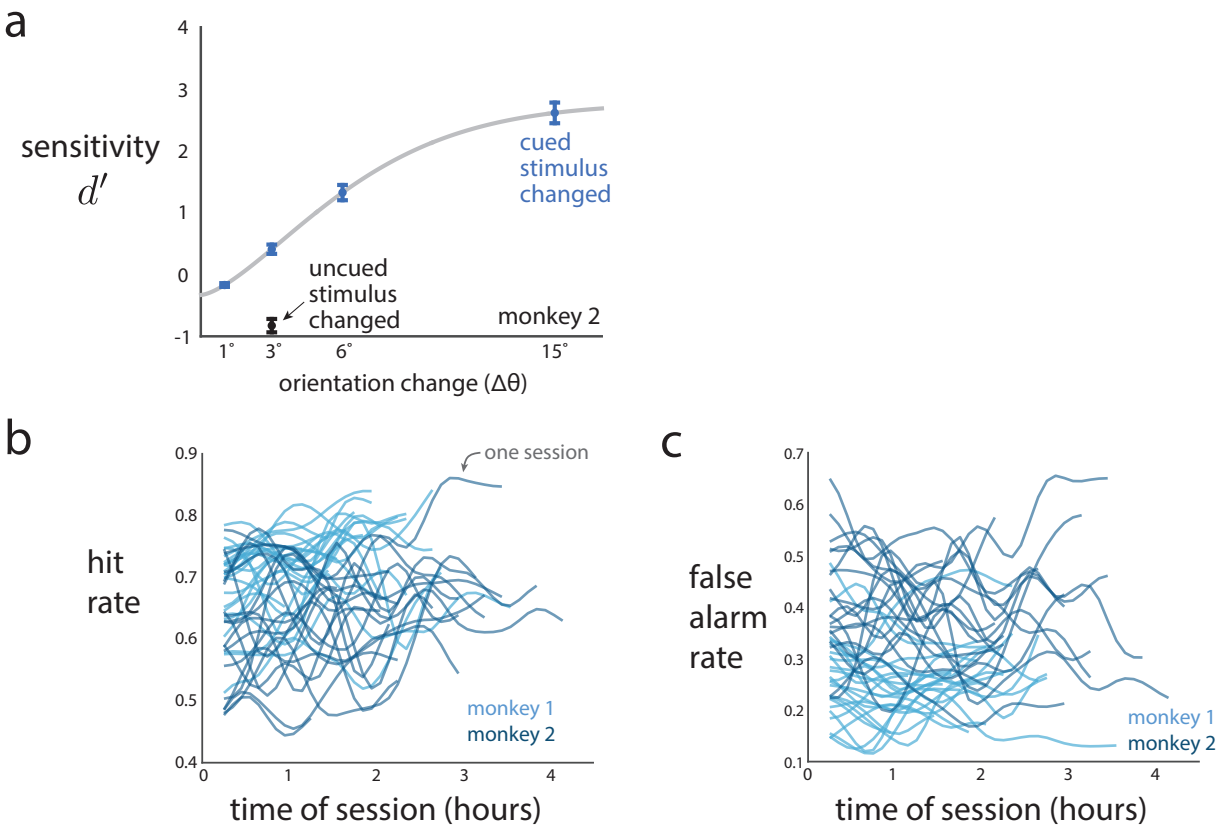
Neuron, Volume *108*

Supplemental Information

Slow Drift of Neural Activity as a Signature of Impulsivity in Macaque Visual and Prefrontal Cortex

Benjamin R. Cowley, Adam C. Snyder, Katerina Acar, Ryan C. Williamson, Byron M. Yu, and Matthew A. Smith



Supplementary Figure 1: Behavior during the orientation change detection task. Related to Figure 1.

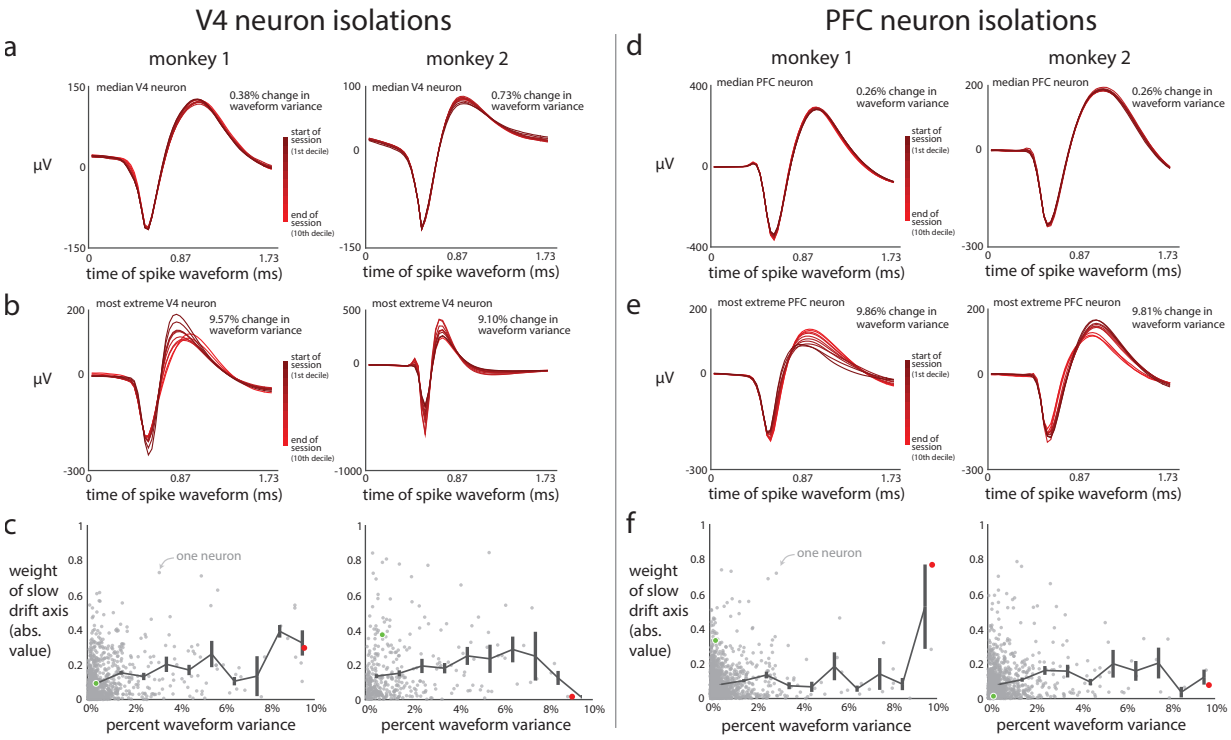
a. The animal better detects larger stimulus changes as well as stimulus changes that are cued than uncued. Task performance (measured as sensitivity d') versus the change in orientation $\Delta\theta$ (either clockwise or counterclockwise) of the cued stimulus (blue) and the uncued stimulus (black). Results shown here are for monkey 2 (see Fig. 1**b** for results of monkey 1). Larger $\Delta\theta$ are easier for the animal to detect, and cued stimulus changes ($\Delta\theta = 3°$, blue) are easier to detect than uncued stimulus changes ($\Delta\theta = 3°$, black). Data are fit with a Weibull function (gray). Dots indicate means over sessions, error bars indicate ± 1 s.e.m.

