Sensory readout accounts for adaptation

Timothy C. Sheehan[1], John T. Serences[1,2,3]

[1]Neurosciences Graduate Program, [2]Department of Psychology, and [3]Kavli Institute for Brain and Mind, University of California, San Diego, California 92093

Abstract

Sensory responses and behavior are strongly shaped by stimulus history. For instance, perceptual reports are sometimes biased towards previously viewed stimuli (*serial dependence*). Previous behavioral studies suggest that serial dependence is implemented via modulations in visual cortex, but neural evidence is lacking. We recorded fMRI responses while human participants performed a delayed orientation discrimination task. While behavioral reports were *attracted* to the previous stimulus, response patterns in sensory areas were *repelled*. We reconciled these opposing biases using a model where both sensory encoding and readout are shaped by stimulus history. Neural adaptation reduces redundancy at encoding and leads to the repulsive biases that we observed in visual cortex. Serial dependence is not implemented in visual cortex but rather by readout mechanisms that account for adaptation during encoding. The model suggests the visual system improves efficiency via adaptation while still optimizing behavioral readout based on the temporal structure of natural stimuli.

Keywords: neural adaptation, sensory decoding, Bayesian inference, serial dependence

Introduction

Natural stimuli are known to demonstrate statistical dependencies across both space and time, such as a prevalence of vertical and horizontal (cardinal) orientations and a higher probability of small orientation changes in sequential stimuli[1–3]. These regularities can be leveraged to improve the efficiency and accuracy of visual information processing. We use the term “encoding” to refer to the initial conversion of external sensory information into neural activity patterns and the term “readout” to refer to the readout of these encoded signals to shape behavior. At encoding, regularities yield attenuated neural responses to frequently occurring stimuli (“adaptation”), reducing metabolic cost and redundancy in neural codes. At readout, regularities support the formation of Bayesian priors that can be used to bias perception in favor of higher probability stimuli. On their own, both adaptation and Bayesian readout can explain a variety of behavioral phenomena such as improved precision around the cardinal axes (oblique effect) and why we remember objects as being closer to the average exemplar (contraction bias) [4–6]. While the effects of sensory history on sensory coding and behavior have been studied extensively, it is unclear how changes at encoding shape readout and behavior.

Adaptation increases coding efficiency by modulating sensory tuning properties as a function of the recent past. For instance, reducing the gain of neurons tuned to a recently seen adapting stimulus reduces the temporal autocorrelation of activity when similar stimuli are presented sequentially. In turn, reducing these autocorrelations improves the overall efficiency of sensory codes: fewer spikes are dedicated to encoding redundant stimuli, and the presence of a novel stimulus can be more easily detected as it will be accompanied by a sudden increase in activity [2,7–15]. Importantly, adapted representations early in the processing stream (e.g. in LGN) are inherited by later visual areas [11,16,17].

Although adaptation increases coding efficiency, it comes at a cost to perceptual fidelity as adaptation can lead to repulsion away from the adapting stimulus for features such as orientation and motion direction [18–20]. For instance, after continuously viewing and adapting to motion in one direction, stationary objects will appear to be moving in the opposite direction (i.e., current perceptual representations are *repelled* away from recent percepts). However, this potentially deleterious aftereffect is accompanied by better discriminability around the adapting stimulus, which may be more important than absolute fidelity from a fitness perspective [14,21–24].

In contrast to the repulsive biases associated with neural adaptation, perception is sometimes attracted to recently attended items in conditions where weak stimuli are attended – a phenomenon termed "serial dependence". As serial dependence can impact immediate perceptual reports and the relative perception of simultaneously presented items, some have suggested that it reflects modulations in early stages of sensory processing [25–27]. In line with this idea, one fMRI study demonstrated that early sensory biases match 'attractive' behavioral reports [28]. However, consecutive stimuli were always either the same or orthogonal orientations, conditions where serial dependence effects on behavior are negligible [26,29,30]. Thus, without sampling the entire stimulus feature space, it is unclear how to integrate this finding with related empirical and theoretical work on serial dependence. Counter to studies suggesting a sensory locus of serial dependence, other behavioral results have found that serial dependence does not occur immediately after encoding but instead emerges only, and increases with, a working memory maintenance period [31–33]. This observation suggests that serial dependence is not the product of early sensory coding [32] and instead might be implemented by a later readout or memory maintenance circuit [34,35]. There is evidence that such a readout mechanism is Bayesian, as the influence of the "prior" (the previous stimulus) is larger when sensory representations are less precise due to either external or internal noise [27,35]. Thus, the collective evidence is mixed, with some studies pointing towards an early sensory locus and others to later stages of readout and memory storage.

This lack of consensus suggests that assessing interactions between sensory and readout stages of processing may be key to better understanding the impact of stimulus history on perception. For example, previous work suggests that readout does not account for neural adaptation that happens during encoding, as adaptive repulsive biases cascade across layers of the visual processing hierarchy and penetrate behavioral reports [11,17,36]. These studies, however, did not consider paradigms where the adapting stimulus was behaviorally relevant. Attending to relevant stimuli may shape how readout stages account for the current state of adaptation, possibly inducing attractive serial dependence. To assess this possibility, we utilized multivariate fMRI decoding techniques to characterize how representations in early visual areas change as a function of stimulus history during a delayed orientation discrimination task (Figure 1A). We replicated classic "serial dependence" findings where behavioral reports were attracted to the orientation of the previous stimulus. We found that this attractive behavioral bias was not accompanied by attractive biases in visual cortex, as predicted by early sensory models of serial dependence. Rather, we observed *repulsive biases* in early visual cortex, consistent with adaptation. To explain these results, we examined several possible read-out mechanisms and found that only decoding schemes that account for adaptation can explain attractive serial dependence in behavior. More generally, these results explain how the visual system can reduce energy usage without sacrificing precision by optimizing encoding and behavioral readout relative to the temporal structure of natural environments.

Results

Behavior

To probe the behavioral effects of serial dependence, we designed a delayed discrimination task where participants judged whether a bar was tilted CW or CCW relative to the orientation of a remembered grating (Figure 1A). We first report the results from a behavior-only study (n=47) followed by an analysis of neural activity for a cohort completing the same task in the scanner (n=6). Task difficulty was adjusted for each participant by changing the magnitude of the probe offset ($\delta\theta$) from the remembered grating and was titrated to achieve a mean accuracy of ~70% (accuracy 69.8±0.82%, $\delta\theta$: 4.95±0.27°; all reported values mean ±1SEM unless otherwise noted). Fixing subjects at this intermediate accuracy level helped to avoid floor/ceiling effects and improved our sensitivity to detect perceptual biases while keeping participants motivated.

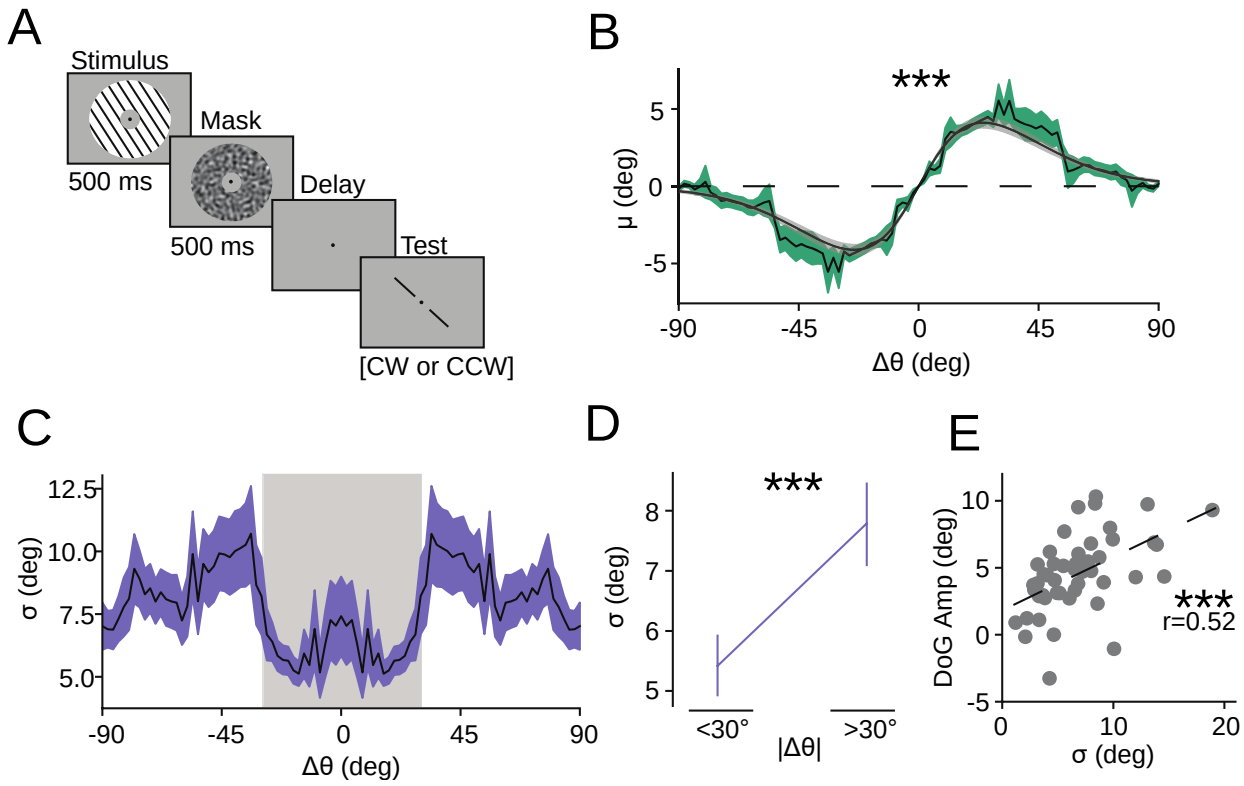


Figure 1 Behavior. **A**: Task Schematic. An orientated stimulus is followed by a probe bar that is rotated <15° from the stimulus. Participants judged whether the bar was CW or CCW relative to the stimulus in a binary discrimination task. **B**: Behavioral bias, green: average model-estimated bias as a function of $\Delta\theta=\theta_{n-1}-\theta_n$ (± SEM across participants); gray: average DoG fit to raw participant responses sorted by Δθ (± 1SEM across participants). **C**: Behavioral σ as a function of Δθ, shaded region corresponds to |Δθ|<30°. **D**: Behavioral variance is significantly less for |Δθ|<30°. **E:** Bias is correlated with variance across participants. ***, p<.001

To quantify the pattern of behavioral responses, we modelled the data as the product of a noisy encoding process described by a Gaussian function centered on the presented orientation with standard deviation σ and bias μ. Optimal values for σ and μ were found by maximizing the likelihood of responses for probes of varying rotational offsets from the remembered stimulus, thus converting pooled binary responses into variance and bias measured in degrees (see Response Bias, Figure S1). This allowed us to measure precision for individual participants and also allowed us to measure how responses were

biased as a function of the orientation difference between the remembered gratings on consecutive trials $\Delta\theta = \theta_{n-1} - \theta_n$, an assay of serial dependence. To increase power and remove systematic response biases, we 'folded' responses relative to Δθ such that all analysis have either rotational (Figure 1B) or horizontal (Figure 1C) symmetry [34].

Responses were clearly biased towards the previous stimulus (Figure 1B, green curve), which we quantified by fitting a Derivative-of-Gaussian (DoG) function to the raw response data for each participant (gray curve; amplitude: 4.53°±0.42°, $t(46) = 7.8$, $p = 5.9*10^{-10}$, one sample t-test; full width at half max (FWHM): 42.9°±1.8°, see Serial Dependence). The magnitude and shape of serial dependence is consistent with previous reports [26,29]. We next examined how response precision (σ) varied as a function of Δθ and found that responses were more precise around small changes (Figure 1C), again consistent with previous reports [37]. We quantified this difference in precision by splitting trials into 'close' and 'far' bins (greater than or less than 30° separation) and confirmed that responses following 'close' stimuli were more precise ($t(46)=-3.72$, $p=0.0003$, paired 1-tailed t-test, Figure 1D , see Response Precision). Note that the choice of 30° was arbitrary and all threshold values between 20° and 40° yielded a significant ($p<.05$) result.

Previous work has shown that serial dependence is greater when stimulus contrast is lower [27] and when internal representations of orientation are weaker due to stimulus independent fluctuations in encoding fidelity [35]. We tested a Bayesian interpretation of these findings by asking whether less precise individuals are more reliant on prior expectations and therefore more biased. Indeed, we found a positive correlation between DoG amplitude and σ, (Figure 1E, $r(45)=0.52$, $p=.0001$, 1-tailed Pearson's correlation).

Finally, we confirmed that our behavioral results were not driven by an artifact of our modeling procedure by replicating the relationships using raw response probabilities (Figure S2). We also confirmed these effects were not driven by a subset of low/high performing individuals or trial counts (Figure S3) and were not due to inhomogeneities in the stimulus sequences used for some participants (Figure S4).

Stimulus history effects in visual cortex

To examine the influence of stimulus history on orientation-selective response patterns in early visual cortex, six participants completed between 748 and 884 trials (mean 838.7) of the task in the fMRI scanner over the course of four, two-hour sessions (average accuracy of 67.7% ±0.4% with an average probe offset, $\delta\theta$, of 3.65°). As with the behavior-only cohort, these participants showed strong attractive serial dependence (Figure 2A, green) that was significantly greater than 0 when parameterized with a DoG function (amplitude=3.50°±0.27°, $t(5)=11.93$, $p=.00004$; FWHM=35.9°±2.34°, Figure 2A black dotted line). This bias was not significantly modulated by inter-trial interval, delay period, or an interaction between the two factors (all p-values > 0.5, mixed linear model grouping by participant). Similar to the behavioral cohort, we found that variance was generally lower around small values of Δθ. We quantified variance in the same manner as the behavioral cohort (flipping responses to match biases and down-sampling the larger group) and found that responses were more precise following close (<30°) relative to far stimuli (>30°, $t(5) = -9.96$, $p=0.00009$, 1-tailed paired t-test, Figure 2B). This pattern was significant ($p<0.05$) for thresholds between 20° and 40°, and these findings were generally replicated when

analyzed using a model-free approach as described for the behavior-only cohort (Figure S2D-F, except that the threshold analysis no longer reached significance).

