} \tag{S23}$$

with $\mathbf{M}_r^{\mathrm{CW}}, \mathbf{M}_r^{\mathrm{CCW}}$ denote the diagonal matrices with the elements corresponding to $M^{\mathrm{CW}}(m,r), M^{\mathrm{CCW}}(m,r)$ evaluated at the discretization points of m. Additionally, we added the decision lapse rate parameter w_λ, resulting in a modified set of conditional distributions:

$$\mathbf{p}_\lambda^{\mathrm{CW}} = (1-w_\lambda)\cdot\mathbf{p}^{\mathrm{CW}} + w_\lambda\cdot\mathbf{p}^{\mathrm{CCW}} \tag{S24}$$

$$\mathbf{p}_\lambda^{\mathrm{CCW}} = w_\lambda\cdot\mathbf{p}^{\mathrm{CW}} + (1-w_\lambda)\cdot\mathbf{p}^{\mathrm{CCW}} \tag{S25}$$

Finally, we propagated these conditional distributions to t_{EM} and convolved them with the production error associated with the estimation report, obtaining the joint density of discrimination and estimation:

$$p\left(\hat{c}_j = \mathrm{CW}, \hat{\theta}_j | \theta_j, r_j, \vartheta_j\right) = e^{w_P^2/2 \cdot \mathbf{H}_{\mathrm{Diffusion}}} \cdot e^{\mathbf{L} \cdot (t_{\mathrm{EM}} - t_{\mathrm{DM}})} \cdot \mathbf{p}_\lambda^{\mathrm{CW}} \tag{S26}$$

$$p\left(\hat{c}_j = \mathrm{CCW}, \hat{\theta}_j | \theta_j, r_j, \vartheta_j\right) = e^{w_P^2/2 \cdot \mathbf{H}_{\mathrm{Diffusion}}} \cdot e^{\mathbf{L} \cdot (t_{\mathrm{EM}} - t_{\mathrm{DM}})} \cdot \mathbf{p}_\lambda^{\mathrm{CCW}} \tag{S27}$$

where w_P is the width of the production error distribution.

We fit our dynamical models to maximize the joint likelihood of the N_{trial} behavior data points, consisting of discrimination-estimation report pairs, for each individual:

$$p\left(\mathrm{Data}\right) = \prod_{j=1}^{N_{\mathrm{trial}}} p\left(\hat{c}_j, \hat{\theta}_j | \theta_j, r_j, \vartheta_j\right) \tag{S28}$$

where $\hat{c}_j, \hat{\theta}_j$ denote the participants' behavioral report of discrimination and estimation for trial j, respectively, and $\theta_j, r_j, t_{\mathrm{DM}}^j$ denote each trial j's stimulus and reference orientations, and decision-making timing condition, respectively. We fit the model to the data using the Nelder-Mead optimization implemented in scipy[3], with 20 iterations, with random initial conditions drawn from the constrained ranges of the model parameters.

2.4 Model parameters

The parameters to be fitted are specified as follows ($*w_K$ was set to 0 for the reduced model).

Symbol	Parameter	Constrained range
w_K	Drift rate for memory*	$[0, 15]$ $(^\circ/s)$
w_D	Diffusion rate for memory	$[0, 15]$ $(^\circ/s^{1/2})$
w_β	Decision-induced bias	$[-15, 15]$ $(^\circ)$
w_r	Reference bias weight	$[0, 15]$
w_E	Encoding variability	$[0, 15]$ $(^\circ)$
w_P	Production variability	$[0, 15]$ $(^\circ)$
w_λ	Decision-making lapse rate	[0, 0.15]
s	Encoding function shape parameter	[0, 1]

3 Recurrent neural network model

See the **Recurrent neural network models** section in the **Methods** for notation.

3.1 Task structure

For the train episode, in each trial, a fixation epoch of 0.1s with no inputs was followed by a stimulus epoch of 0.6s, the first-delay epoch of 0.3s, a DM epoch of 0.6s, a second-delay epoch of 0.3s, and an estimation report epoch of 0.1s, whose ending terminated a single trial. For the generalization episode, respecting the event block structure of the human task, we used the first-delay epoch of 1.8s for early DM conditions and 4.2s for late DM conditions, and a second-delay epoch of 4.2s for early DM conditions and 1.8s for late DM conditions, with other epoch lengths held the same.