We wondered if the time course of behavior was stereotypical across sessions (e.g., both hit rate and false alarm rate slowly increased for all sessions). We found that hit rate and false alarm rate slowly changed over the course of the session but not in a consistent way across sessions (e.g., all rates do not strictly increase).

b. Hit rate, measured as the fraction of stimulus changes that were correctly detected.

c. False alarm rate, measured as the fraction of unchanged stimulus presentations to which the animal incorrectly made a saccade. For **b** and **c**, each line is the running estimate of the rate for one session (see STAR Methods).

The results in **b** and **c** indicate that hit rate and false alarm rate did not fluctuate in a stereotypical manner (e.g., strictly increasing). This suggests that the slow changes in rates were not due to a general increase in fatigue (or satiation, etc.) over the course of the session. The reasoning is as follows. As the animal gradually becomes fatigued, the animal likely becomes less engaged in the task and more prone to "guessing"—increasing the animal's overall likelihood to make a saccade in order to receive reward for guesses that happened to be correct. Thus, if changes in rates were due to fatigue or satiation, we would expect to see both hit rate and false alarm rate gradually increase throughout the session. However, hit rate and false alarm rate did not strictly increase for every session. Still, it is possible that the animal had differing periods of fatigue and high engagement throughout each session, consistent with a fluctuating hit rate and false alarm rate.



Supplementary Figure 2: Ensuring the recording stability of the neuron isolations. Related to Figure 2.

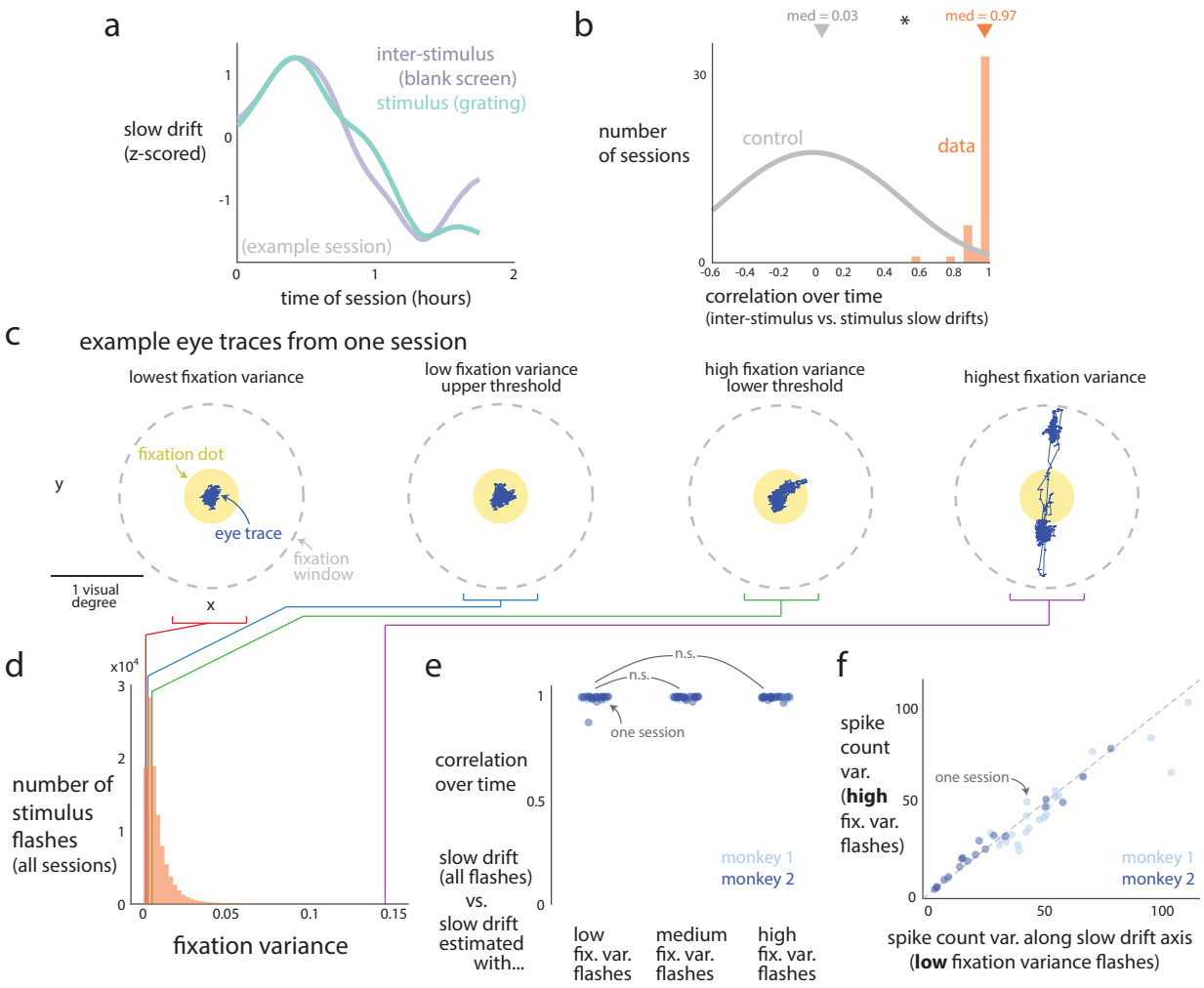
One potential source of the slow drift could have been recording instability, which might have caused the recorded spike waveforms to gradually change during a session. A changing spike waveform could have affected our spike sorting procedure, leading to a spurious slow drift in neural activity. To ensure this was not the case, we assessed how much a neuron's spike waveform changed throughout a session with the percent waveform variance metric (see STAR Methods).

a. Spike waveforms for V4 neurons with the median percent waveform variance for each monkey. Each trace represents the average waveform for one decile of the session.

b. Same conventions as in panel **a** except for V4 neurons we analyzed with the largest percent waveform variance (neurons with a percent waveform variance greater than 10% were removed from our analyses).

c. Having removed neurons with percent waveform variance greater than 10%, we performed two analyses to verify that the identified slow drift was not a product of recording instabilities in the remaining neurons. First we asked whether neurons with highly stable waveforms (i.e., percent waveform variance < 1%) contributed to the slow drift. We found many neurons with a very low percent waveform variance but large weight magnitude (i.e., there are many gray dots with < 1% waveform variance and > 0.2 weight magnitude). Thus, even the most stably-recorded neurons contribute substantially to the identified slow drift. Second, we binned the percent waveform variances into 1% bins, and computed the mean weight magnitude across neurons for each bin (black line). This mean value was relatively flat across bins, suggesting neurons at all percent waveform variance levels contributed to the identified slow drift. Black error bars indicate ± 1 s.e.m., and the green and red dots denote the median and most extreme neurons, respectively. These results indicate that the identified V4 slow drift in neural activity could not have been caused by neural recording instabilities.

d-**f**. PFC neurons were subjected to the same neuron removal criteria ($<$ 10% waveform variance) as were the V4 neurons. All conventions were the same as in **a**-**c**. These results indicate that any slow drift observed in the activity of the analyzed PFC neurons was not caused by neural recording instabilities.