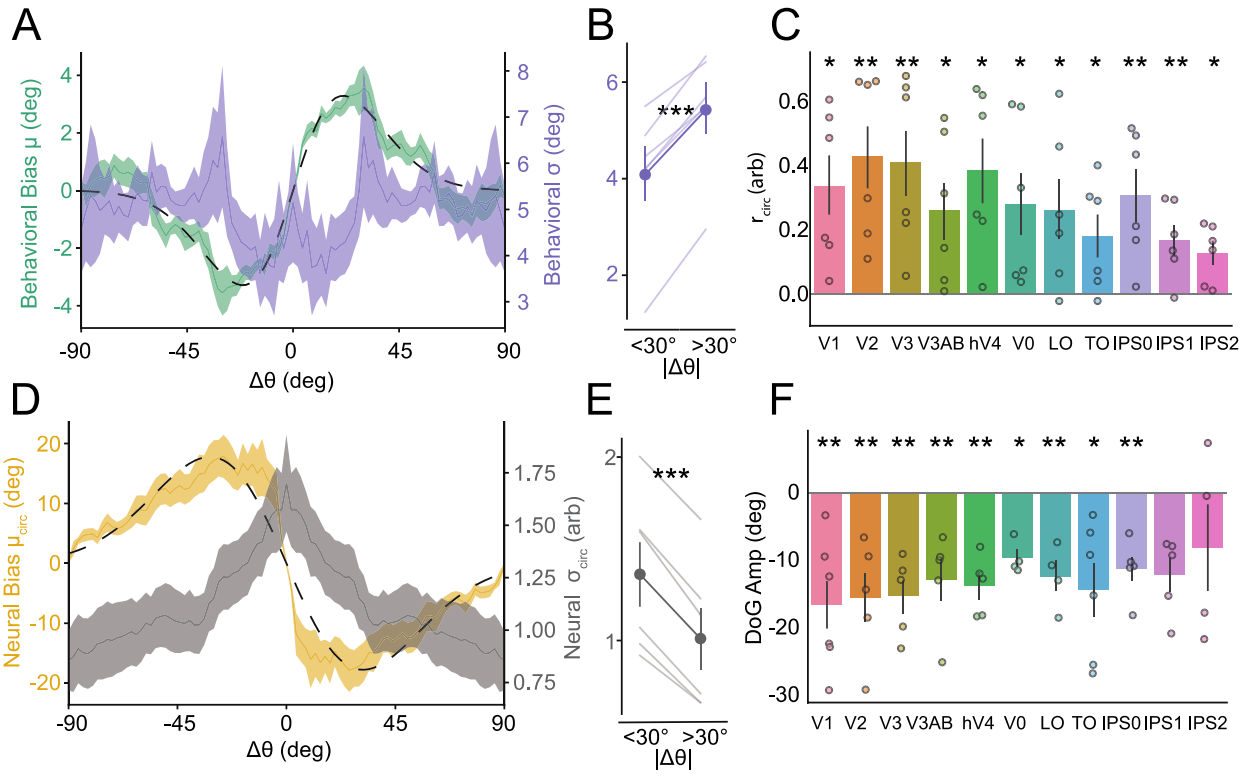


Figure 2 *Behavioral and Neural Bias.* ***A****: Left-axis, Behavioral serial dependence. Shaded green: average* model-estimated bias as a function of Δθ (± SEM across participants); dotted black line: average DoG fit to raw participant responses sorted by Δθ. Right-axis, variance. Purple shaded line: model-estimated variance as a function of Δθ (± SEM across participants). **B**: Behavioral *σ* is significantly less for |Δθ|<30°. **C**: Decoded orientation was significantly greater than chance when indexed with circular correlation for all ROIs examined. Error bars indicate ±SEM across participants. Dots show data from individual participants. **D**: Left-axis, decoding bias. Shaded yellow line: decoded bias (μ_{circ} of decoding errors) sorted by Δθ (± SEM across participants); dotted black line: average DoG fit to raw decoding errors sorted by Δθ. Right-axis, decoded σ_{circ} . Shaded gray line: average decoding variance (σ_{circ}) as a function of Δθ (± SEM across participants). Note that σ_{circ} can range from [0, inf] and has no units. **E:** Decoded variance is significantly less for |Δθ|<30° **F:** Decoded errors are significantly repulsive when parameterized with a DoG for most ROIs. *, p<.05; **, p<.01; ***, p<.001.

To characterize activity in early visual areas, independent retinotopic mapping runs were completed by each subject to identify regions of interest (ROIs) consisting of: V1, V2, V3, V3AB, hV4, Ventral-, Temporal- and Lateral-Occipital Areas (VO, TO, and LO), and intraparietal sulcus areas IPS0-2. In addition, a separate localizer task was used to sub-select the voxels that were most selective for the spatial position and orientation of the stimuli used in our task (see Voxel Selection).

To examine how visual representations are affected by stimulus history, we trained a decoder on the orientation of the sample stimulus on each trial based on BOLD activation patterns in each ROI. We used a decoder that accounts for stimulus related noise correlations (see Orientation Decoding, [35,38]) using a leave-one-run-out cross-validation across sets of 68 consecutive trials (4 blocks of 17 trials) that had orientations pseudo randomly distributed across all 180° of orientation space. We first quantified

single-trial decoding performance using circular correlation (r_{circ}) between the decoder-estimated orientations and the actual presented orientations and found that all ROIs had significant orientation information (Figure 2C). For this and all of the remaining main figures, we used the average of four TRs (spanning 4.8-8.0s) following stimulus presentation to avoid the influence of the probe stimulus (which came up ≥6s into the trial and thus should not influence responses in the 4.8-8.0s window after accounting for hemodynamic delay). That said, our ability to decode orientation was not specific to the exact TRs selected, or whether the decoder was trained on the task data or an independent localizer task (Figures S5-S6).

The high SNR of the BOLD decoder allowed us to examine residual errors on individual trials. When measuring the bias of these decoding errors (μ_{circ}) as a function of stimulus history (Δθ), we unexpectedly observed a strong *repulsive* bias reflecting neural adaptation (V3, Figure 2G yellow, see Neural Bias). This bias was significant when quantified with a DoG (amplitude=-19.76°±3.06°, $t(5)$=-5.90, p=.0010; FWHM=47.7°±1.71°, Figure 2G black dotted-line). All ROIs except IPS0 and IPS1 had a significantly negative amplitude (p<.05) and the average DoG amplitude across ROIs was also significant ($t(10)$=-7.65, p=.00001, Figure 2F). This repulsive pattern suggests that serial dependence is not a direct result of biases in early sensory areas. Importantly, representations in early visual areas (V1-V3AB) showed a repulsive bias for all participants regardless of the specific decoding technique used and when the decoder was instead trained on an independent orientation localizer (Figure S7-8). This suggests that the repulsive bias is also found in the "sensory" code and is not specific to working memory maintenance. This repulsive pattern held throughout the duration of the trial, suggesting it was not a transient phenomenon (Figure S7A). In accordance with the large effects of the previous stimulus on current trial representations, we observed above chance decoding for the identity of the previous stimulus in 9/11 ROIs using the same TRs and decoding techniques as used for the current stimulus (Figure S9).

We also examined how the precision of neural representations changed as a function of stimulus history. In sharp contrast to behavior, σ_{circ} exhibited a monotonic trend such that neural decoding was *least* precise when the previous stimulus was similar (Figure 2D, gray curve, see Neural Variance). We quantified this difference in sensory uncertainty in a similar manner to the behavioral data and found that variance in the sensory representations was significantly greater following a similar stimulus (<30°, $t(5)$=13.33, p=.00002, paired 1-tailed t-test, V3, Figure 2E). This pattern was significant (p<.01) in 8/11 ROIs and the difference in precision was significant across ROIs ($t(10)$=6.92, p=.00002, Figure S10A-B). The results did not change qualitatively when we utilized decoded uncertainty derived directly from the posterior rather than the circular standard deviation of decoded responses [38] (Figure S10C-D), or when we used other thresholds between 20° and 40°. The repulsion of sensory representations and the corresponding reduction in decoding precision around the previous orientation is consistent with neural adaptation where recently active units are attenuated, thus leading to lower SNR responses in visual cortex.

Encoder-Decoder Model

We observed an attractive bias and low variability around the current stimulus feature in behavior, and a repulsive bias and high variability around the current feature in the fMRI decoding data.

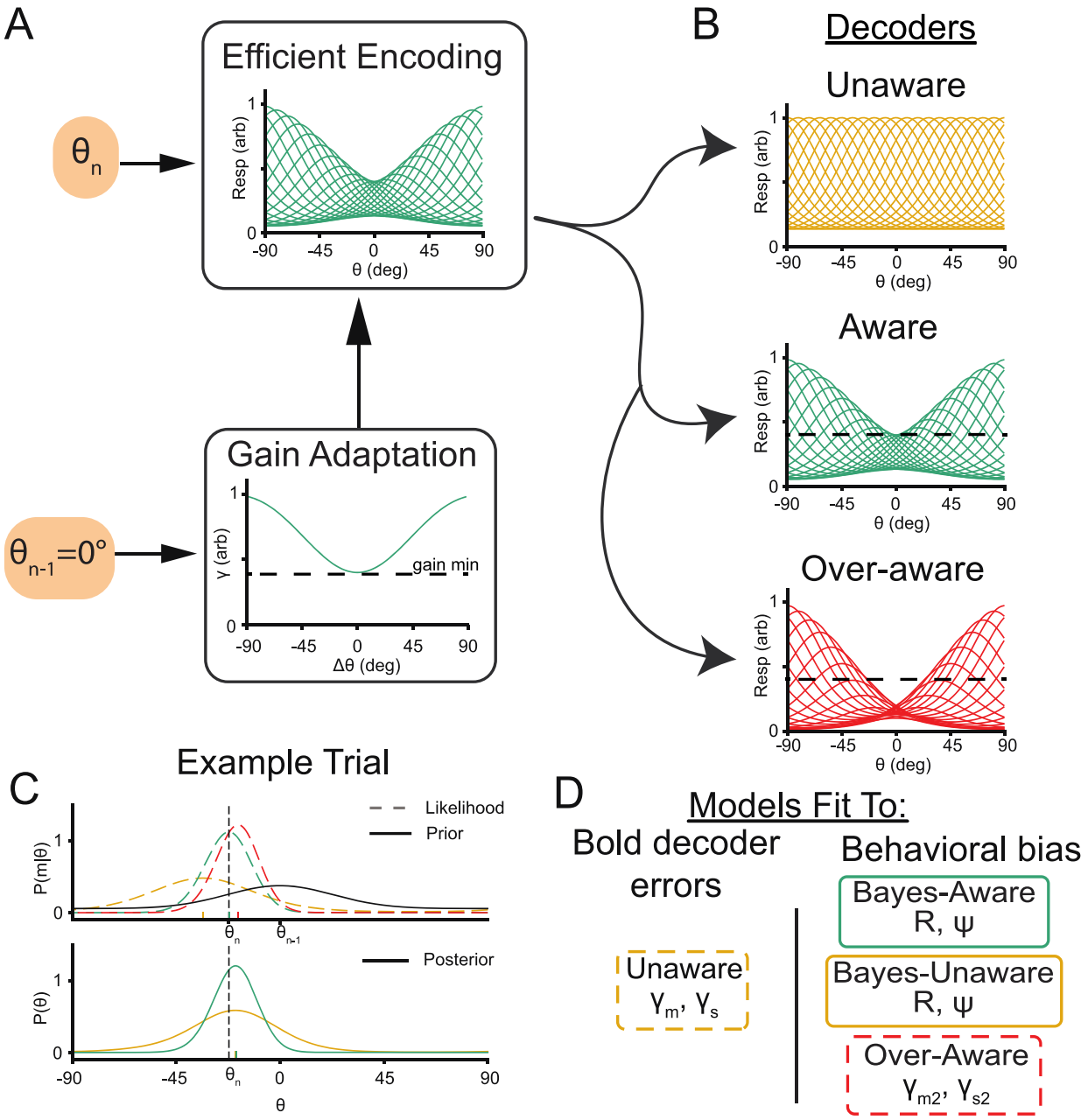


Figure 3 Encoder-Decoder model schematic. **A**: Encoding. Units with von Mises tuning curves encodes incoming stimuli. The gain of individual units undergoes adaptation such that their activity is reduced as a function of their distance from the previous stimulus. **B**: Decoding. This activity is then read out using a scheme that assumes one of three adaptation profiles. The unaware decoder assumes no adaptation has taken place, the aware decoder assumes the true amount of adaptation while the over-aware decoder over-estimates the amount of adaptation (note center tuning curves dip lower than the minimum gain line from encoding). **C**: Example stimulus decoding. Top: The resulting likelihood function for the unaware readout (dotted yellow line) has its representation for the current trial (θ_n=-30°) biased away from the previous stimulus (θ_{n-1}=0°). The aware readout (dotted green line) is not biased, while the over-aware readout is biased towards the previous stimulus. These likelihood functions can be multiplied by a prior of stimulus contiguity (solid black line) to get a Bayesian posterior (bottom) where Bayes-unaware and Bayes-aware representations are shifted towards the previous stimulus. Tick marks indicate maximum likelihood or decoded orientation. **D**: Summary of models and free parameters being fit to both BOLD decoder errors and behavioral bias.

187

Thus, the patterns of bias and variability observed in the behavioral data are opposite to the patterns of bias and variability observed in visual cortex. To better understand these opposing effects, we reasoned that representations in early visual cortex do not directly drive behavior but instead are read out by later cortical regions that determine the correct response given the task [39–41]. In this construction, the decoded orientations from visual cortex represent only the beginning of a complex information processing stream that, in our task, culminates with the participant making a speeded button press response. Thus, we devised a two-stage encoder-decoder model to describe observations in both early visual cortex and in behavior (see modeling).

The encoding stage of the model consists of a simulated population of orientation-selective neurons with von Mises tuning curves evenly tiling the feature space. The gain of these tuning curves undergoes adaptation such that units tuned to the previous stimulus (θ_{n-1}) will have their activity reduced on the current trial (Figure 3A). Note that these neurons are assumed to have Poisson firing rates and that their responses are noiseless while training the model.

The decoding stage reads out this activity using one of three strategies (Figure 3B). The *unaware* decoder assumes no adaptation has taken place and results in stimulus likelihoods $p(m|\theta)$ that are repelled from the previous stimulus (Figure 3C, yellow). This adaptation-naïve decoder is a previously hypothesized mechanism for behavioral adaptation [36] and likely what gives rise to the repulsive bias we observe in visual cortex using a fMRI decoder that is agnostic to stimulus history (Figure 2D).

	Fit To:	BOLD Decoder	Behavior		
Stage:		**Unaware**	**Bayes unaware (Prior*unaware)**	**Bayes aware (Prior*aware)**	**Over-aware**
Encoding	γ_m	**X**			
	γ_s	**X**			
Decoding	γ_{m2}	0	0	γ_m	**X**
	γ_{s2}	1	1	γ_s	**X**
Bayes	R	5	**X**	**X**	5
	ψ	N/A	**X**	**X**	N/A

*Table 1 Cells correspond to parameters for proposed decoders, with '**X**' indicating free parameters adjusted to fit empirical data. γ_m controls the amplitude and γ_s controls the width of gain adaptation (Figure 3A). These parameters are fit by minimizing the residual sum of squared errors between the unaware decoder and the BOLD decoder output. γ_{m2} and γ_{s2} are the assumed adaptation parameters at decoding. These terms are either set to assume no adaptation (unaware), match the true amount of adaptation (aware) or are free parameters adjusted to maximize the likelihood of responses (over-aware, Figure 3B). Last, R adjusts the average Poisson firing rate and ψ controls the variance of the prior distribution (Figure 3C). These parameters are adjusted for decoders using a Bayesian prior while R is set to the arbitrary value of 5 for non-Bayesian decoders (it has no effect on bias for non-Bayesian decoders). Increasing R increases the precision of the likelihood function and reduces the relative influence of the prior. Increasing ψ increases the range of Δθ over which the prior has an influence.*

Alternatively, the *aware* decoder (Figure 3C, green) has perfect knowledge of the current state of adaptation and can thus account for and 'un-do' biases introduced during encoding. Finally, the *over-aware* decoder knows the identity of the previous stimulus but over-estimates the amount of gain modulation that takes place, resulting in a net attraction to the previous stimulus (Figure 3C, red). We additionally combined a formal prior based on temporal contiguity with the stimulus likelihood from the previously described decoders [35]. In our implementation, a Bayesian prior centered on the previous stimulus (Figure 3C, black) is multiplied by the decoded likelihood to get a Bayesian posterior (Figure 3C, bottom). We applied this prior of temporal contiguity to both the *aware* decoder as well as the *unaware* decoder to test the importance of awareness at decoding. We did not apply a prior to the *over-aware* model to balance the number of free parameters between the various decoders and to see if the *over-aware* model could achieve attractive serial dependence without a Bayesian prior (Table 1). In total we explored three separate decoder models: the *Bayes-unaware* and *Bayes-aware* models which apply a prior to their respective likelihoods as well as the *over-aware* model which outputs the maximum likelihood (Figure 3D).