For the inputs to the network, during the stimulus epoch, we used stimulus input vector $\mathbf{u}_\Theta(t)$ representing the orientation Θ, which consists of a ring of units, wherein the unit i has preferred orientation θ_i, was given as follows:

$$(\mathbf{u}_\Theta)_i = \gamma_\Theta \cdot \exp\left(\kappa_\Theta \left(\cos(\Theta - \theta_i) - 1\right)\right) \tag{S29}$$

where $\gamma_\Theta = 1$ denotes the strength of the stimulus and $\kappa_\Theta = 5$ the concentration parameter. On the other hand, with r and r_i denoting the reference orientation and unit i's preferred orientation in the reference

receiving population, reference input $\mathbf{u}_R(t)$ during the DM epoch was given as a one-hot vector:

$$(\mathbf{u}_R)_i = \gamma_R \cdot \mathbf{1}_{r=r_i} \tag{S30}$$

where $\gamma_R = 2$, respecting the aspect of our experiment that there was much higher 'certainty' for the visual orientation of reference than that of stimulus. Separate populations of $\mathbf{r} = [\mathbf{r}_\Theta; \mathbf{r}_R]$, namely $\mathbf{r}_\Theta$ and $\mathbf{r}_R$, received $\mathbf{u}_\Theta$ and $\mathbf{u}_R$, respectively, since in our paradigm, the spatial configuration of reference and stimulus was designed to drive the distinct populations with distinct spatial 'receptive fields'. For simplicity, we assumed each of $\mathbf{r}_\Theta$ and $\mathbf{r}_R$ to be 48-dimensional.

In line with our experiment, we used $N_\Theta = 24$ discretized stimuli ranging from $0°$ to $172.5°$ by the discretization size of $7.5°$. Also, we constrained the range of reference to the proximity of stimulus, $|r - \theta| \leq 30°$ by the discretization size of $7.5°$, resulting in 9 possible cases of relative reference. We excluded the case where the reference orientation was the same with the stimulus ($r = \theta$) during the train episode to facilitate training but included the case during the generalization episode.

3.2 Training procedure

Using the procedure described above, we trained recurrent connection weight $\mathbf{J}$. To facilitate the interpretation of emergent recurrent computations, the input and output connection weights, except for DM output weight, were fixed as cosine circulant matrices:

$$\mathbf{J}_\Theta(i,j) = \mathbf{J}_R(i,j) = \gamma_I \cdot \cos\left(2\pi(i-j)/N_\Theta\right) \tag{S31}$$

$$\mathbf{J}_{\mathrm{EM}}(i,j) = \gamma_{\mathrm{EM}} \cdot \cos\left(2\pi(i-j)/N_\Theta\right) \tag{S32}$$

where $\gamma_I = 1$ and $\gamma_{\mathrm{EM}} = 0.4$. Since both $\mathbf{r}_\Theta$ and $\mathbf{r}_R$ were 48-dimensional, we used $\mathbf{J}_\Theta$ and $\mathbf{J}_R$ as $N_{\mathrm{rec}} \times 48$. $\mathbf{J}_{\mathrm{EM}}$ was $24 \times N_{\mathrm{rec}}$, wherein 24 was the length of the response vector $\mathbf{z}^{\mathrm{EM}}$. $\mathbf{J}^{\mathrm{DM}}$ was set as a $2 \times N_{\mathrm{rec}}$ zero-or-one voting matrix:

$$\mathbf{J}_{\mathrm{DM}}(i,j) = \mathbf{1}_{i \in \mathcal{I}_{\mathrm{CW}}} \cdot \mathbf{1}_{j=1} + \mathbf{1}_{i \in \mathcal{I}_{\mathrm{CCW}}} \cdot \mathbf{1}_{j=2} \tag{S33}$$

where we chose the voting to be balanced, to be specific, $\mathcal{I}_{\mathrm{CW}} = \{1, \cdots, 24\} \cup \{49, \cdots, 72\}$ and $\mathcal{I}_{\mathrm{CW}} = \{25, \cdots, 48\} \cup \{73, \cdots, 96\}$.