Supplementary Figure 3: One possible source of the slow drift in V4 activity is the change in rate of small eye movements, known as microsaccades, during fixation. Related to Figure 2.

These microsaccades may cause changes in V4 activity either by making small shifts to the visual stimulus relative to V4 receptive fields or via corollary discharge. Here, we found this not to be the case. Instead, the slow drift, as an impulsivity signal, may influence the rate of microsaccades, consistent with our findings that the slow drift correlates to other behavioral variables (Fig. 3).

a. If the slow drift arose from the feedforward effects of microsaccades (i.e., shifting the visual stimulus relative to V4 receptive fields), then we reasoned that when no stimulus was being presented (i.e., a blank screen except for the fixation dot), we should find little to no slow drift. To test this, we compared estimates of the slow drift when a stimulus was being presented (green, 'stimulus') and when no stimulus was presented (purple, 'inter-stimulus'). The former was the same slow drift as that estimated in Figure 2, while the latter was estimated in the same manner but using V4 spike counts (taken in a 200 ms window) starting 200 ms before stimulus onset (during which no stimulus was presented and 100 to 300 ms after the previous stimulus offset). For this example session, we found a strong correlation over time ($\rho = 0.96$, $p < 0.002$, permutation test) between the slow drift during stimulus presentation (green) and during the inter-stimulus period (purple).

b. Correlation over the course of the session between the slow drift during stimulus presentation and during the inter-stimulus period, all sessions ('data', orange). The median correlation of the data ($\rho = 0.973$, aggregated for both monkeys) was significantly higher ($p < 0.002$, permutation test) than that between two smooth, random walks with the same time scale as the slow drift (median $\rho = 0.032$, see STAR Methods). (continued on next page...)

Supplementary Figure 3: (...continued from previous page)

Results for individual monkeys held the same trends (median $\rho = 0.97, 0.97$ for monkeys 1, 2). We also confirmed that the sizes of the slow drifts (cf. Fig. 5**c**) were significantly correlated across sessions ($\rho = 0.46, 0.71$ for monkey 1, 2; $p < 0.05$ for each monkey, permutation test). These results indicate that the slow drift is still present in V4 activity even when no stimulus is present. This suggests that the slow drift does not arise in a feedforward manner from shifts of the visual stimulus relative to the V4 receptive fields caused by microsaccades.

c. To further test if microsaccades influence the slow drift, we measured to what extent the eyes moved during each stimulus flash. We defined the measure 'fixation variance' as the variance of the eyes' locations in visual degrees over time during fixation. For an example session, the stimulus flash with the lowest fixation variance had a stable eye trace (left panel, blue trace), while the flash with the highest fixation variance had salient microsaccades (right panel, blue trace visits two distinct spatial locations). The fixation cue dot is in yellow, and the fixation window (the boundary that if crossed, indicated a break in fixation) is a gray dashed line.

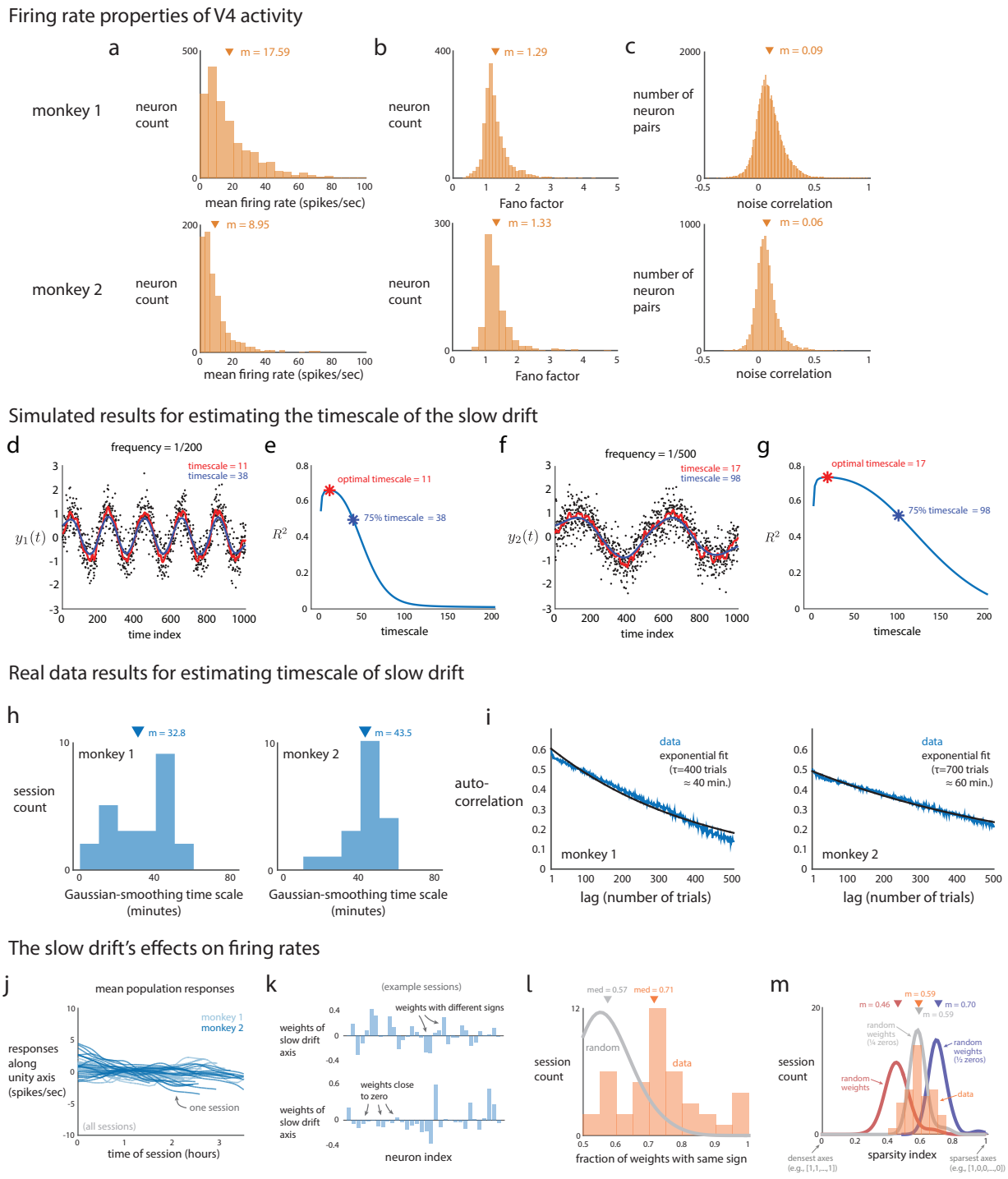
If the slow drift arose from microsaccades, then we expected that we should not find evidence of the slow drift in V4 activity for stimulus flashes with low fixation variance (i.e., when little to no eye movement occurred). We grouped stimulus flashes into those with a low fixation variance (below a low fixation variance upper threshold), a high fixation variance (above a high fixation variance lower threshold), and a medium fixation variance (between the two thresholds). The thresholds were chosen such that a third of stimulus flashes had low fixation variance (second panel from the left, example eye trace for a stimulus flash with fixation variance equal to the upper threshold for low fixation variance flashes), while a third of stimulus flashes had a high fixation variance (third panel from the left, example eye trace for a stimulus flash with fixation variance equal to the lower threshold of high fixation variance flashes). For visual clarity, example eye traces were re-centered.