For each participant, we fit the encoder-decoder model in two steps (Figure 3D). All model fitting was performed using the same cross-validation groups as our BOLD decoder and each stage had two free parameters that were fit using grid-search and gradient descent techniques. We first report results from the encoding stage of the model. The gain applied at encoding was adjusted to minimize the residual sum of squared errors (RSS) between the output of the *unaware* decoder and the residual errors of our BOLD decoder. The *unaware* readout of the adapted encoding process (Figure 4A, yellow) provided a good fit to the average decoding errors obtained with the BOLD decoder (Figure 4A, black outline, ρ=0.99) and across individual participants (S11A, ranges: ρ= [0.92,0.98]). The *unaware* readout provided a better fit to the outputs of our neural decoder than the presented orientation (Figure 4C, $t(5)$=5.94, p=.001, paired one tailed t-test) because it captured a significant proportion of the variance in decoding errors as a function of $\Delta\theta$ (Figure 4D, $t(5)$=9.34, p=.0001, one-tailed t-test).

We next considered three readout schemes of this adapted population to maximize the likelihood of our behavioral responses (Figure 3D). The *Bayes-aware* decoder is consistent with previous Bayesian accounts of serial dependence[38], but additionally asserts that Bayesian inference occurs after encoding and that readout must account for adaptation. Alternatively, the *Bayes-unaware* decoder tests whether this awareness is necessary to achieve attractive serial dependence. Both models were able to achieve attractive biases that were positively correlated with average behavioral biases (Figure 4B), and individual biases (Figure S11B-C), but the Bayes-aware model was significantly more likely given participant responses (Figure 4E, $t(5)$=12.8, p=$5.3*10^{-5}$). We also considered the *over-aware* model to determine if a mismatch between expected and true levels of adaptation can explain attractive serial dependence without the need to invoke a formal Bayesian prior. This model also outperformed the *Bayes-unaware* model (Figure 4E, $t(5)$=3.69, p=.014) but was not significantly different from the Bayes-aware model (p=0.18, all t-tests paired, two-tailed).Finally, we examined the variance of the *unaware* decoder as well as the three readout schemes fit to behavior (Table 1) to see if they were able to reproduce patterns similar to the BOLD decoder and the behavioral responses, respectively. As model coefficients were fit independent of observed variance, correspondence between model performance and BOLD/behavioral data would provide convergent support for the best model. While the models were trained using noiseless activity at encoding, we simulated responses using Poisson rates to induce variability.

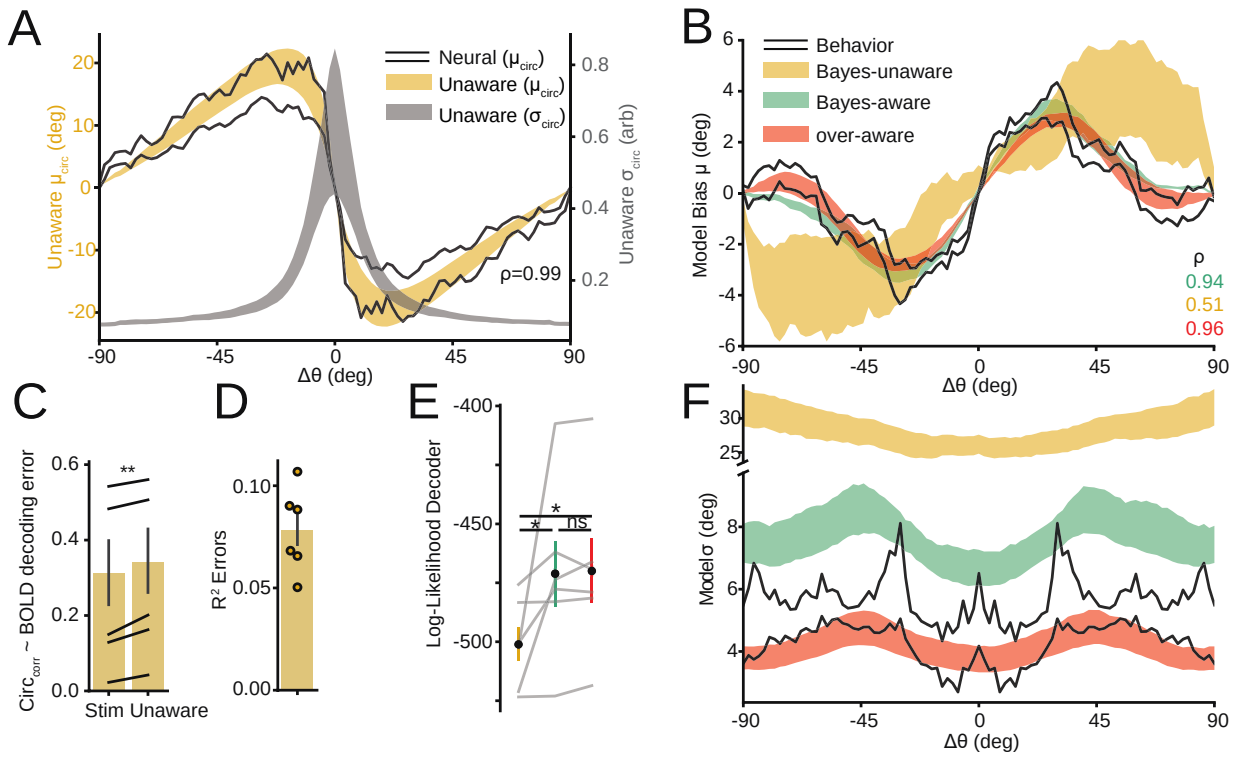


Figure 4 Model performance. **A,C,D**- Neural Decoder; **B,E,F**- Behavior. **A**: Unaware decoder (yellow) provides a good fit to neural bias (black outline). Decoded variance decreases monotonically with distance from previous stimulus. (± SEM across participants). **B**: Perceptual bias (black outline) was well fit by the Bayes-aware and over-aware models but not the Bayes-Unaware model (± SEM across participants). **C:** Unaware decoder provided a better fit to decoded responses than the presented orientation. **D**: The unaware decoder was able to explain a significant proportion of decoding error variance. **E**: Likelihood of observed responses for best fit model for each participant. Bayes-unaware significantly worse than other models. Same color scheme as **B**. **F**: Perceptual σ had a similar shape and magnitude to Bayes-aware and over-aware model fits. Bayes-unaware model output was much less precise and had a different form. Same legend as **B**. *, p<.05; **, p<.01; ***, p<.001

We simulated 1000 trials from each cross-validated fit and pooled the model outputs. We first confirmed that the variance of the *unaware* decoder was highest following small changes of Δθ (Figure 4A, gray; Figure S12C $t(5)=3.4$, p=.01, paired 1-tailed t-test <30° vs >30°) matching the output of our neural decoder (Figure 2G) and providing additional support for gain adaptation causing the observed repulsion in the fMRI data. Next, we examined the different behavioral decoders and found that, matching real behavioral responses, both the *Bayes-aware* and *over-aware* decoders were significantly more precise following small values of Δθ (Figure 4F; Figure S12C, Bayes-aware, $t(5)=-3.19$, p=.012, over-aware, $t(5)=-6.64$, p=.0006) while the *Bayes-unaware* decoder did not show this trend ($t(5)=-1.99$, p=.052). Notably, the overall magnitude of variance observed with the *Bayes-unaware* decoder was also much higher than that observed in the real behavioral data (Figure 4F) and thus provided a significantly worse fit relative to either of the aware decoders (p<.005, Figure S12A-B, paired t-tests comparing Jensen-Shannon divergence of error distributions). Together, the variance data provides additional evidence in favor of adaptation driving the repulsive biases that were observed in the BOLD data and awareness of the current state of adaptation being a requisite condition for attractive serial dependence.

Discussion

In this study, we sought to understand the neural underpinning of attractive serial dependence, and how changes in tuning properties at encoding shape behavior. Based on previous behavioral and neural studies, we expected to observe attractive biases in line with observed behavior and decoding from early visual areas [26,28,29]. Instead, we found that representations were significantly repelled from the previous stimulus starting in primary visual cortex and continuing through IPS (Figure 2F). This repulsion is consistent with bottom up adaptation beginning either at or before V1 and cascading up the visual hierarchy [11,42]. As repulsive biases are clearly in the opposite direction as behavioral biases, we built a model to link these conflicting patterns. The critical new insight revealed by the model is that only readout schemes that account for adaptation can explain attractive serial dependence. More generally, our BOLD data strongly point against an early sensory or 'perceptual' account of serial dependence and instead suggest that serial dependence is driven by post-perceptual or mnemonic circuits [34,43].

Two previous studies have examined how sensory representations are shifted by serial dependence. An fMRI study observed *attractive* biases in early visual areas that corresponded to behavioral performance, but the study was limited as it only used two orthogonal orientations (θ={45,135°}) [28]. As shown in the present results (Figures 1B, 2A) and in other studies[26,29,32,44,45], serial dependence is absent at these offsets. In addition, the stimuli were rendered at low contrast and were embedded in visual noise, making them difficult to accurately encode. Thus, the observed history bias may be more akin to perceptual 'priming' as opposed to attractive serial dependence[46,47], as individuals may have been able to detect stimuli faster and more reliably when they observed a similar stimulus in the recent past (particularly on trials where they failed to accurately encode the near-threshold orientation presented on the current trial). Further, as their decoder was trained with only two stimuli, they could not build an explicit model of orientation representations. Thus, it is unclear if enhancements at encoding correspond to a shift in orientation-selective information in voxel-activation patterns or rather to a reduction in variability. More in line with the present experiment, a second study found that population representations in FEF were *repelled* while saccades were attracted to the location of the previous stimulus[32]. Perhaps because this effect was observed in a later visual area, the authors explained their finding as a consequence of residual attentional shifts from the previous trial. Our finding of repulsive biases as early as V1 is more consistent with bottom up adaptation as attention effects tend to become more pronounced later in the visual hierarchy (and Papadimitriou and colleagues also acknowledge this as an alternative mechanism)[48–51].

In line with classic accounts, adaptation in visual cortex should lead to a reduction in energy usage during encoding[10]. However, our modeling results highlight the importance of an aware decoder, which may offset adaptation-related efficiency gains. Instead the main advantage of adaptation may be to decorrelate inputs, thus enhancing the discriminability of incoming stimuli [9,10]. The resulting biases may have little fitness cost relative to the advantage of being aware of stimulus changes potentially signaling threat or food. Indeed an optimal processing stream may emphasize differences at encoding and only favor stability once a stimulus has been selected by attention for more extensive post-perceptual processing [43]. This motif of pattern separation followed by pattern completion would not be unique to adaptive visual processing. For example, similar mechanisms have been proposed as a critical component of long term memory processing in the hippocampus and associative memory formation in

the fly mushroom body [52]. Thus, the biases introduced by adaptation may be beneficial in part because they expand the dimensionality of the representational space.

In our model, we did not explicitly define how awareness of adaptation is implemented. However, some representation of information about stimulus history appears to be a minimum requirement. The identity of the previous stimulus for spatial position and angle has previously been shown to be decodable from the spiking activity of single units in the frontal eye field (FEF) and large-scale activity patterns in human EEG [32,53]. We additionally demonstrate that information about the previous trial is encoded in patterns of fMRI activity in human visual cortex (Figure S9). These signals could potentially be represented concurrently with representations of the current stimulus in the same populations of sensory neurons, but separating activity representing current and previous stimuli may prove difficult under this scheme. An alternate, and potentially more appealing, account holds that representations of stimulus history are maintained outside of early visual areas, consistent with findings from mouse parietal cortex [4]. This anatomical segregation could disambiguate incoming sensory drive from representations of stimulus history.

For the decoding stage of our model, we established that only readout schemes that are aware of adaptation can explain attractive serial dependence. The *Bayes-aware* model is an extension of previously proposed models that employ an explicit prior but that did not consider effects of adaptation at encoding[35]. In contrast, the *over-aware* model is a novel account that can achieve similar performance without needing an explicit prior based on stimulus history. While model fit metrics did not readily distinguish one of these two models as superior, the *over-aware* model may prove to be more flexible. For instance, one of our fMRI participants showed significant repulsion from far stimuli, an observation also reported by others [29,31]. While the *over-aware* model can fit this repulsive regime, the *Bayes-aware* model is incapable of generating repulsive patterns (compare models fits for subj #3, Figure S11). This limitation of a purely Bayesian account of serial dependence is also observable in prior work (Figure 6B in [35]).

The *over-aware* – or more generally a "flexibly-aware" – decoder may also account for phenomena not covered in the present study. While behavioral (repulsive) adaptation is assumed to result from an unaware decoder [8,36], the magnitude of neural adaptation may be much larger than the resulting behavioral repulsion observed [12,19,23]. Thus, behavioral adaptation may arise when adaptation outweighs awareness (an *'under-aware'* decoder) which could arise in paradigms where inducing stimuli are task irrelevant and presented for long periods of time [19,54]. By contrast, *over-aware* decoders may arise in laboratory paradigms that involve attending to and holding in memory a weak stimulus [26,27].

In this study, we extended previous descriptions of serial dependence by quantifying how both bias and variance are shaped by stimulus history. We report a robust pattern of perception being most precise following small changes in successive stimulus features (Figure 1C-D, 2A-B). This relationship violates a perceptual 'law' proposing that bias is inversely proportional to the derivative of discrimination thresholds[55]. This 'law' would assert that our attractive bias should come with a less precise representation following small changes (or a repulsive bias to account for our enhanced precision). We argue that serial dependence is not 'violating' this law, but rather believe this is further evidence for serial dependence being a post-perceptual phenomenon. Neural representations exhibit repulsive biases, expanding the perceptual space and allowing greater discriminability. When these

representations are read out by an aware decoder, the bias is undone but the enhanced discriminability remains (Figure 4F).

Acknowledgements

Funded by NEI R01-EY025872 to JTS, and NIMH Training Grant in Cognitive Neuroscience (T32-MH020002) to TS. Thanks to Chaipat Chunharas for critical discussions in experimental design and assistance with scanning and to Anika Jollorina and Shuangquan Feng for assistance with behavioral data collection. Code for Bayesian decoder adapted from code provided by Ruben van Bergen. Thanks to Marcelo Mattar for helpful comments on our model and to Margaret Henderson, Sunyoung Park, and Kirsten Adam for thoughtful comments on earlier versions of the manuscript.

Works Cited

1. Dong, D. W. & Atick, J. J. Statistics of natural time-varying images. *Netw. Comput. Neural Syst.* **6**, 345–358 (1995).
2. Felsen, G., Touryan, J. & Dan, Y. Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. *Netw. Comput. Neural Syst.* **16**, 139–149 (2005).
3. 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).
4. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).
5. Ashourian, P. & Loewenstein, Y. Bayesian Inference Underlies the Contraction Bias in Delayed Comparison Tasks. *PLoS ONE* **6**, e19551 (2011).
6. 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).
7. Barlow, H. B. Possible Principles Underlying the Transformations of Sensory Messages. in *Sensory Communication* (ed. Rosenblith, W. A.) 216–234 (The MIT Press, 2012). doi:10.7551/mitpress/9780262518420.003.0013.
8. Benucci, A., Saleem, A. B. & Carandini, M. Adaptation maintains population homeostasis in primary visual cortex. *Nat. Neurosci.* **16**, 724–729 (2013).
9. Clifford, C. W. G. *et al.* Visual adaptation: Neural, psychological and computational aspects. *Vision Res.* **47**, 3125–3131 (2007).
10. Clifford, C. W. G., Wenderoth, P. & Spehar, B. A functional angle on some after-effects in cortical vision. *Proc. R. Soc. Lond. B Biol. Sci.* **267**, 1705–1710 (2000).
11. Dhruv, N. T. & Carandini, M. Cascaded Effects of Spatial Adaptation in the Early Visual System. *Neuron* **81**, 529–535 (2014).

12. Dragoi, V., Rivadulla, C. & Sur, M. Foci of orientation plasticity in visual cortex. *Nature* **411**, 80–86 (2001).

13. Dragoi, V., Sharma, J. & Sur, M. Adaptation-Induced Plasticity of Orientation Tuning in Adult Visual Cortex. *Neuron* **28**, 287–298 (2000).