For each network, before training, we initialized the recurrent connection weight $\mathbf{J}$ as zero. The joint loss $\mathcal{L}$ was computed by time-averaging the cross entropies between the network output and the "ground-truth" target output.

$$\mathcal{L}_{\mathrm{DM}} = \left\langle \sum_{j \in \{\mathrm{CW},\mathrm{CCW}\}} M^{\mathrm{DM}}(t) \cdot z_j^{\mathrm{DM}}(t) \cdot \log\left(1/q_j^{\mathrm{DM}}(t)\right) \right\rangle_t \tag{S34}$$

$$\mathcal{L}_{\mathrm{EM}} = \left\langle \sum_{j \in \{1,\cdots,N_{\mathrm{in}}\}} M^{\mathrm{EM}}(t) \cdot z_j^{\mathrm{EM}}(t) \cdot \log\left(1/q_j^{\mathrm{EM}}(t)\right) \right\rangle_t \tag{S35}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{DM}} + \mathcal{L}_{\mathrm{EM}} \tag{S36}$$

where the semantic notation $j \in \{\mathrm{CW}, \mathrm{CCW}\}$ corresponds to $j \in \{1, 2\}$. The term M_t^{DM} was a zero-or-one mask for the loss evaluation that was non-zero only during a decision period at the end of each trial, and M_t^{EM} was non-zero except for the initial fixation epoch. The loss was minimized by backpropagation using Tensorflow[4] using Adam optimizer with a learning rate of 0.02. We iterated 300 times per network training. For each iteration, 128 trials (*i.e.*, inputs-desired output pairs) were randomly generated. Each trial included a random combination of stimulus orientation and relative reference orientation.

For the behavioral and "neural" analyses of trained RNNs, we trained 50 independent RNNs, all of which were included in the analyses. To inspect the stimulus-specific and decision-consistent biases displayed by the trained RNNs in the early and late DM conditions in the generalization episodes, we utilized the output vectors $\mathbf{z}^{\mathrm{DM}}, \mathbf{z}^{\mathrm{EM}}$. For the DM and EM readouts, $\hat{c}^{\mathrm{RNN}}, \hat{\theta}^{\mathrm{RNN}}$, we used the following:

$$\hat{\theta}^{\mathrm{RNN}} = \arctan\left(\sum_{j=1}^{24} z_j^{\mathrm{EM}} \sin 2\theta_j \Big/ \sum_{j=1}^{24} z_j^{\mathrm{EM}} \cos 2\theta_j \right) \tag{S37}$$

$$\hat{c}^{\mathrm{RNN}} = \mathrm{sign}\left(z_1^{\mathrm{DM}} - z_2^{\mathrm{DM}}\right) \tag{S38}$$

Analyses on the decision and estimation readouts, $\hat{c}^{\mathrm{RNN}}, \hat{\theta}^{\mathrm{RNN}}$, were the same as those on the human discrimination and estimation task behaviors $\hat{c}, \hat{\theta}$.

3.3 State space analysis

We trained independent, 50 "homogeneous" RNNs, assuming the veridical encoding of stimulus input $\mathbf{u}_\Theta$, *i.e.*, $\Theta|\theta = \theta$, along with our original, "heterogeneous" RNNs. All the aforementioned details in training the networks, other than the input-level encoding, were held the same across homogeneous and heterogeneous RNNs.

To study the mechanism underlying the bias dynamics of RNN population activities during the decision-making epoch, we used PCA in capturing the low dimensional behavior of RNN. To capture the dynamics shared across the trained individual networks in the state space analyses, we computed the "mean" model by taking the mean of individually trained $\mathbf{J}$, one model for each of the trained homogeneous and heterogeneous RNNs, respectively. To isolate the relationship between drift dynamics and decision-induced biases, we generated the trials from homogeneous and heterogeneous RNNs, without network noise (ξ) or input-level variability (*i.e.*, $\Theta|\theta = \theta$).

To inspect the functional properties of the sub-populations in RNN, we further separated the recurrent vector $\mathbf{r}$ based on their distinct output projection profiles for the DM output, as we used the zero-one voting mapping $\mathbf{J}_{\mathrm{DM}}$. That is, we separated $\mathbf{r}^{\mathrm{CW}}$, for the CW-projecting population, and $\mathbf{r}^{\mathrm{CCW}}$, for the CCW-projecting population. Since we already used the separation based on input projection profiles, namely, $\mathbf{r}_\Theta$ and $\mathbf{r}_R$, we thus use the combined notation (*e.g.*, $\mathbf{r}_R^{\mathrm{CW}}$ denotes the R-receiving and CW-projecting population). For the 'winning' and 'losing' populations, we chose a near-reference case of $\theta - r = 7.5°$. Since we did not assume noise for the state space analysis, the winning population was $\mathbf{r}^{\mathrm{CW}}$, and the losing population was $\mathbf{r}^{\mathrm{CCW}}$.