d. The fixation variances across stimulus flashes for both monkeys were mostly small, while flashes with microsaccades (i.e., fixation variance > 0.1) were rare. Fixation variances corresponding to the four example flashes from one session in **c** are indicated with colored lines.

e. We asked whether the slow drift in V4 activity was present during stimulus flashes with low fixation variance. If microsaccades influence the slow drift (in a feedforward manner or as a corollary discharge), then there should be no slow drift present for stimulus flashes with low fixation variance. We estimated the slow drift in V4 activity for all flashes, and compared this slow drift to a slow drift estimated with a third of the flashes with low fixation variance ('low fix. var. flashes'). These slow drifts were highly correlated over the course of the session ('low fix. var. flashes', mean $\rho = 0.993$ across sessions and monkeys). We also estimated the correlations between the slow drift (estimated with all flashes) and the slow drift estimated with either medium fixation variance flashes ('medium fix. var. flashes', mean $\rho = 0.996$) or high fixation variance flashes ('high fix. var. flashes', mean $\rho = 0.996$). These distributions of correlations were not significantly larger than those for the low fixation variance flashes ('n.s.' comparisons, $p = 0.368$, $p = 0.386$ for medium and high fixation variance flashes, respectively, permutation test). Thus, we found that the slow drift was present during stimulus flashes with low fixation variance, supporting the idea that the slow drift does not arise from microsaccades either in a feedforward manner or as a corollary discharge.

f. We also asked whether V4 activity along the slow drift axis more strongly varied for low fixation variance flashes versus high fixation variance flashes. If microsaccades influence the slow drift, then we expect that the variance of V4 activity along the slow drift axis for low fixation variance flashes should be less than that for high fixation variance flashes. This is because activity for low fixation variance flashes should vary little (as little to no eye movement occurs) while activity for high fixation variance flashes would vary greatly due to the microsaccades. We found that the variance of residual V4 activity (raw spike counts minus the repeat-averaged response) projected along the slow drift axis (estimated with all flashes) for high fixation variance flashes was no larger than that for low fixation variance flashes (dots not above dashed line, $p = 0.668$, permutation test).

Taking the results in **b**, **e**, and **f** together, we conclude that the slow drift does not arise as a consequence of changes to the rate of microsaccades, either in a feedforward manner or as a corollary discharge. However, the slow drift, as an impulsivity signal, likely influences the rate of microsaccades, which we speculate about in the Discussion.



Supplementary Figure 4: Firing rate properties and slow drift of V4 activity. Related to Figure 2 and STAR Methods. (continued on next page...)

Supplementary Figure 4: (...continued from previous page)

Properties of V4 neural activity were consistent with those observed in previous studies (reviewed in Cohen & Kohn, 2011). Spike counts were taken in a 400 ms bin starting 50 ms after stimulus onset. Results for monkey 1 (left column) and monkey 2 (right column) were computed separately for each stimulus orientation (i.e., 45° and 135°) and aggregated across orientations and sessions. **a**. Mean firing rate. **b**. Fano factors (spike count variance divided by mean spike count). **c**. Noise correlations for each pair of simultaneously-recorded neurons. Triangles indicate means ('m').

Previous work has also found that noise correlations between pairs of V4 neurons tend to decrease when the animal employs attention (Cohen & Maunsell, 2009; Mitchell et al., 2009). We confirmed that for our recorded V4 neurons, the mean noise correlation decreased between valid (i.e., 'attended') and invalid (i.e., 'unattended') trials (Δ mean $r_{\text{sc}} = -0.014, -0.010$ for monkey 1, 2). The magnitudes of these changes were consistent with those in previous work (Cohen & Maunsell, 2009; Mitchell et al., 2009).

We asked to what extent hit rate, false alarm rate, and the slow drift slowly evolved over time. Thus, we sought a way to estimate the timescales of these variables. Here, we provide intuition of our approach for estimating timescale with simulated data.

d. We simulate a signal $y_1(t) = \sin(2\pi \frac{1}{200} t) + \epsilon$, where $t = 0, \ldots, 1000$ represents time (e.g., in minutes) and $\epsilon \sim \mathcal{N}(0, \sigma)$ for $\sigma = 0.5$ is the noise that is independent at each time point. We simulated $1,000$ time points of this signal (black dots). To estimate this signal's timescale, we used Gaussian smoothing and considered different standard deviations as candidate timescales (red and blue lines denote smoothed estimates with different timescales). We chose this approach over a Fourier analysis, because this approach allows for the unequal spacing between time points, which occurred in the neural data (i.e., trials were initiated by the animal and inter-trial periods need not have the same durations).

e. For each candidate timescale, we computed the cross-validated performance R^2 by predicting each held-out data point by a Gaussian weighted average of its neighboring data points (held-out data were chosen randomly for 10 fold cross-validation). We found that the candidate timescale that achieved the largest R^2 (the optimal timescale, red asterisk) was small, and that a much longer timescale still achieved 75% of the largest R^2 (blue asterisk). This suggests that although a small candidate timescale (in this case, 11) may be optimal for Gaussian smoothing (because it can capture both low and high frequency components), it does not necessarily reflect the timescale of the most dominant component (e.g., a low frequency of $\frac{1}{200}$). Instead, the timescale of this dominant, low-frequency component is reflected by how slowly the R^2 curve drops off as the candidate timescale increases beyond the optimal timescale. Indeed, the 75% timescale of 38 better reflects the low frequency of $\frac{1}{200}$. We note that the timescale of 38 reflects the window of the Gaussian weighted average and is *not* expected to be equal to the period of the sinusoid (in this case, 200).