14. Durant, S., Clifford, C. W. G., Crowder, N. A., Price, N. S. C. & Ibbotson, M. R. Characterizing contrast adaptation in a population of cat primary visual cortical neurons using Fisher information. *J. Opt. Soc. Am. A* **24**, 1529 (2007).

15. Kohn, A. & Movshon, J. A. Adaptation changes the direction tuning of macaque MT neurons. *Nat. Neurosci.* **7**, 764–772 (2004).

16. Gardner, J. L. *et al.* Contrast Adaptation and Representation in Human Early Visual Cortex. 14 (2005).

17. Patterson, C. A., Wissig, S. C. & Kohn, A. Adaptation Disrupts Motion Integration in the Primate Dorsal Stream. *Neuron* **81**, 674–686 (2014).

18. Dekel, R. & Sagi, D. Tilt aftereffect due to adaptation to natural stimuli. *Vision Res.* **117**, 91–99 (2015).

19. He, S. & MacLeod, D. I. A. Orientation-selective adaptation and tilt after-effect from invisible patterns. *Nature* **411**, 473–476 (2001).

20. Moradi, F., Koch, C. & Shimojo, S. Face Adaptation Depends on Seeing the Face. *Neuron* **45**, 169–175 (2005).

21. Abbonizio, G., Clifford, C. & Langley, K. Contrast adaptation may enhance contrast discrimination. *Spat. Vis.* **16**, 45–58 (2002).

22. Clifford, C. W. G., Wyatt, A. M., Arnold, D. H., Smith, S. T. & Wenderoth, P. Orthogonal adaptation improves orientation discrimination. *Vision Res.* **41**, 151–159 (2001).

23. Jin, D. Z., Dragoi, V., Sur, M. & Seung, H. S. Tilt Aftereffect and Adaptation-Induced Changes in Orientation Tuning in Visual Cortex. *J. Neurophysiol.* **94**, 4038–4050 (2005).

24. Phinney, R. E., Bowd, C. & Patterson, R. Direction-selective Coding of Stereoscopic (Cyclopean) Motion. *Vision Res.* **37**, 865–869 (1997).

25. Cicchini, G. M., Mikellidou, K. & Burr, D. Serial dependencies act directly on perception. *J. Vis.* **17**, 6 (2017).

26. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).

27. Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K. & Whitney, D. Serial dependence in position occurs at the time of perception. *Psychon. Bull. Rev.* **25**, 2245–2253 (2018).

28. St. John-Saaltink, E., Kok, P., Lau, H. C. & de Lange, F. P. Serial Dependence in Perceptual Decisions Is Reflected in Activity Patterns in Primary Visual Cortex. *J. Neurosci.* **36**, 6186–6192 (2016).

29. Fritsche, M., Mostert, P. & de Lange, F. P. Opposite Effects of Recent History on Perception and Decision. *Curr. Biol.* **27**, 590–595 (2017).

30. Cicchini, G. M., Mikellidou, K. & Burr, D. C. The functional role of serial dependence. *Proc. R. Soc. B Biol. Sci.* **285**, 20181722 (2018).

31. Bliss, D. P., Sun, J. J. & D'Esposito, M. Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* **7**, 14739 (2017).

32. Papadimitriou, C., White, R. L. & Snyder, L. H. Ghosts in the Machine II: Neural Correlates of Memory Interference from the Previous Trial. *Cereb. Cortex* bhw106 (2016) doi:10.1093/cercor/bhw106.

33. Papadimitriou, C., Ferdoash, A. & Snyder, L. H. Ghosts in the machine: memory interference from the previous trial. *J. Neurophysiol.* **113**, 567–577 (2015).

34. Barbosa, J. *et al.* Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).

35. van Bergen, R. S. & Jehee, J. F. M. Probabilistic Representation in Human Visual Cortex Reflects Uncertainty in Serial Decisions. *J. Neurosci.* **39**, 8164–8176 (2019).

36. Seriès, P., Stocker, A. A. & Simoncelli, E. P. Is the Homunculus "Aware" of Sensory Adaptation? *Neural Comput.* **21**, 3271–3304 (2009).

37. Cicchini, G. M. & Burr, D. C. Serial effects are optimal. *Behav. Brain Sci.* **41**, e229 (2018).

38. van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).

39. Crick, F. & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375**, 121–123 (1995).

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

41. Grunewald, A., Bradley, D. C. & Andersen, R. A. Neural Correlates of Structure-from-Motion Perception in Macaque V1 and MT. *J. Neurosci.* **22**, 6195–6207 (2002).

42. Patterson, C. A., Duijnhouwer, J., Wissig, S. C., Krekelberg, B. & Kohn, A. Similar adaptation effects in primary visual cortex and area MT of the macaque monkey under matched stimulus conditions. *J. Neurophysiol.* **111**, 1203–1213 (2014).

43. Pascucci, D. *et al.* Laws of concatenated perception: Vision goes for novelty, decisions for perseverance. *PLOS Biol.* **17**, e3000144 (2019).

44. Kiyonaga, A., Scimeca, J. M., Bliss, D. P. & Whitney, D. Serial Dependence across Perception, Attention, and Memory. *Trends Cogn. Sci.* **21**, 493–497 (2017).

45. Fischer, C. *et al.* Context information supports serial dependence of multiple visual objects across memory episodes. *Nat. Commun.* **11**, 1932 (2020).

46. Bar, M. & Biederman, I. Subliminal Visual Priming. *Psychol. Sci.* **9**, 6 (1998).

47. Tulving, E. & Schacter, D. Priming and human memory systems. *Science* **247**, 301–306 (1990).

48. Kastner, S. Mechanisms of Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI. *Science* **282**, 108–111 (1998).

49. Luck, S. J., Chelazzi, L., Hillyard, S. A. & Desimone, R. Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex. *J. Neurophysiol.* **77**, 24–42 (1997).

50. Sprague, T. C. & Serences, J. T. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat. Neurosci.* **16**, 1879–1887 (2013).

51. Treue, S. Visual attention: the where, what, how and why of saliency. *Curr. Opin. Neurobiol.* **13**, 428–432 (2003).

52. Cayco-Gajic, N. A. & Silver, R. A. Re-evaluating Circuit Mechanisms Underlying Pattern Separation. *Neuron* **101**, 584–602 (2019).

53. Bae, G.-Y. & Luck, S. J. Reactivation of Previous Experiences in a Working Memory Task. *Psychol. Sci.* **30**, 587–595 (2019).

54. Gibson, J. J. & Radner, M. Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *J. Exp. Psychol.* **20**, 453–467 (1937).

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

56. Roth, Z. N., Heeger, D. J. & Merriam, E. P. Stimulus vignetting and orientation selectivity in human visual cortex. *eLife* **7**, e37241 (2018).

57. SciPy 1.0 Contributors *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

58. Stein, H. *et al.* Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat. Commun.* **11**, 4250 (2020).

59. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).

Methods

Participants

Behavioral study: 56 participants (male and female) were drawn from a subject pool of primarily undergraduate students at UC San Diego. All subjects gave written consent to participate in the study in accordance with the UC San Diego IRB, and were compensated either monetarily or with class credit. Of these 56 participants, 9 were removed from further analysis for completing less than 200 trials (2) or getting less than 60% of trials correct (7). We included the remaining 47 participants who completed on average 421 trials, range: [204, 988], in our lab over the course of 1 to 3 sessions.

fMRI study: 6 participants (3 female, mean age 24.6 ±0.92) participated in four, 2-hour scanning sessions. Each subject completed between 748 and 884 trials (mean 838.7). For two participants, one session had to be repeated due to technical difficulties that arose during scanning.

Behavioral Discrimination Task

Participants in the behavior-only study completed the task on a desktop computer in a sound attenuated room. Subjects were seated with a chin rest to stabilize viewing 50 cm from a 39 by 29 cm CRT monitor (1600x1200 px) with a visual angle of 42.6° (screen width). Each trial consisted of a full-field oriented grating (1000 ms) which had to be remembered across a delay period (3,500 ms) before a test. At test, the participant judged whether a line was slightly clockwise (CW) or counter-clockwise (CCW) relative to the remembered orientation (max response time window: 3,000ms, Figure 1A). The oriented grating consisted of a sine wave grating (spatial frequency 1.73 cycles/°, 0.8 Michelson contrast) multiplied by a 'donut' mask (outer diameter Ø=24.3°, inner Ø=1.73°). The stimulus was then convolved with a 2D Gaussian filter (1.16° kernel, SD = 0.58°) to minimize edge artifacts [56]. Phase and orientation were randomized across trials, and the stimulus was phase-reversed every 250ms. After the offset of the oriented grating, a mask of filtered noise was presented for 500ms. The mask was generated by band passing white noise [low 0.22, high 0.87 cycles/°], multiplying by the same donut mask, and convolving with a 2D Gaussian filter (0.27° kernel, SD = 0.11°). The mask was phase reversed once after 250 ms. A black fixation point (diameter .578°) was displayed throughout the extent of the block and turned white for 500 ms prior to stimulus onset on each trial. The probe was a white line (width 0.03°, length 24.3°) masked by the same donut. Subjects indicated whether the probe line was CW or CCW from the remembered orientation by pressing one of two buttons ('Q', 'P') with their left and right pointer fingers. The next trial started after a 1000ms inter trial interval (ITI). For some behavioral participants (n=9) delay and ITI were varied between 0.5-7.5s without notable effects on performance.

First, subjects completed a training block to ensure that they understood the task. Next, they completed a block of trials where difficulty was adjusted by changing the probe offset ($\delta\theta$) between the stimulus and probe to achieve 70% accuracy. This $\delta\theta$ was used in subsequent blocks and was adjusted on a per-block basis to keep performance at approximately 70%. Participants completed an average of 5.76±0.24 blocks [min = 3, max = 9]. Some participants completed the task with slight variations in the distribution and sequence of orientations presented. For completeness we include those details here. Note, however, we additionally report a set of control analyses in which we repeat all of our main analyses excluding blocks with binned stimuli and find no relevant difference in behavior. For most participants, stimuli were pseudo-randomly distributed across the entire 180° space such that they were uniformly distributed across blocks of 64 trials (n=25). However, some participants saw stimuli that were binned (with some jitter) every 22.5° to purposefully avoid cardinal and oblique orientations (11.25°,

33.75°, 56.25°, etc.) and the trial sequence was ordered so that a near oblique orientation was always followed by a near cardinal orientation (n=7). This was implemented to maximize our ability to observe serial dependencies in our binary response data as it is typically strongest around orientation changes of 20° and is more pronounced around oblique orientations [37]. The remaining participants completed both blocks with uniform and blocks with binned stimuli (n=14). All participants were interviewed after the study and reported that stimuli were non-predictable and that all orientations felt equally likely. For our main analysis we include all trials from all participants, irrespective of whether they participated in uniform blocks, binned blocks, or both.

fMRI Discrimination Task

In the scanner, participants completed the behavioral task outlined above with slight modifications. fMRI participants completed the task using a fiber-optic button box while viewing stimuli through a mirror projected onto a screen mounted inside of the bore. The screen was 24 by 18 cm and was viewed at a distance of 47 cm (width: 28.6° visual angle; 1024x768 px native resolution). The stimulus timing was the same except that the sample-to-probe delay period was either 5, 7 or 9 s and the ITIs were uniformly spaced between 5s and 9s and shuffled pseudo-randomly on each run of 17 trials. The oriented gratings had a spatial frequency of 1.27 cycles/°, outer Ø=21.2°, inner Ø=2.37° and were smoothed by a Gaussian filter (0.79° kernel, sd=0.79°). The noise patch (SF low 0.16, high 0.63 cycles/°) was also smoothed by a Gaussian filter (0.29° kernel, sd=0.11°). The probe stimulus was a white line (width = 0.03°).

fMRI participants completed 44-52 blocks of 17 trials spread across 4, two-hour scanning sessions for a total of 748-884 trials. As in the behavior-only task described above, 4 out of 6 fMRI subjects had some blocks of trials where the stimuli were binned in 22.5° increments and ordered in a non-independent manner (21-24 blocks/participant). However, all of the fMRI subjects also participated in blocks with a uniform distribution of orientations across the entire 180° space (24-52 blocks/participant). For our main analysis we include all trials from all participants. However, as with the behavioral analyses, we also report control analyses in which we repeat all of our main analyses excluding blocks with non-random stimuli.

fMRI Localizer Task

Interleaved between the main task blocks, participants completed an independent localizer task used for voxel selection where they were presented with a sequence of grating stimuli at different orientations. Stimuli had a pseudo-randomly determined orientation that either matched the spatial location occupied by the *donut* stimuli used in our main task (outer diameter Ø=21.2°, inner diameter Ø=2.37°) or were a smaller foveal oriented Gabor corresponding to the 'hole' in the *donut* stimuli (diameter Ø=2.37°). Participants were instructed to attend to one of three features orthogonal to orientation depending on the block: detect a contrast change across the entire stimulus, detect a small grey blob appearing over part of the stimulus, or detect a small change in contrast at the fixation point. Each stimulus was presented for 6000 ms and was separated by an ITI ranging from 3-8s.

Response Bias

Each trial consisted of a stimulus and a probe separated by a probe offset ($\delta\theta$) that was either positive (probe is CW of stimulus) or negative. Participants judged whether the probe was CW or CCW relative to the remembered orientation by making a binary response. To quantify the precision and the response bias, we fit participant responses with a Gaussian cumulative density function with parameters

μ and σ corresponding to the *bias (mean)* and *standard deviation* of the distribution. The likelihood of a given distribution was determined by the area under the curve (AUC) of the distribution of CW (CCW) offsets between the stimulus and the probe ($\delta\theta$) on trials where the participant responded CW (CCW; see Figure S1). In extreme cases, a very low standard deviation (σ) value with no bias would mean that all $\delta\theta$ would lie outside the distribution and the participant would get every trial correct (Figure S1A). A high negative bias (μ) value would mean that $\delta\theta$ would always lie CW relative to the distribution and the participant would respond CW on every trial (Figure S1B). The best fitting parameters were found using a bounded minimization algorithm (limited memory BFGS) on the negative log likelihood of the resulting responses (excluded the small number of trials without a response) given the generated distribution [57]. We included a constant 25% guess rate in all model fits to ensure the likelihood of any response could never be 0 (critical for later modelling). While this was critical to fitting our model to raw data, the specific choice had no qualitative effect on our behavioral findings besides making the σ values smaller compared to having a 0% guess rate. By having a constant guess rate rather than varying it as a free parameter we were able to directly compare σ values across participants as a measure of performance. Realistic model parameters and the effects of bias on response likelihood are also demonstrated (Figure S1 C-D).

Serial Dependence

To quantify the dependence of responses on previous stimuli, we analyzed response bias and variance as a function of the difference in orientation between the previous and current orientation ($\Delta\theta = \theta_{n-1} - \theta_n$). We performed this analysis using a sliding window of 16°. To improve power, we 'folded' our response data such that, when examining bias at 30° we included values from 22°- 38° as well as responses from -22°- (-38°) by inverting both the responses and probe offsets ($\delta\theta$) for the negative values of Δθ. This procedure removes any systematic responses biases (e.g., favoring CW responses) and, as a result, the figures presenting serial dependence have rotational symmetry across the origin[34,58].

We additionally fit a Derivative of Gaussian (DoG) function to parameterize the bias of participant responses. The DoG function is parameterized with an amplitude A and width w

$$y = xAwce^{-(wx)^2} \quad [1]$$

where $c = \sqrt{2e}$ is a normalization constant. For the purpose of fitting to our participant responses, x is Δθ and y corresponds to μ in our response model. For each participant we adjusted three parameters: A, w, and σ to maximize the likelihood of participant responses. We report the magnitude of our fits as well as the resulting full width at half max (FWHM) estimated numerically.