Since we aimed to inspect the RNN dynamics on the common ground, we chose to project the population activities of mean homogeneous and heterogeneous RNNs onto the shared subspace spanned by $\mathbf{r}_\Theta$ of the mean homogeneous RNNs. We first stacked the activities of $\mathbf{r}_\Theta = (\mathbf{r}_\Theta^{\mathrm{CW}} + \mathbf{r}_\Theta^{\mathrm{CCW}})/2$ of the mean homogeneous RNNs for each 'condition' (the combination of $N_\Theta = 24$ stimulus orientations and $N_R = 9$ relative reference orientations) to form a column-mean-centered data matrix $\mathfrak{D}$ shaped $(N_\Theta \cdot N_R \cdot T) \times 24$, where T denotes the number of the total time steps of RNN outputs. We then constructed the PC projection matrix $\mathbf{V}_{\mathfrak{D}}$ based on the singular value decomposition of the data matrix $\mathfrak{D}$.

$$\mathfrak{D} = \mathbf{U}_{\mathfrak{D}} \mathbf{S}_{\mathfrak{D}} \mathbf{V}_{\mathfrak{D}}^\top \tag{S39}$$

Next, we projected the population activities of $\mathbf{r}_R^{\mathrm{CW}}$, $\mathbf{r}_R^{\mathrm{CCW}}$, and $\mathbf{r}_\Theta^*$ of each of the mean homogeneous and heterogeneous RNNs onto the two-dimensional subspace constructed with $\mathfrak{D}$, with the projection matrix $\mathbf{V}_{\mathfrak{D}}$. This enabled us to unify the subspace for the two RNNs we compared.

To analyze the state space, we fitted the following origin-centered ellipse to the starting points of each of the population activities projected onto the two-dimensional subspace,

$$\frac{(x \cos E_\theta + y \sin E_\theta)^2}{E_a^2} + \frac{(x \sin E_\theta - y \cos E_\theta)^2}{E_b^2} = 1 \tag{S40}$$

where E_a and E_b are radii and E_θ is the rotation angle of the ellipse. To draw arrows, we considered the vector from the starting point to the ending point of each of the two-dimensional activities during the reference presentation and projected it onto the tangent line at the starting point, followed by the scaling by 0.2 for $\mathbf{r}_R^{\mathrm{CW}}$, $\mathbf{r}_R^{\mathrm{CCW}}$ and 0.5 for $\mathbf{r}_\Theta$.

4 Analysis of near-reference variability

4.1 Definition of near-reference variability

In a similar spirit to the previous study that characterized near-reference variability as a signature of decision-induced bias[5], we computed variability of the distribution of estimation task error, $\epsilon = \hat{\theta} - \theta$, for each of the relative reference conditions. As a robust measure of the variability of estimation error distributions, we used the interquartile range (IQR).

4.2 Analysis of different datasets

For across-data set comparison, we used publicly shared data of Zamboni et al.[6] and Fritsche & de Lange[7] for the "reference repulsion" tasks (*i.e.*, when the intervening reference serves as a boundary for decisions), and Rademaker et al.[8] for the "reference attraction" tasks (*i.e.*, when the intervening reference serves as a distractor). Across the data sets, we split the data into 7 bins based on (stimulus - reference), and calculated IQR for each bin. For Rademaker et al.[8], wherein the authors used an orthogonal set of stimulus and distractor orientations, we constrained the range of interest of (stimulus - reference) within $[-25°, 25°]$. Relative reference was considered "Near" when (stimulus - reference) is within $[-8°, 8°]$, across the data sets.

Supplementary References

[1] Wei, X.-X. & Stocker, A. A. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience* **18**, 1509–1517 (2015).

[2] Wei, X.-X. & Stocker, A. A. Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences* **114**, 10244–10249 (2017).

[3] Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods* **17**, 261–272 (2020).

[4] Abadi, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[5] Luu, L. & Stocker, A. A. Post-decision biases reveal a self-consistency principle in perceptual inference. *Elife* **7**, e33334 (2018).

[6] Zamboni, E., Ledgeway, T., McGraw, P. V. & Schluppeck, D. Do perceptual biases emerge early or late in visual processing? decision-biases in motion perception. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20160263 (2016).

[7] Fritsche, M. & de Lange, F. P. Reference repulsion is not a perceptual illusion. *Cognition* **184**, 107–118 (2019).

[8] Rademaker, R. L., Chunharas, C. & Serences, J. T. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature neuroscience* **22**, 1336–1344 (2019).