f. To confirm this intuition, we simulated $y_2(t)$ in the same way as $y_1(t)$ except with a lower frequency of $\frac{1}{500}$. We expected a slower drop off in R^2 for $y_2(t)$ than $y_1(t)$.

g. We found that the optimal timescale (red asterisk) for $y_2(t)$ was not representative of the dominant frequency of $\frac{1}{500}$. However, the 75% timescale of 98 did reflect this low frequency, indicating the drop in R^2 was slower for a signal with a lower frequency (compare blue curves in **b** and **d**), as expected. Thus, we choose to report the candidate timescale that achieves 75% of the largest R^2 for two reasons. First, this timescale appears to be a more interpretable estimate of the timescale of the most dominant component of the signal (e.g., a timescale of 98 better represents $y_2(t)$ in **c** than a timescale of 17). Second, this timescale is more likely to avoid the preference of Gaussian smoothing to choose smaller candidate timescales as the optimal. We only used this approach to report the timescales in the main text for hit rate, false alarm rate, and the slow drift; the estimated timescales do not affect results presented in any other figure.

(continued on next page...)

Supplementary Figure 4: (...continued from previous page)

Using the above method to estimate timescale (**d**-**g**), we found that the timescale of the V4 slow drift was approximately 40 minutes.

h. We estimated the timescale of the V4 slow drift with Gaussian smoothing (same estimation procedure as that for estimating the timescale of hit rate and false alarm rate, see STAR Methods). We found that this timescale for the slow drift of V4 activity (i.e., V4 responses projected onto the slow drift axis) was approximately 40 minutes for each monkey (mean $m = 32.8$, 43.5 minutes for monkeys 1, 2).

i. We also computed the timescale of the slow drift using a different method. We performed an autocorrelation on the responses projected onto the slow drift axis (blue, averaged over sessions). We fit an exponential function (black) to the autocorrelation curve (blue), and found decay rates of $\tau = 400, 700$ trials for monkeys 1, 2. On average, 100 trials took 10.2, 9.4 minutes for monkeys 1, 2, indicating that the timescale of the slow drift was ~50 minutes (~40, ~60 for monkeys 1, 2). One caveat is that the autocorrelation requires that responses have equal spacing between time points, but this was not the case for our data: trials could be separated by different lengths in time. Taken together, the results of the Gaussian smoothing analysis and the autocorrelation analysis suggest that much of the response variation can be explained by a slow drift with a timescale of 40 minutes or greater.

We further asked how each neuron contributes to the slow drift. Because the slow drift represents a weighted average across neurons, the effects of the slow drift may be missed if one simply averages firing rates across neurons (i.e., equal weights) if the weights of the slow drift are not equal (i.e., both positive and negative weights). We found this to be the case.

j. Time courses of the mean population rate for all sessions. The mean population rate is a proxy for the number of On/Off states (Engel et al., 2016) or synchronized/desynchronized states (Beaman et al., 2017) and has been used previously to summarize population activity (Okun et al., 2015). The mean population rate was computed by projecting population activity onto the unity axis $\mathbf{u} = \frac{1}{N}[1, 1, \ldots, 1]$ for N neurons and then Gaussian smoothed. The slow drift was processed in the same manner except that the slow drift axis was identified with PCA (Fig. 2**d**). The slow drift varied substantially more than the corresponding mean population rate ($\sigma^2_{\text{slow drift}}/\sigma^2_{\text{pop. rate}}$ median = $6.9, 2.9$ for monkeys 1, 2 and were significantly above 1, $p < 0.05$ for both monkeys, paired permutation test). In addition, the slow drifts were not correlated with their corresponding mean population responses (mean $\rho = 0.04 \pm 0.16, -0.04 \pm 0.21$ for monkeys 1, 2, where ± 1 indicates 1 s.e.m.). These results indicate that the slow drift is not apparent when directly averaging the activity across neurons. Furthermore, the slow drift does not appear to be directly related to the observed On/Off states or synchronized/desynchronized states observed in previous work (Engel et al., 2016; Beaman et al., 2017).

k. Weight vectors of the slow drift axis for two example sessions. In the following two analyses, we asked whether the weights had the same signs and how many weights were close to zero.