Response Precision

In addition to quantifying how responses were biased as a function of stimulus history, we also estimated how precise responses were depending on their unsigned distance from the previous stimulus (|Δθ|). We used the same 'folding' procedure described in the previous section and only included trials on the right half of our bias/variance plots (eg. Figure 1C, $\Delta\theta > 0$) to avoid double counting trials. Values from the bin with more samples (typically 'far') were resampled (31 repetitions) without replacement with the number of samples in the smaller bin and the median chosen to control for sample number differences.

Scanning

fMRI task images were acquired over the course of four 2-hour sessions for each participant in a General Electric Discovery MR750 3.0T scanner at the UC San Diego Keck Center for Functional Magnetic Resonance Imaging. Functional echo-planar imaging (EPI) data were acquired using a Nova Medical 32-channel head coil (NMSC075-32- 3GE-MR750) and the Stanford Simultaneous Multi-Slice (SMS) EPI sequence (MUX EPI), with a multiband factor of 8 and 9 axial slices per band (total slices 72; 2-mm3 isotropic; 0-mm gap; matrix 104 x 104; field of view 20.8 cm; TR/TE 800/35 ms; flip angle 52°; in-plane acceleration 1). Image reconstruction and un-aliasing was performed on cloud-based servers using reconstruction code from the Center for Neural Imaging at Stanford. The initial 16 repetition times (TRs) collected at sequence onset served as reference images required for the transformation from k-space to the image space. Two 17s runs traversing k-space using forward and reverse phase-encoding directions were collected in the middle of each scanning session and were used to correct for distortions in EPI sequences using FSL top-up (FMRIB Software Library) for all runs in that session (Andersson et al. 2013, Jenkinson et al. 2012). Reconstructed data was motion corrected and aligned to a common image. Voxel data from each run was de-trended (8TR filter) and z-scored.

We also acquired one additional high-resolution anatomical scan for each subject (1 x 1 x 1-mm3 voxel size; TR 8,136 ms; TE 3,172 ms; flip angle 8°; 172 slices; 1-mm slice gap; 256x192-cm matrix size) during a separate retinotopic mapping session using an Invivo eight-channel head coil. This scan produced higher quality contrast between gray and white matter and was used for segmentation, flattening, and visualizing retinotopic mapping data. The functional retinotopic mapping scanning was collected using the 32-channel coil described above and featured runs where participants viewed checkerboard gratings while responding to an orthogonal feature (transient contrast changes). Separate runs featured alternating vertical and horizontal bowtie stimuli; rotating wedges; and an expanding donut to generate retinotopic maps of the visual meridian, polar angle, and eccentricity respectively (see Sprague and Serences, 2013). These images were processed using FreeSurfer and FSL functions and visual regions of interest (ROI) were manually drawn on surface reconstructions (for areas: V1-V3, V3AB, hV4, IPS0-IPS2, VO, LO, and TO).

Voxel Selection

To include only voxels that showed selectivity for the location of the oriented grating stimulus used in our main experimental task, we used responses evoked during the independent *localizer* task (see fMRI Localizer Task). For all analysis we used TRs 5-11 (4-8.8s) following stimulus onset. First, voxels were selected based on their response to the spatial location of the grating stimulus by performing a t-test on the responses of each voxel evoked by the donut and the donut-hole stimuli, selecting the 50% of the voxels most selective to the donut for a given ROI. Of the voxels that passed this cutoff, we then performed an ANOVA across 10° orientation bins and selected the 50% of voxels with the largest F-score thus retaining ~25% of the initial voxel pool. These selected voxels were used in all main analysis.

Orientation Decoding (BNC)

We performed orientation decoding on BOLD activation patterns using a sliding temporal window of 4 TRs. For most analysis we focused on a 3.2s (4 TR) window centered 6.4 s after stimulus presentation. The Bayesian Noise Correlation (BNC) decoder assumes voxels are composed of populations of neurons with tuning functions centered on one of 8 orientations evenly tiling the 180° space. The response of population i to stimulus θ is given by:

$$c_i(\theta) = \max\left(0, cos^5(\theta - \varphi_i)\right) \quad [2]$$

where φ_i is the center of the tuning function. The response of voxel j is defined as a weighted sum of these hypothetical populations:

$$B_j = \sum_i^8 c_i w_i \quad [3]$$

Or in matrix notation,

$$B = CW \quad [4]$$

Where B *(trial x voxel)* is the resulting BOLD activity, C *(trial x channel)* is the hypothetical population response, and W *(channel x voxel)* is the weight matrix. The weight matrix W is estimated as:

$$\widehat{W} = C^{-1}B \quad [5]$$

where C^{-1} (*channel x trial*) is the pseudo-inverse of C (implemented using the NumPy pinv function). This model was used to generate a linear estimate of voxel responses. The resulting residuals correspond to voxel noise.

$$\hat{B} = C\widehat{W} \quad [6]$$

$$B_{Noise} = B - \hat{B} \quad [7]$$

Noise correlations are known to contribute to observed activity and can be detrimental to our resulting decoding capabilities [59]. To reduce the impact of noise correlations across similarly tuned populations, we implemented a Bayesian decoder that explicitly models these correlations [38]. Briefly, we modeled noise as coming from 3 distinct components: *global noise* shared across all voxels, *channel noise* shared across neurons with similar tuning, and *voxel noise* explaining residual fluctuations in individual voxels (see [38] for more details). The magnitude of these noise sources was estimated through maximizing the likelihood of the observed residuals using a multivariate Gaussian defined by (number of voxels) + (1 global) + (1 channel) parameters. After estimating noise sources, we could estimate the posterior probability distribution given our fit weights $\widehat{W}$ and noise parameters $\hat{\Omega}$:

$$P\left(\theta \middle| B; \widehat{W}, \hat{\boldsymbol{\Omega}}\right) = \frac{P\left(\theta \middle| B; \widehat{W}, \hat{\Omega}\right)p(\theta)}{\int P\left(\theta \middle| B; \widehat{W}, \hat{\Omega}\right)p(\theta)\, d\theta} \quad [8]$$

For each trial we then selected the θ most likely to have given rise to response B given $\widehat{W}\ and\ \hat{\Omega}$ as our decoded orientation and used its vector length as a proxy for model certainty. The encoding and noise parameters of our model were fit to a subset of data and used to estimate responses for held-out trials of the task data. We used leave-1-block-out cross-validation where each block was a set of 4 consecutive runs (64 trials). These blocks featured orientations that were evenly distributed across the entire 180° space to ensure a balanced training set. We performed additional analysis training a model on the localizer task and testing on the memory task as well as cross-validating

548 within the localizer task. These models had lower SNR than models trained on the task but showed
549 qualitatively similar results as our task trained neural decoder.

550

551 ## Orientation Decoding (IEM)

552 For some analysis we additionally include the outputs of an Inverted Encoding Model (IEM). The
553 IEM uses the same encoding model as the BNC decoder (eq. [2-5]) but does not generate a specific
554 model of noise covariance. We instead inverted our estimated weight matrix ($\widehat{W}$) to estimate the
555 channel response on held out trials, $\hat{C} = B\widehat{W}^{-1}$, where $\widehat{W}^{-1}$ is the pseudo-inverse of $\widehat{W}$. The circular
556 mean of $\hat{C}$ was taken as the orientation estimate.

557 ## Neural Bias

558 To quantify how BOLD representations were biased by sensory history we computed the circular
559 mean of decoding errors (θ_{error} = wrap($\theta_{decode} - \theta_{stim}$)):

$$\mu_{circ} = angle(\vec{R}), \quad [9]$$

560

$$\vec{R} = \frac{1}{nTrials}\sum_{k=0}^{nTrials} e^{i\theta_{error}^{k}}. \quad [10]$$

561

562 We estimated this bias using the same 16° sliding window as a function of Δθ used for visualizing
563 response bias from participant responses. We additionally quantified the magnitude of the bias in
564 decoding errors by fitting a DoG function to the raw decoding errors by minimizing the residual sum of
565 squares (RSS) and reporting the amplitude term.

566 ## Neural Variance

567 To quantify the variance of decoded orientations from visual areas, we computed the circular
568 standard deviation:

$$\sigma_{circ} = \sqrt{-2\ln|\vec{R}|}. \quad [11]$$

569

570 This was estimated on both binned decoding errors (eq. 10), or on the single trial posterior estimate
571 from our orientation decoder (eq. 8). This was visualized using the same sliding window analysis as well
572 as in reference to whether it was close or far from the previous stimulus. Both pooled and single trial
573 estimates are reported and give similar results.

574 ## Modeling

575 We sought to develop a model that could explain both neural and behavioral biases as a
576 function of stimulus history. For the fMRI data, we focused on explaining changes in encoding that could
577 lead to the observed biases in the output of the BOLD decoder that was specifically designed to be
578 '*unaware*' of stimulus history. To explain the behavioral data, we assumed that a decoder would receive
579 inputs from the same population of sensory neurons that we measured with fMRI and that the decoder
580 would read out this information in a manner that gives rise to attractive serial dependence. We
581 considered readout models that were either *unaware*, *aware*, or *over-aware* of adaptation and

additionally applied a Bayesian inference stage, which integrates prior expectations of temporal stability, to the *unaware* and *aware* decoders [35]. We then compared performance between these competing models to see which could best explain our behavioral data.

Our full models consisted of two stages: an encoding stage where the gain of artificial neurons was changed as a function of the previous stimulus (adaptation) and a decoding stage where the readout from this adapted population was modified. The encoding population consisted of 100 neurons with von Mises tuning curves evenly tiling the 180° space. The expected unadapted population response is:

$$Resp_N(\theta_n) = R\ \gamma_N e^{\kappa \cos(\Phi - \theta_n) - 1} \quad [12]$$

Where γ_N is the scalar 1 for constant gain without adaptation, Φ is the vector of tuning curve centers, θ_n is the orientation of the current stimulus, κ=1.0 is a constant controlling tuning width, and R is a general gain factor driving the average firing rate. We implemented sensory adaptation by adjusting the gain of tuning curves relative to the identity of the previous stimulus, θ_{n-1} (Figure 3A, *Gain Adaptation*):

$$\gamma_A(\theta_{n-1}) = \gamma_N - rect(\gamma_m \cos^3(\gamma_s(\Phi - \theta_{n-1}))) \quad [13]$$

Where γ_m is the magnitude of adaptation, γ_s scales the width of adaptation, and *rect* is the half-wave rectifying function. The responses of the adapted population thus depend on both the current and previous stimulus (Figure 3A, *Efficient Encoding*):

$$Resp_A(\theta_n, \theta_{n-1}) = R\ \gamma_A e^{\kappa \cos(\Phi - \theta n) - 1} \quad [14]$$

Unaware decoder: We first considered a model in which an adapted orientation-encoding representation is being decoded by an *unaware* readout mechanism (Figure 3B). The likelihood of each orientation giving rise to the observed response profile across *N* neurons was estimated assuming activity was governed by a Poisson process:

$$P_{unaware}(\ Resp_A|\theta) = \exp\left(\sum_{i=1}^{N} \log P_{Poisson}\left(Resp_A^i(\theta); Resp_N^i(\theta)\right)\right) \quad [15]$$

$$P_{Poisson}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad [16]$$

Where $Resp_N^i(\theta)$ is the expected response of the unadapted neuron i to stimulus θ and $P_{Poisson}(k;\lambda)$ is the probability of observing k spikes given an expected firing rate of λ. The decoded orientation is then the θ giving rise to the maximum likelihood (MLE).

Aware decoder: In addition to the unaware decoder, we also evaluated the ability of a decoder that was aware of the current state of adaptation to explain behavior. The *aware* decoder differs from the *unaware* decoder in that its assumed activity level for each unit is modulated as a function of stimulus history:

$$P_{aware}(\,Resp_A|\theta_n;\theta_{n-1}) = \exp\left(\sum_{i=1}^{N}\log P_{Poisson}\left(Resp_A^i(\theta_n,\theta_{n-1}),Resp_A^i(\theta_n,\theta_{n-1})\right)\right) \quad [17]$$

Note that here the rate parameter $k \equiv \lambda \equiv Resp_A$ such that the observed and expected values perfectly align with the presented orientation. $P_{aware}(\,Resp_A|\theta_n;\theta_{n-1})$ is dependent on sensory history and is non-biased.

Over-Aware decoder: Our final decoding scheme we call the *over-aware decoder*. This model can test whether serial dependence can be achieved without an explicit stage of Bayesian inference introduced in the next section. The decoder has an assumed adaptation defined by a unique set of free parameters $\gamma_{m2}\ and\ \gamma_{s2}$ which shapes a separate gain adaptation:

$$\gamma_{OA}(\theta_{n-1}) = \gamma_N - rect(\gamma_{m2}\cos^3(\gamma_{s2}(\Phi-\theta_{n-1}))) \quad [18]$$

which in turn shapes the response profile of $Resp_{OA}$ in the same manner as $Resp_A$. The likelihood profile is then defined as:

$$P_{over-aware}(\,Resp_A|\theta) = \exp\left(\sum_{i=1}^{N}\log P_{Poisson}\left(Resp_A^i(\theta);Resp_{OA}^i(\theta,\theta_{n-1})\right)\right) \quad [19]$$

where our expected (assumed) rate λ is designated by $Resp_{OA}$. By having a larger assumed adaptation than implemented at encoding (through either $\gamma_{m2} > \gamma_m$ or $\gamma_{s2} > \gamma_s$) the net effect of the over-aware decoder should be behavioral attraction.

Bayesian Inference: In addition, we explored the effect of applying an explicit Bayesian prior based on temporal contiguity to the likelihood functions derived from these different readout schemes. This type of prior has been previously used to explain behavioral biases without considering how encoding might also be affected by stimulus history [35]. Specifically, the prior is defined by the transition probability between consecutive stimuli and is defined as a mixture model of a circular Gaussian and a uniform distribution:

$$P_T(\theta_n|\theta_{n-1}) = \frac{1}{Z}e^{-\frac{angle(\theta,\theta_{n-1})^2}{2\psi^2}} \quad [20]$$

$$P_{Bayesian}(\theta_n|\theta_{n-1}) = P_{SAME}P_T(\theta|\theta_{n-1}) + (1-P_{SAME})\frac{1}{2\pi} \quad [21]$$

With P_{SAME} set to 0.64 (as found empirically in [35]), Z as a normalization constant so P_T integrates to 1, and ψ is a free parameter describing the variance of the transition distribution. This prior (Figure 3C, black line) is multiplied by the *unaware* likelihood (Figure 3C, yellow dashed-line): to get the posterior estimate of our *Bayesian-unaware* decoder (Figure 3C, yellow solid-line):

$$P_{Bayesian-unaware}(\,\theta_n|Resp_A;\,\theta_{n-1}) = P_{Bayesian}(\theta|\theta_{n-1})P_{unaware}(\,Resp_A|\theta_n) \quad [22]$$

We can additionally examine a *Bayesian-aware* decoder by substituting its respective likelihood function. We did not examine a *Bayesian-over-aware* model so that all decoding models would have the same number of free parameters and so that we could directly evaluate the need for an explicit prior.

Model Fitting: The *encoding* stage of the model has two free parameters and for each subject these parameters were optimized to minimize the residual sum of squares (RSS) between our measured fMRI decoding errors and the decoding errors of our *unaware* decoder. For simplicity we only fit our model to decoding errors from V3 as it had the highest SNR, but other early visual ROIs showed similar results. After fitting the *encoding* stage of the model, we then separately fit the three competing *decoding* models to best account for the behavioral data: *Bayes-unaware, Bayes-aware,* and *over-aware* (two free parameters each). The output of this readout stage was treated as the behavioral bias (μ) and the free parameters were optimized to maximize the likelihood of the observed responses (assuming constant standard deviation σ estimated empirically for each participant). For the purposes of fitting the model, the firing rates of the modelled neurons were deterministic (no noise process). Having noiseless activity had no effect on the expected bias (verified with additional simulations) and served to make model fitting more reliable and less computationally intensive. Both stages of the model were fit using the same cross-validation groups as our neural decoder. To ensure all models had a sufficient chance of achieving a good fit to behavioral data, we implemented a grid search sampling 30 values along the range of each variable explored (900 locations total) followed by a local search algorithm (Nelder-Mead) around the most successful grid point. We found dense sampling of the initial parameter space was especially important for our *Bayes-unaware* model.