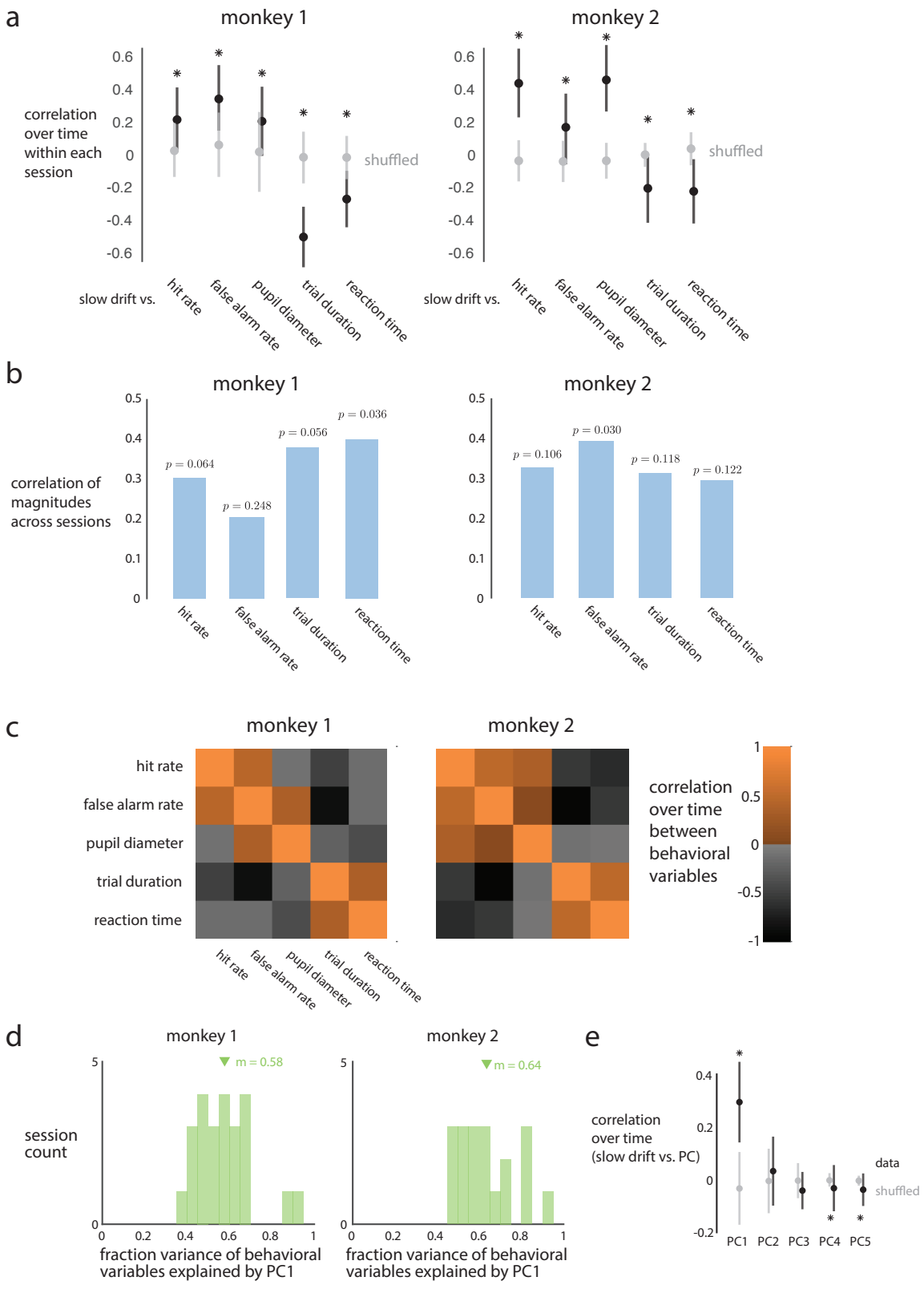
l. We assessed to what extent the slow drifts of neurons were positively or negatively correlated by analyzing the signs of the weights. If some neurons were negatively correlated, it suggests that taking the population average may miss large effects of the slow drift. We found this to be the case, as only 70% of neurons had weights with the same sign. We found this by computing the fraction of weights with the same sign for each session. A fraction of 1 indicates that all neurons have weights with the same sign. A fraction of 0.5 indicates that half the neurons have positive weights, while the other half have negative weights. The median fraction for all sessions was 0.71 ('data'). For reference, the distribution of fractions for weights whose signs were randomly flipped with equal probability ('random', smoothed histogram) had a median of 0.57.

(continued on next page...)

Supplementary Figure 4: (...continued from previous page)

m. Another important consideration was to what extent the weights of the slow drift axes were sparse (i.e., how many weights had values far from zero). This is important because it could have been the case that only 1 to 2 neurons had large weights and the rest of the weights were close to zero (i.e., high sparsity), implying that the slow drift is not a population effect but rather an effect of a small number of specific neurons. However, this was not the case, as we found that roughly 75% of the recorded neurons slowly drifted by performing the following analysis. We defined a sparsity index for slow drift axis $\mathbf{s} \in \mathbb{R}^N$ for N neurons as the angle between $[|\mathbf{s}_1|, \ldots, |\mathbf{s}_N|]$ (where $\mathbf{s}_i$ is the ith element of $\mathbf{s}$ and $|\cdot|$ is the absolute value function) and $[1, 1, \ldots, 1]$ (i.e., the unity axis) divided by the maximally-sparse angle. The maximally-sparse angle is the angle between the $[1, 1, \ldots, 1]$ axis and the $[1, 0, \ldots, 0]$ axis (i.e., a unit axis). A sparsity index of 1 indicates that $\mathbf{s}$ is highly sparse with many weights that are close to zero. A sparsity index of 0 indicates that $\mathbf{s}$ is dense (i.e., many weights far from zero). For reference, we computed the sparsity index for randomly-generated vectors where each weight was drawn from a standard Gaussian (red). We computed two other reference distributions whose vectors were generated in the same way as that for the red distribution but a fraction of weights were forced to be zero (gray: $\frac{1}{4}$ of weights are zero, blue: $\frac{1}{2}$ of weights are zero). The distribution of sparsity indices for the slow drift axes (orange, 'data') overlapped the most with the $\frac{1}{4}$-weights distribution (gray), indicating that the activity of roughly 75% of the recorded neurons slowly drifted.

Taking **l** and **m** together, we conclude that ~50% of neurons increase or decrease their activity together, ~25% either increase or decrease their activity in the opposite manner of the first 50%, and ~25% have little to no drift.



Supplementary Figure 5: The slow drift covaries with slow fluctuations in behavior. Related to Figure 3.
(continued on next page...)

Supplementary Figure 5: (...continued from previous page)

We found similar trends for individual monkeys as aggregated results (Fig. 3**b** and **c**). Same plotting conventions as in Figure 3**b** and **c**.

a. Correlations over time within each session between the V4 slow drift and behavioral variables for each monkey. Correlations match those of aggregated results (Fig. 3**b**). Asterisk denotes significance over chance levels ($p < 0.05$, permutation test).

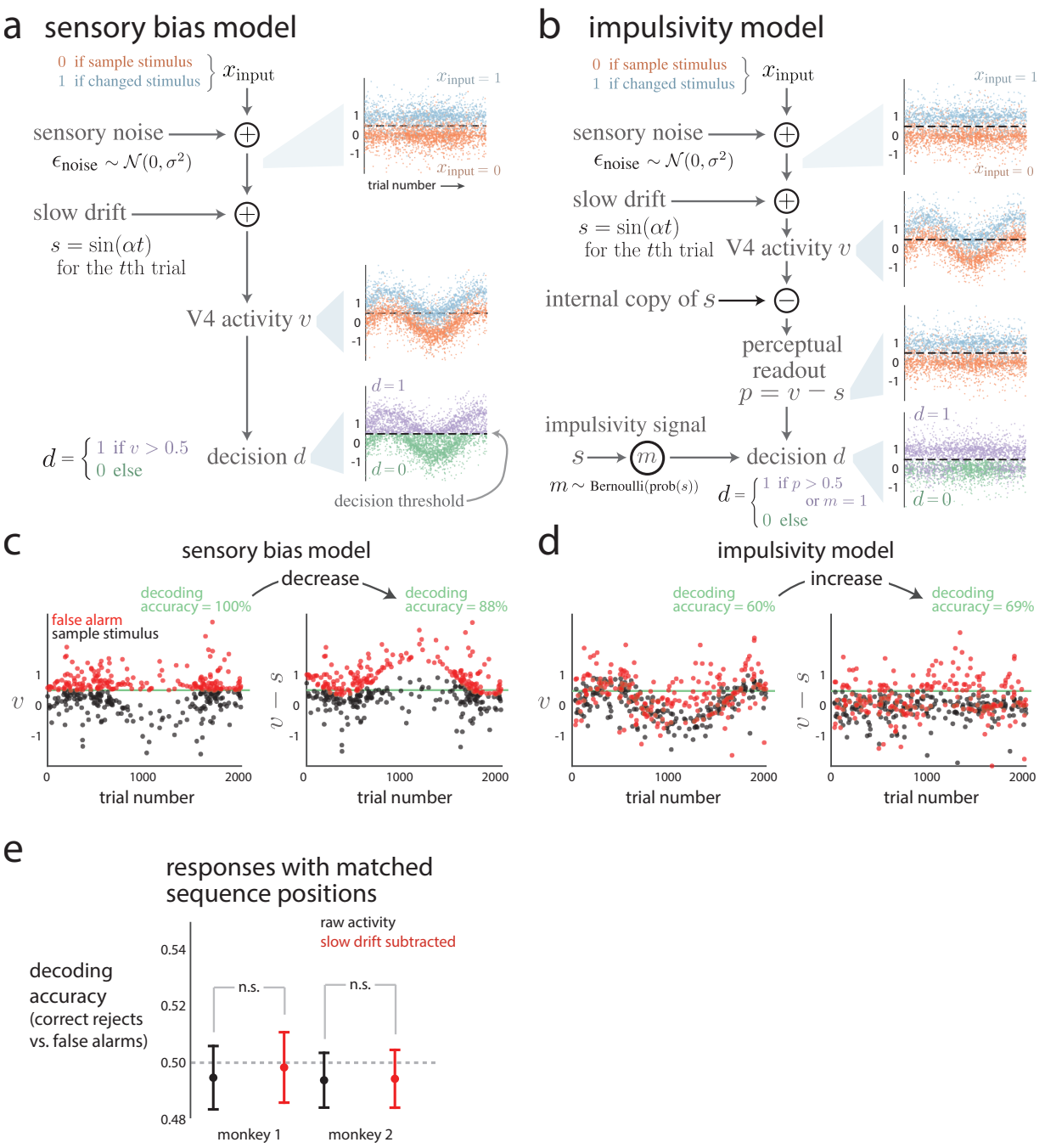
b. Correlations between magnitude of slow drift and magnitudes of behavioral variables for each monkey. All correlations are positive and close to significant (p-values are shown, permutation test).

We further analyzed correlations between behavioral variables. Behavioral variables primarily covaried along one dominant pattern, and the slow drift was most correlated with this pattern.

c. We investigated the correlations over the time within each session between the five behavioral measures for each monkey. We found a similar trend as that of the correlations between the slow drift and behavioral variables (Fig. 3**b**). Namely, hit rate, false alarm rate, and pupil diameter were mostly positively correlated; trial duration and reaction time were positively correlated; and these two groups were negatively correlated.

d. We next asked to what extent the covariance matrices in **c** could be explained by one co-fluctuation pattern among the behavioral variables. We applied PCA to the behavior variables, and found that the top principal component (PC1) explained 58%, 64% of the variance of the behavioral variables for monkeys 1, 2. This indicates that there is a dominant co-fluctuation pattern among the behavioral variables. We confirmed that the weights of this dominant pattern matched the pattern of correlations in **c** (i.e., the weights were close to $[1, 1, 1, -1, -1]$, which matched the groupings of correlations).

e. Finally, we asked to what extent did the slow drift covary with different co-fluctuation patterns of the behavioral variables. The slow drift covaried most strongly with the first principal component (PC1), significantly greater than that of shuffled data ($p < 0.002$, permutation test). This relationship was weaker for the other PCs (PC2-PC5), and in the opposite direction for the weakest PCs (PC4 and PC5, asterisk corresponds to $p < 0.05$, permutation test). Black and gray dots indicate median correlations across sessions. Error bars indicate bootstrapped 90% confidence intervals. These results indicate that a single dominant pattern explains the co-fluctuations of the behavioral variables, and that the slow drift is most correlated with this pattern.



Supplementary Figure 6: Further intuition and controls for the sensory bias model and the impulsivity model. Related to Figure 6 and STAR Methods.

We illustrate the simulation procedure of the sensory bias model, the impulsivity model, and the decoding analysis in Figure 6**e** and **j**. Simulations have the same trial structure as that of the orientation-change detection task. Each trial consists of a sequence of stimulus inputs $x_{\text{input}} \in \{0, 1\}$, where $x_{\text{input}} = 0$ indicates no stimulus change and $x_{\text{input}} = 1$ indicates a stimulus change. The trial sequence starts with $x_{\text{input}} = 0$ and has a 40% chance of changing from $x_{\text{input}} = 0$ to $x_{\text{input}} = 1$ for each consecutive input. The correct output of the model is to indicate decision $d = 0$ if $x_{\text{input}} = 0$ and $d = 1$ if $x_{\text{input}} = 1$.
(continued on next page...)

Supplementary Figure 6: (...continued from previous page)

a. Simulation procedure of the sensory bias model. Sensory noise ϵ_{noise} is added to x_{input} (top right panel depicts $x_{\text{input}} + \epsilon_{\text{noise}}$). Slow drift s is then added to form the simulated V4 activity $v = x_{\text{input}} + \epsilon_{\text{noise}} + s$ (middle right panel depicts v). Finally, the decision $d \in \{0, 1\}$ is based on v (bottom right panel depicts v but colored based on decision d). Dashed black lines indicate the decision threshold used to determine decision d. For the sensory bias model, any V4 activity v above the decision threshold of 0.5 leads to a "saccade" (i.e., $d = 1$).

b. Simulation procedure of the impulsivity model. V4 activity is simulated in the same manner as that for the sensory bias model (i.e., top right two panels are the same in **a** and **b**). An internal copy of s is removed from the readout of the V4 activity to form the perceptual readout $p = v - s$ (right panel, third from the top, depicts p). Decision d is based on the perceptual readout p passing a decision threshold (i.e., $d = 1$ if $p > 0.5$) or an impulsivity signal m (i.e., $d = 1$ if $m = 1$). It is more likely for impulsivity signal m to equal 1 if slow drift s has a higher value. An important point is that impulsivity signal m influences decision d independent of perceptual readout p, leading to some decisions $d = 1$ that are not caused by the perceptual readout p passing the decision threshold (bottom right panel, some purple dots are below the dashed line, cf. bottom right panel in **a**).

c. Illustration of the analysis procedure for decoding false alarms from the simulated V4 activity for the sensory bias model. We consider only false alarm trials, and compare simulated V4 activity v corresponding to the last stimulus input of the trial (i.e., when a false alarm occurred because $d = 1$ when $x_{\text{input}} = 0$, red dots) versus the preceding stimulus input (i.e., a correctly rejected stimulus input because $d = 0$ when $x_{\text{input}} = 0$, black dots). The red dots form a subset of the purple dots in the bottom right panel in **a**, and the black dots form a subset of the green dots. We decode V4 activity v (i.e., with the slow drift) and $v - s$ (i.e., with the slow drift subtracted) using a threshold decoder (green lines, linear SVM decoder). For the sensory bias model, subtracting the slow drift *decreases* the decoding accuracy (i.e., decoding accuracy is higher when decoding v than when decoding $v - s$). This is because the slow drift biases the sensory evidence v to be closer or further from the decision threshold (**a**, bottom right panel), and removing this bias discards information that V4 activity has about the decision.