Model Evaluation: For bias of neural and behavioral responses, we evaluated the performance of the two stages of our model separately. These stages must be evaluated in a qualitatively different manner as the neural data gives us an orientation estimate for each trial while the behavioral data consists of binary responses. For the encoding stage, we quantified how well the output of our *unaware* decoder predicted the raw errors of our BOLD decoder using circular correlation. The performance of this model was contrasted with the true presented orientation which is analogous to the representation of an unadapted population. We additionally computed the variance of the neural decoding errors explained by the model bias (R^2). For the decoding stage of our model, we compared the log-likelihood of observed responses for each model.

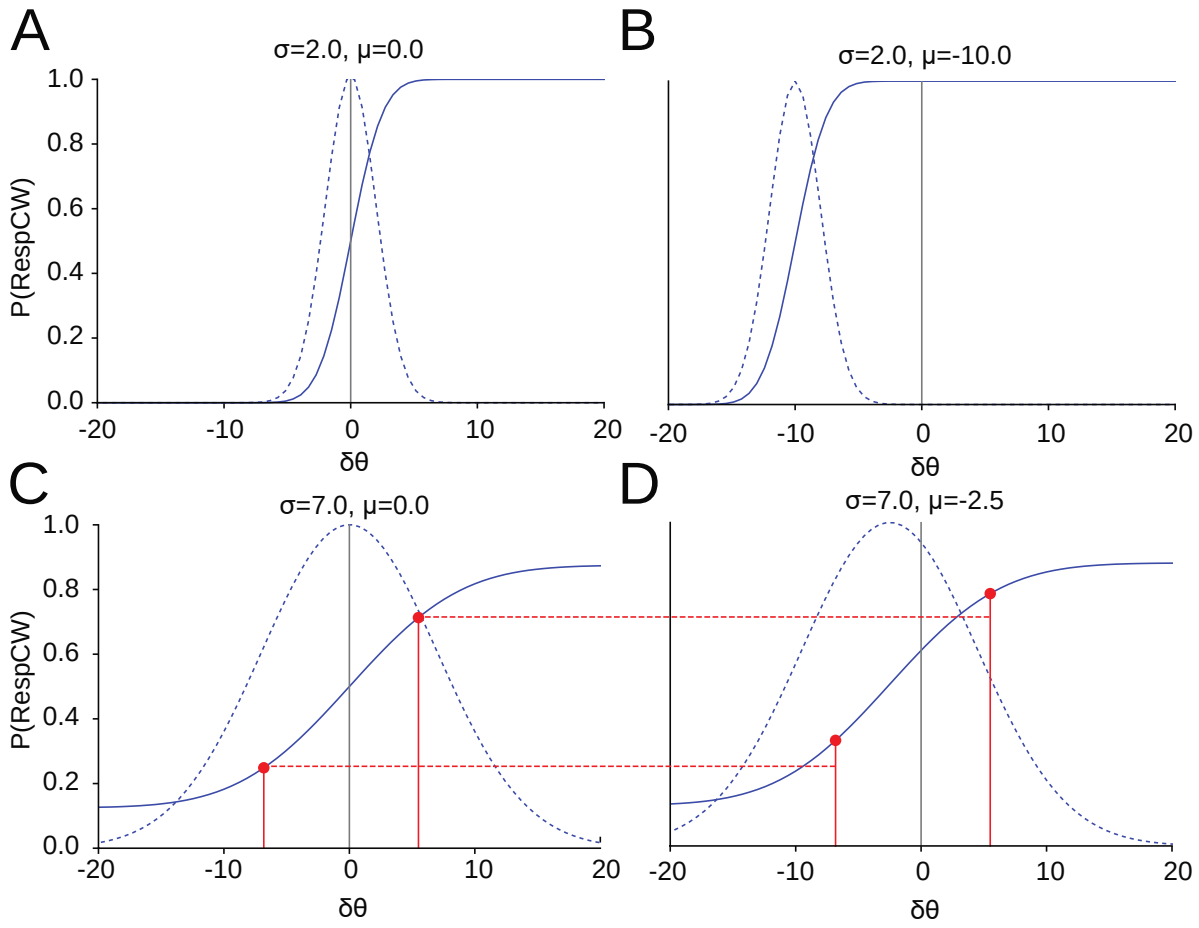
We additionally estimated the variance of our models using neurons with rates generated by a Poisson process. The average bias was unaffected by allowing random fluctuations in activity, but the trial-to-trial variance increased. To get a stable estimate, we simulated 1000 trials for each set of parameters estimated for a cross-validation loop for each participant and pooled these outputs. We compared the overall variance of our models to our single parameter estimate of participant precision using Jensen-Shannon divergence. We additionally examined relative precision of our model for close and far trials in the same manner as participant responses and decoding errors (Response Precision).

Data/Code Availability (upon acceptance for publication)

Code for processing raw data as well as for analyzing decoded representations can be found here (GITHUB). This includes all processing performed on our BOLD data and our implementation of the Bayesian decoder in Python code for running our model as well as the data used to fit models can be

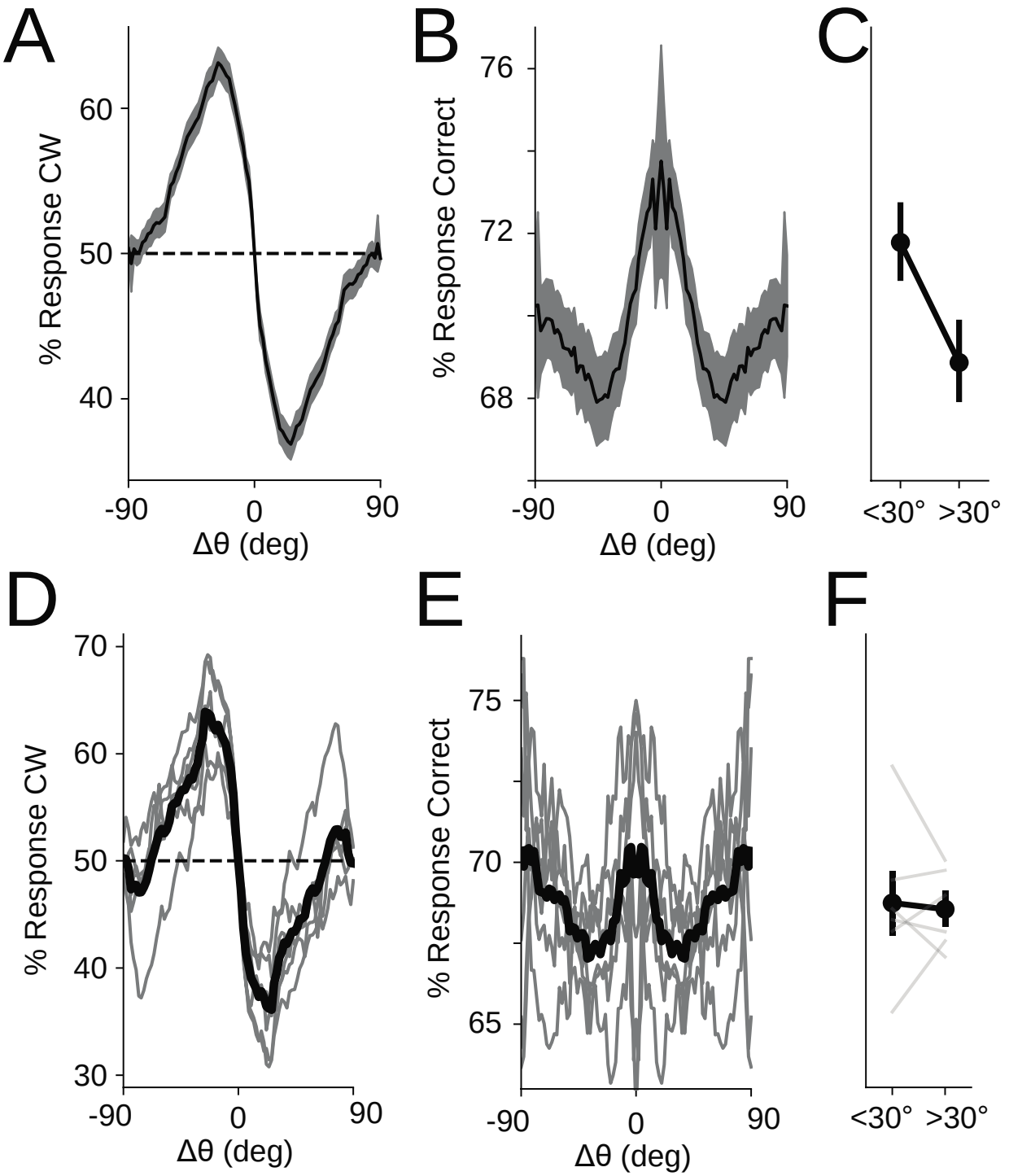
675 found here (GITHUB). Data used in this study will be posted on the first author's Open Science
676 Framework repository (REPOSITORY LINK).

Supplemental Figure 1



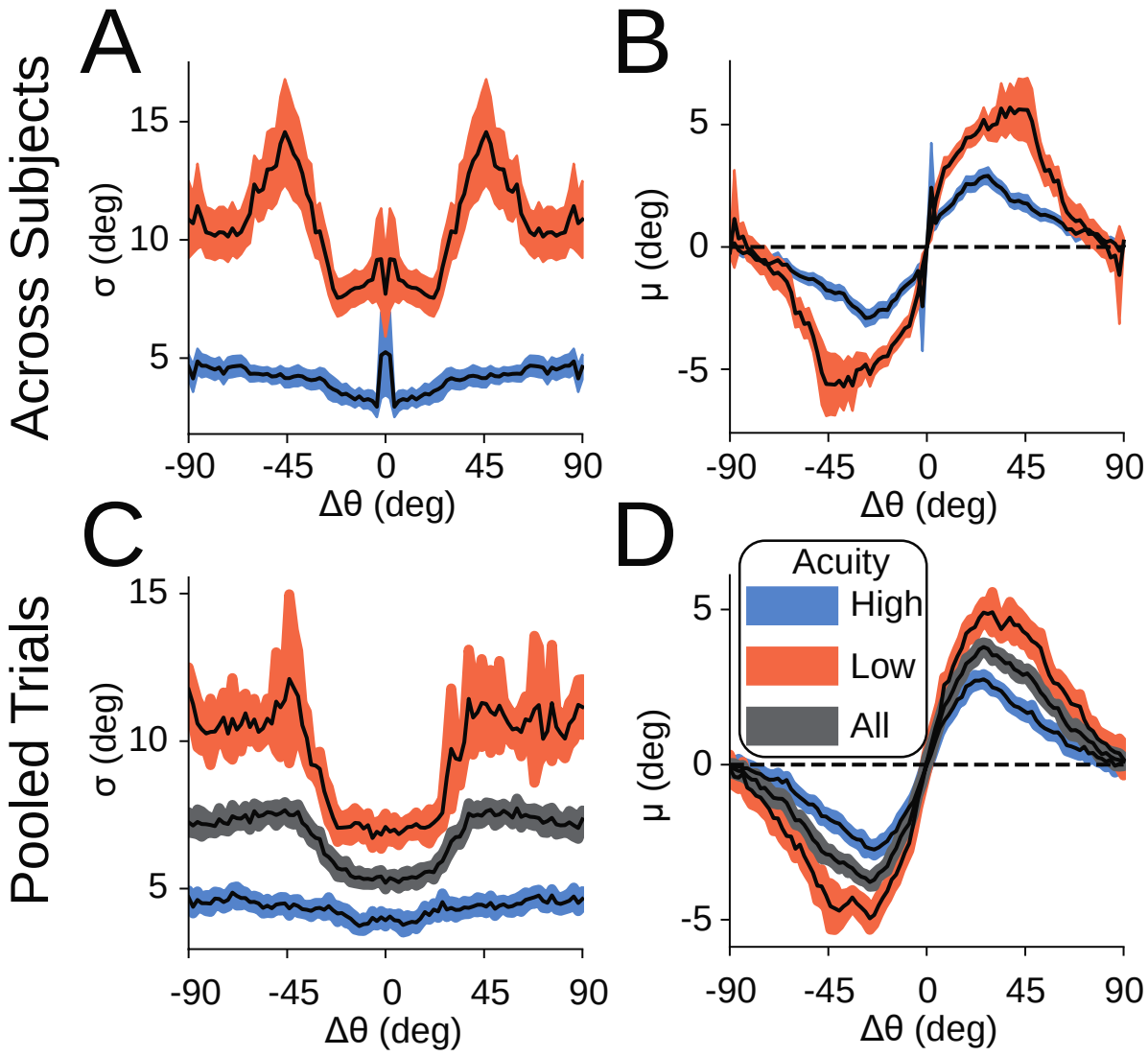
Supplemental Figure 1 Response model. Encoding of stimulus is assumed to be a noisy process whereby the distribution of encoded orientation is described by a Gaussian pdf with mean μ and standard deviation σ. Dashed line is pdf and solid line is the cdf of encoding distribution. Note that participants are reporting the probes orientation relative to the stimulus so more frequent CW responses would correspond to a CCW perceptual bias. **A:** Example estimation curve with no bias and a very small σ. If the difficulty was set to $\delta\theta$=6° (3 sd) than this participant would get essentially all (99.7%) trials correct. **B:** Estimation curve with a μ=-10, this participant would respond CW on almost every trial. **C-D**: Realistic encoding curves. To aid with fitting and to best describe responses, a constant guess rate of 25% was included in the response model fit to participant responses. **C:** An unbiased distribution with two theoretical stimuli on which the participant responded CW. The left response $\delta\theta$=-6° is incorrect. **D:** A CCW biased distribution results in a higher likelihood for all CW responses.

Supplemental Figure 2



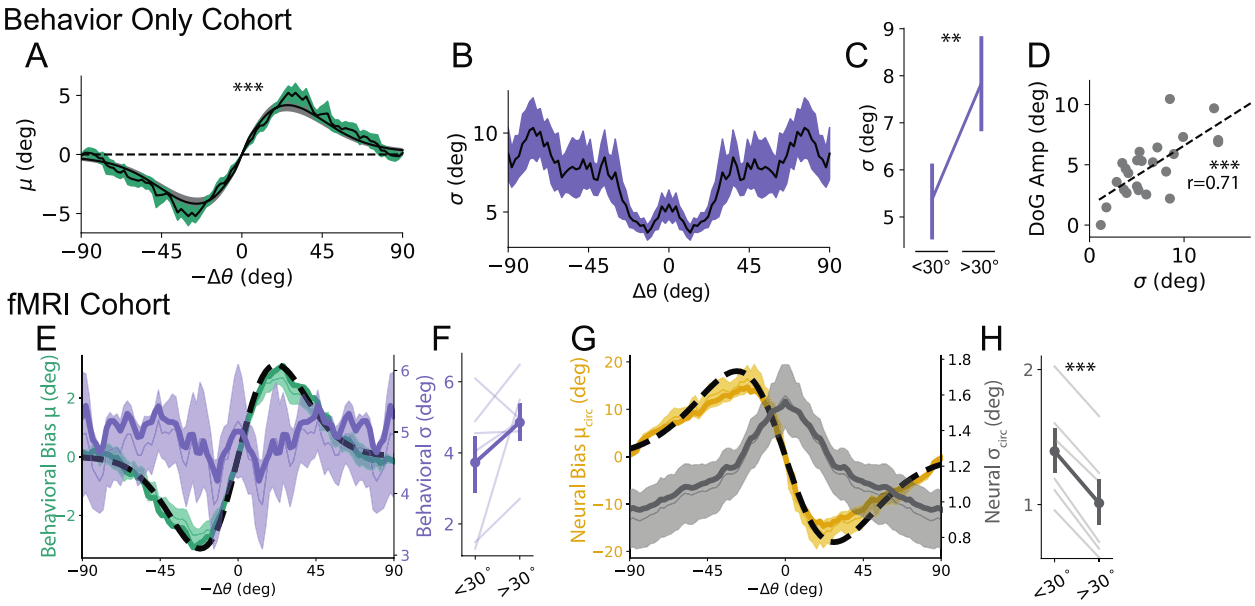
Supplemental Figure 2 Raw responses. **A-C** behavioral participants. **A**: % response CW as a function of Δθ. Note the opposite direction of effect as a CW response means the stimulus was perceived to be CCW of the probe. Shading is SEM across participants **B:** % correct as a function of Δθ. Note that as with all analyses, trials without a response are excluded. **C.** % correct following close or far stimuli. Close sequences led to significantly more correct trials $T(45)=3.54$, $p=.0005$. **D-F** fMRI participants. Analysis the same but showing individual participants. No significant difference for accuracy between close and far stimuli $p=0.40$.

Supplemental Figure 3



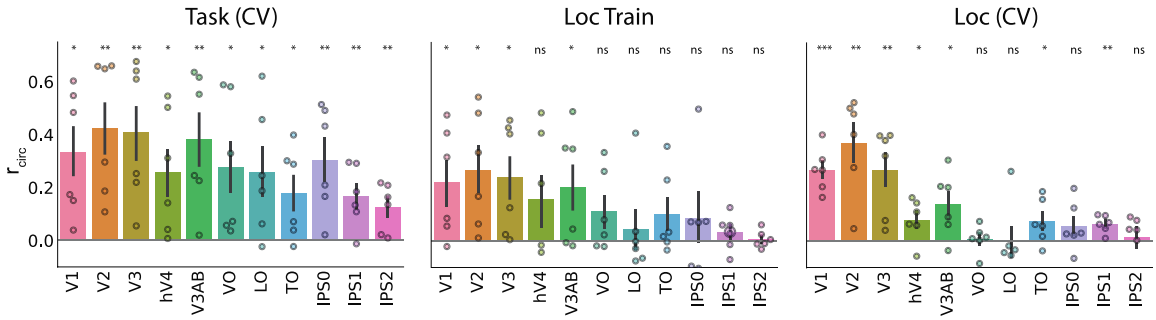
Supplemental Figure 3 Median split bias/variance. **A-B:** average (± SEM across participants) across participants. **A:** Model estimated variance for high and low precision participants. **B.** Model estimated bias is larger for less precise (5.89±0.52°) than more precise (3.57±0.72°) participants, T(44)=2.5, p=.007, unpaired 1-tailed t-test on DoG fits. **C-D:** pooled analysis (± 95% bootstrapped CI). **C:** same as A. Insert shows high acuity participants on own axis. **D:** Same as B.

Supplemental Figure 4



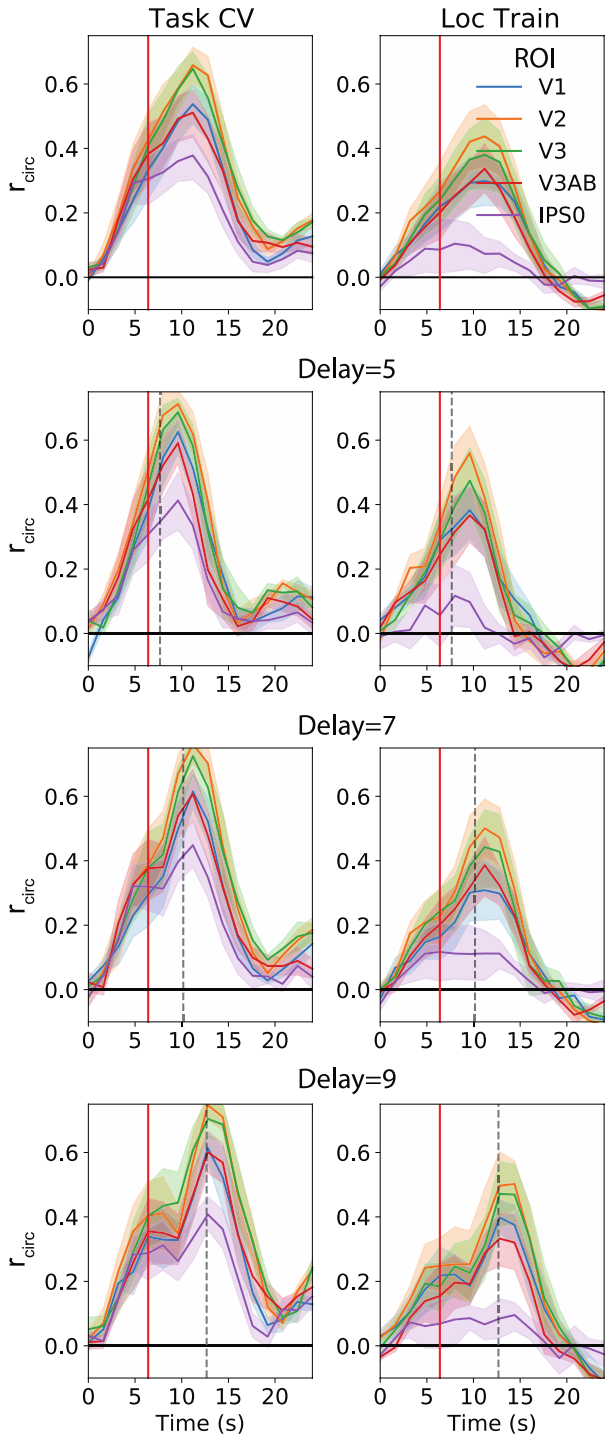
Supplemental Figure 4 Random sequences only. **A-C:** Some participants completed blocks of trials where orientation was not independent across time. Specifically, trials were sorted such that the distance between consecutive stimuli was a multiple of 22.5° (with some jittering). This was intended to maximize our sensitivity to detecting serial dependence as prior experiments in our lab have shown serial dependence typically peaks around 25 degrees. Our behavioral effects ended up being robust and we later opted to just have stimuli be independent across time. Despite no participants overtly noticing any pattern, it is possible that this contrived setup somehow contributed to the behavioral trial history effects that we observed. To assess the impact of this manipulation, we separately analyzed data from only those participants who completed the task with independent stimulus sequences. This cohort (N=25) had an average accuracy of 70.46±1.14° at an average $\delta\theta$ of 4.97±0.35°. **A:** Serial dependence. The average amplitude when parameterized with a DoG was still significantly greater than 0 (amp=4.71±0.49, T(23) = 9.4, p=2.4*10-9; width 0.027±0.0019, FWHM 43.68±1.86°, (mean±SEM). **B-C:** Response variance. Responses were still significantly more precise following similar stimuli t(24)=-2.66, p=0.01. **D:** bias and variance were still positively correlated across participants (r(22)=0.71, p=0.00005. **D-G:** As with our behavior only cohort, some fMRI participants completed blocks of trials where trials were not independent across time. We re-ran a series of control analysis excluding these blocks and found little change to our main findings. **E:** Responses were still systematically attracted to the previous stimulus (DoG Amp: 3.25± 0.34, T(5)=8.85, p=1.53e-04; DoG Width: 36.1±2.9 **F:** Response variance was no longer significantly smaller following small changes but was trending in that direction (T(5)=-1.55, p=.09). **F-G:** Decoded representations showed the same robust pattern of repulsive bias and uncertainty as the full dataset. Together this suggests that our results were not somehow corrupted by the set of trials in which stimuli were not independent.

Supplemental Figure 5



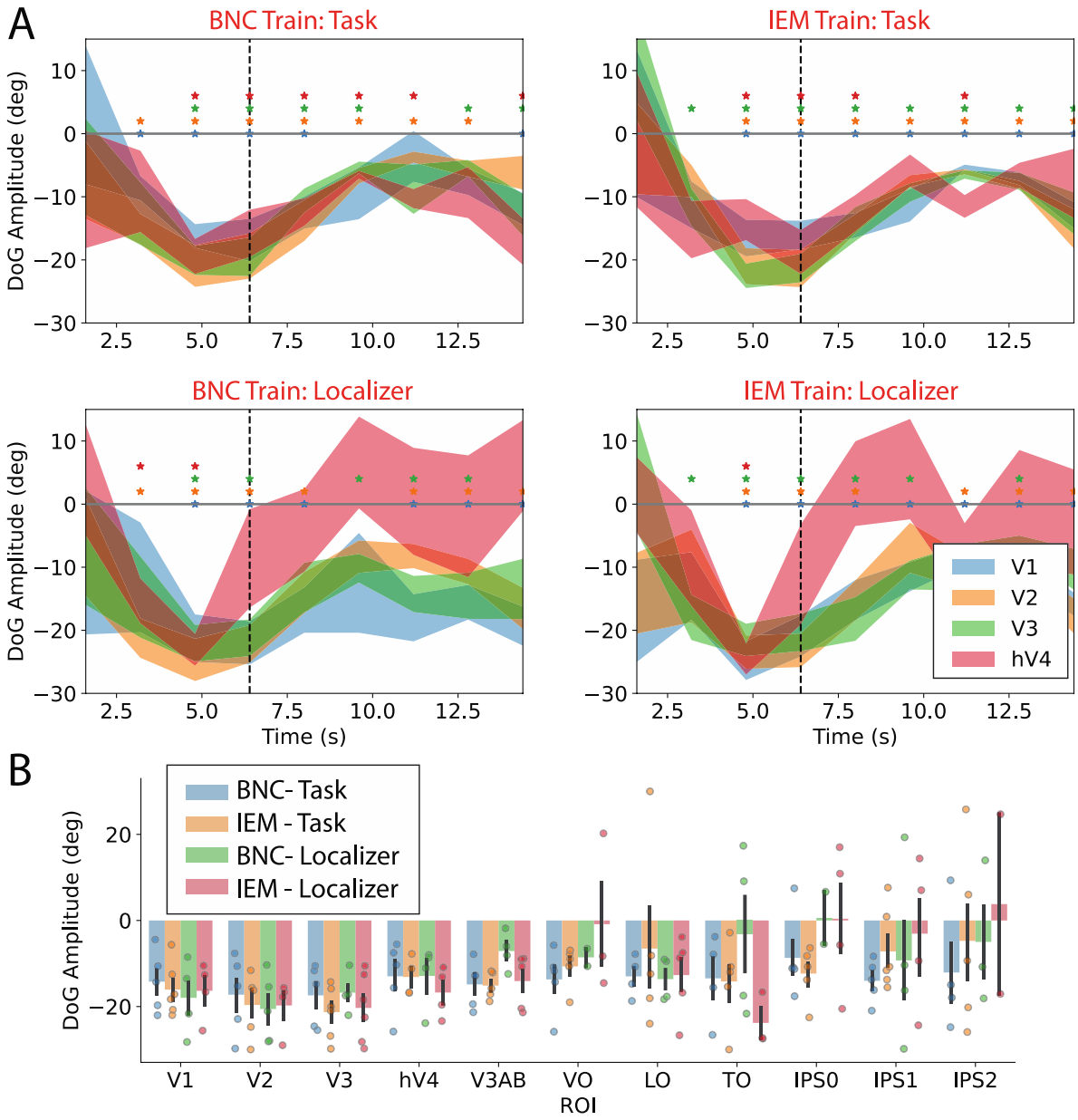
Supplemental Figure 5 Model Performance. Relative performance of model trained and tested on task (**left**), on a separate localizer paradigm and tested on the task (**middle**), or trained and tested within the localizer task (**right**). For the model trained and tested on the localizer data, we could not use orientation information during voxel selection as this would be a circular analysis. Instead, we performed a 75% voxel threshold on donut selectivity for each ROI. See Voxel Selection for selection process for localizers tested on task data. *, p<.05; **, p<.01; ***, p<.001; t-tests on Fisher transformed r_{circ}.

Supplemental Figure 6



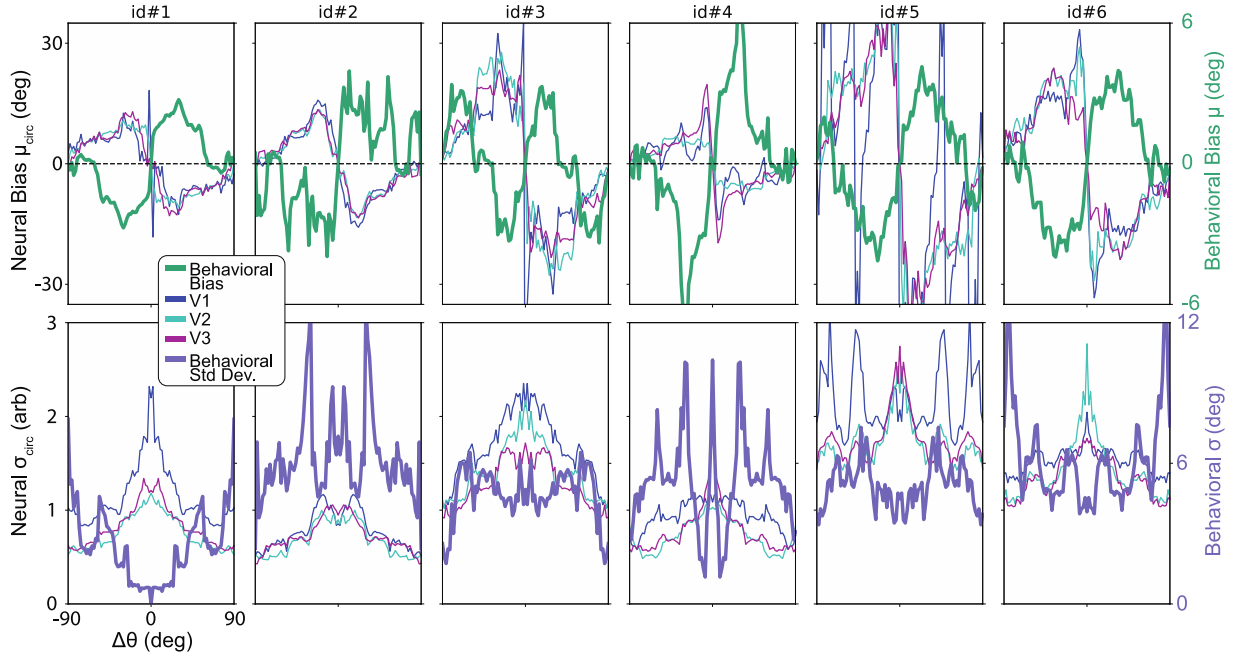
Supplemental Figure 6 Model Performance across time. ***Left column****: model trained on task.* ***Right column:*** *model trained on localizer and tested on all task TRs.* ***Top Row:*** *all trials.* ***Rows 2-4:*** *subset of trials corresponding to delays of 5, 7 and 9s. Shaded lines depict average* r_{circ} *across participants* (± SEM across participants) for 5 ROIs (see legend). Dashed vertical line is average delay time for a given group. Red vertical line is central TR used in main analyses. Time is not shifted to account for hemodynamic lag so even the probe on the shortest delay trials should not affect signal measured at red line.

Supplemental Figure 7



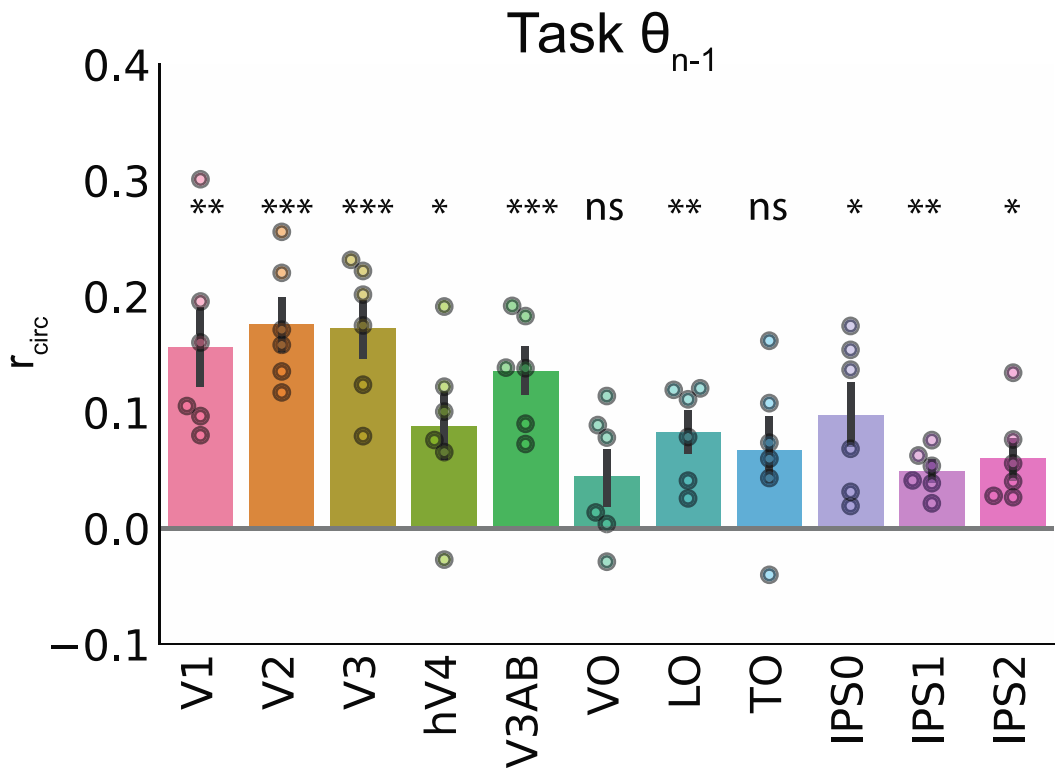
Supplemental Figure 7: Decoded bias across time and ROI. **A:** Decoded bias is significanly repulsive (and never attractive) across the extent of the trial when parameterized with a DoG. For completeness we show two different decoders ('BNC', Bayesian Noise Correlation; and "IEM", Inverted Encoding Model) trained on both the task and localizer data. Time points represent middle of sliding 4 TR window. *, p<.01, uncorrected. **B:** Decoded bias is generally repulsive (and never attractive) across all ROIs and decoding techniques for TR window centered at 6.4s (indicated with dashed line in **A**).

Supplemental Figure 8



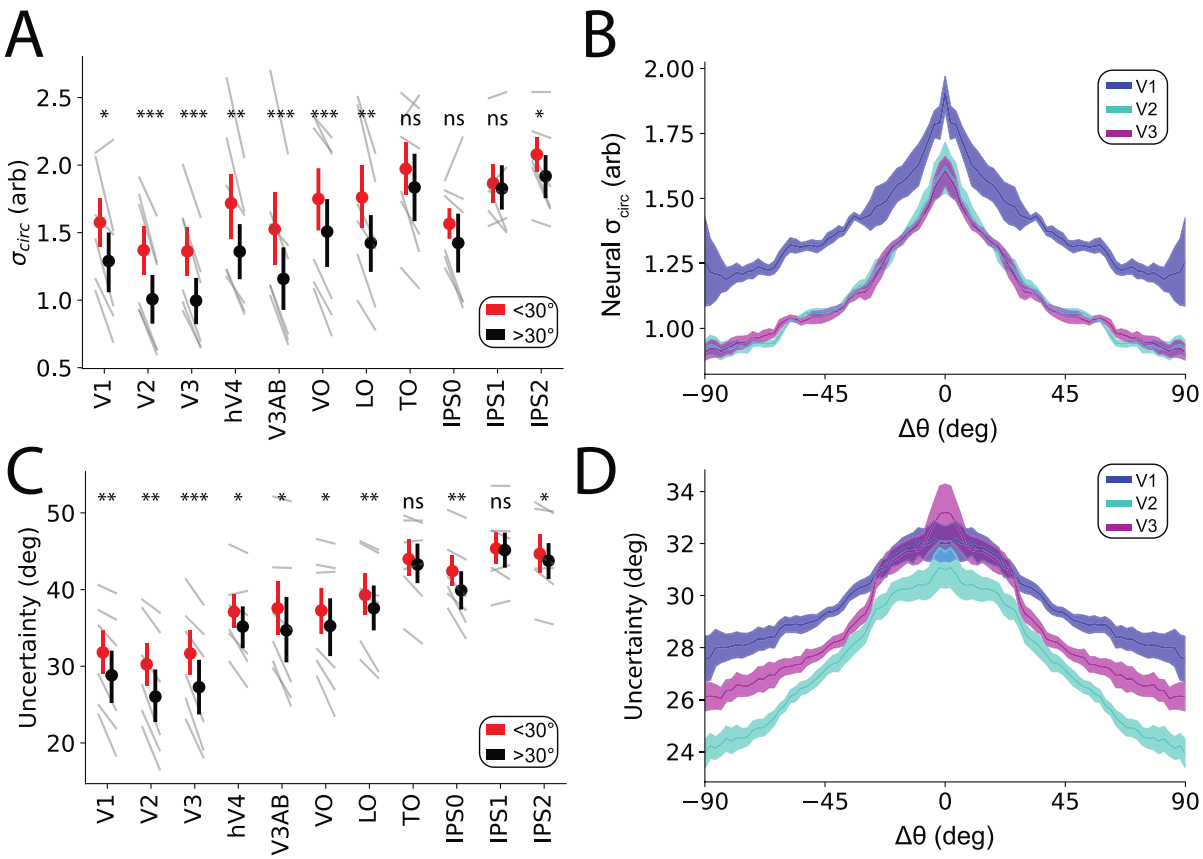
Supplemental Figure 8 Decoded and behavioral bias (**top**) and variance (**bottom**) for individual participants. **Left axis:** Neural data for ROIs V1, V2, and V3 (see **legend**). Decoded orientation is clearly repelled in all participants in V1-V3 and neural σ generally peaks at Δθ=0. **Right axes**: Behavioral data. Responses are clearly attracted for all participants. Note how participant id#3 has peripheral *repulsion* from very distant stimuli.

Supplemental Figure 9



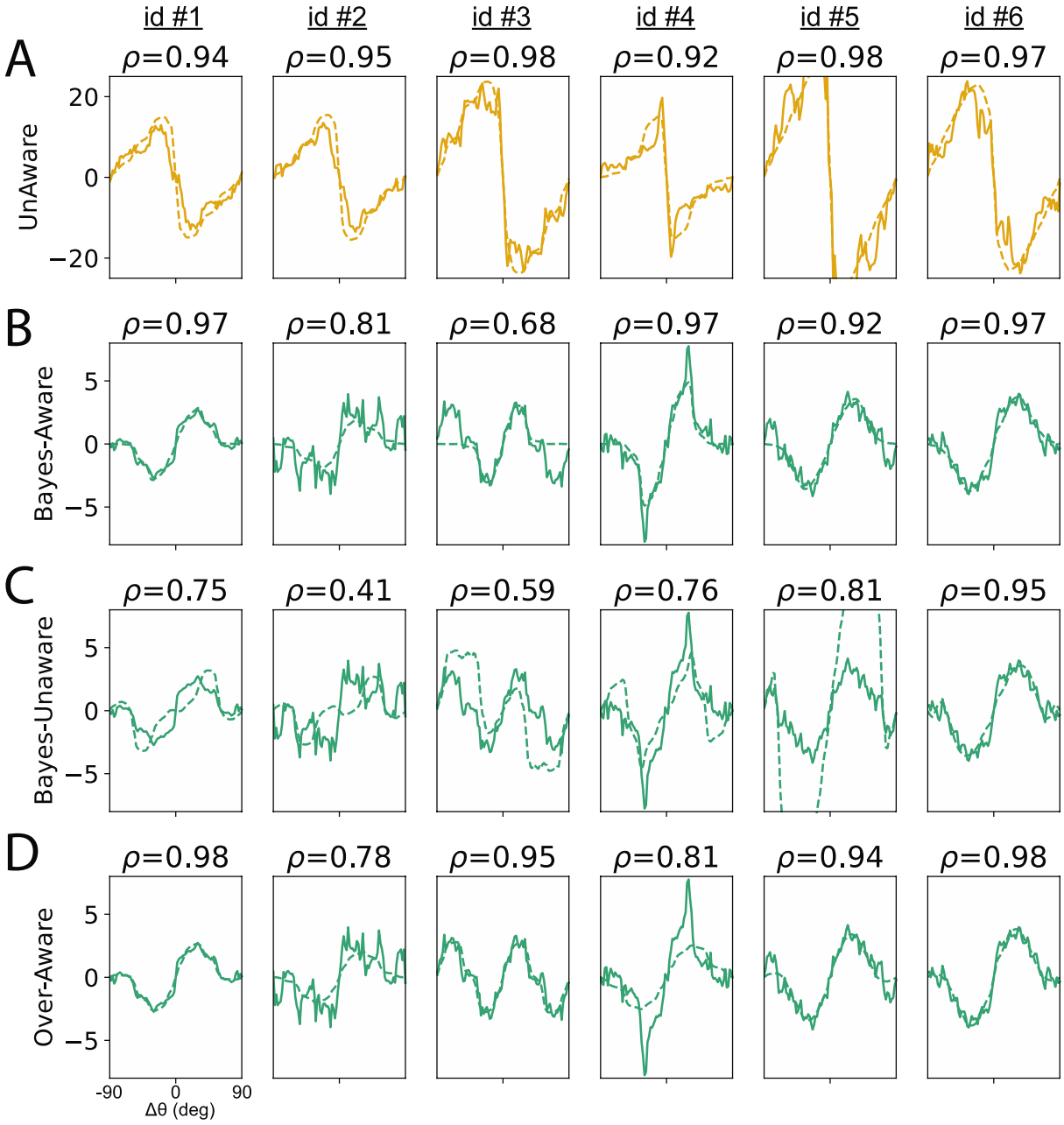
Supplemental Figure 9 Decoding performance for model trained and tested on task data to decode previous trial's stimulus (θ_{n-1}). Performance was significantly above chance in most ROIs. ns, not significant, *, p<.05, **, p<.01, ***, p<.001.

Supplemental Figure 10



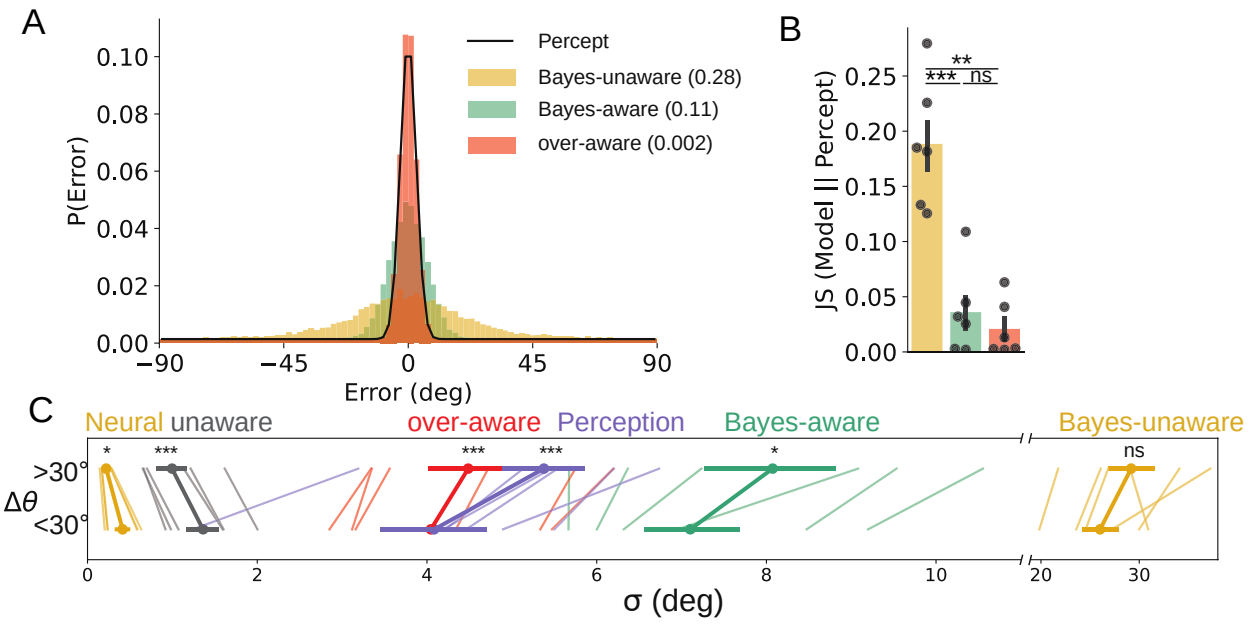
Supplemental Figure 10 Decoded uncertainty as a function of Δθ across ROIs. **A**: σ_{circ} of decoding errors is significantly greater for close (<30°) versus far (>30°) stimuli across early visual ROIs. Points and error bars are mean ±SEM across participants; gray lines depict individual participants. **B:** Sliding σ_{circ} for V1-V3 shows a monotonic relationship. **C-D:** Same as **A-B** but measuring uncertainty directly measured from the single trial posterior (see Neural Variance). Results are qualitatively very similar for both techniques. ns, not significant, *, p<.05, **, p<.01, ***, p<.001.

Supplemental Figure 11



Supplemental Figure 11 Model fits for individual participants (same order as Figure S8). Solid lines correspond to empirical neural (**yellow**) or behavioral (**green**) bias; dashed lines correspond to model fits to BOLD decoding bias (Unaware model, **A**) or behavior (**B-D**). Model fits plotted are average of noiseless biases generated by models fit to each CV fold. Note that a models are fit to raw data, not binned data presented here. Pearson's correlations are reported above each fit between binned and model estimated bias.

Supplemental Figure 12



Supplemental Figure 12 Model Performance. **A:** Distribution of empirically predicted response errors (black line) and simulated model fits for an example participant along with associated Jensen-Shannon divergences. **B:** The *Bayes-unaware* model provided a significantly worse fit to empirical uncertainty than either "aware" model when assessed across participants. **C**. Visualization of all uncertainties split as a function of close and far stimuli. The *unaware* model was significantly less precise following small changes matching neural decoding. The two "aware" models were significantly more precise following small changes matching perception. The Bayes-unaware model did not have significant modulation of decoding uncertainty and had an average uncertainty that was on average 5x that of perception. ns, not significant, *, p<.05, **, p<.01, ***, p<.001.

Table 1

Average fit coefficients for models averaged across CV fits (± SEM across participants) shown in bold. Other parameters either fixed values, drawn from fit to encoding stage for given participant, or are not utilized for a particular model (N/A).

	Fit To:	BOLD Decoder	Behavior		
Stage:		**Unaware**	**Bayes unaware (Prior*unaware)**	**Bayes aware (Prior*aware)**	**Over-aware**
Encoding	γ_m	**0.81 ±0.04**			
	γ_s	**0.56±0.16**			
Decoding	γ_{m2}	0	0	γ_m	**0.69±0.09**
	γ_{s2}	1	1	γ_s	**0.68±0.11**
Bayes	R	5	**1.95±0.57**	**0.17±0.03**	5
	ψ	N/A	**0.60±0.05**	**0.86±0.05**	N/A