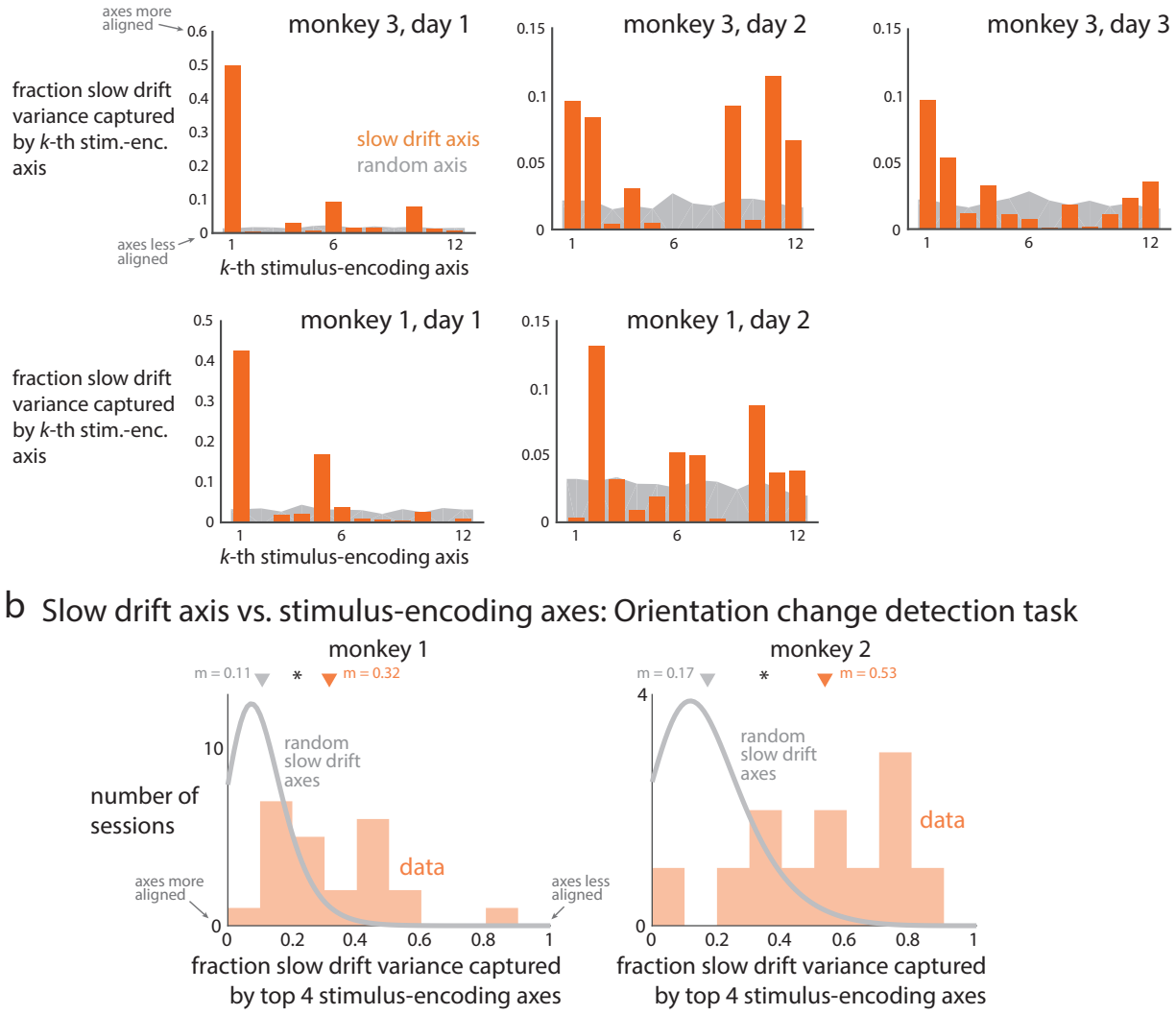
d. Same analysis procedure as in **c**, except for the impulsivity model. Subtracting the slow drift *increases* the decoding accuracy (i.e., decoding accuracy is lower when decoding v than when decoding $v - s$). This is because the slow drift is removed from perceptual readout p (**b**, right panel, third from top). Still, the slow drift influences many decisions unrelated to the perceptual readout (e.g., many false alarms occur even when p is below the decision threshold because $m = 1$). Here, the ability of v to predict the occurrence of a false alarm within a trial comes from the dependence of d on $p > 0.5$ and not from the dependence of d on $m = 1$. This is because we only consider a false alarm and its preceding correctly-rejected stimulus flash within a trial (i.e., the slow drift is held constant within a trial), and not a false alarm and a correctly-rejected stimulus flash from any trial (for which the slow drift may vary).

The decoding accuracies here are higher than those in Figure 6 because here we did not include decision output noise (see STAR Methods) in order to better illustrate the differences between the two models. Including this output noise (which accounts for the fact that we recorded only a small fraction of the neurons in the decision-making circuit, and other unobserved sources of noise are likely) yields overall decoding accuracies of the models (Fig. 6**e** and **j**) more similar to those of the real data (Fig. 6**f**).

(continued on next page...)

Supplementary Figure 6: (...continued from previous page)

e. Controlling for visual response adaptation when decoding false alarms versus correct rejects from V4 responses. It could be the case that adaptation-like effects (Kohn, 2007, *Journal of Neurophysiology* 54:100-200) led to an increase in our decoding accuracy in Figure 6**f**, rather than the false alarm itself. To control for this, for each false alarm trial, we identified a nearby matching trial (which need not be a false alarm trial) whose sequence length M was at least one flash longer than the sequence length K of the corresponding false alarm trial (i.e., $M > K$). We then decoded responses to the two stimulus flashes in positions $K - 1$ and K from the matched trial. Because these matched responses come from two correctly-rejected stimulus flashes, we expect our ability to decode these responses to be at chance, unless adaptation-like effects are at play. Indeed, when we decoded these matched responses, we found decoding accuracy close to chance (dots close to gray dashed line). In addition, we found we found no difference in decoding accuracy ($p = 0.572$, $p = 0.745$ for monkeys 1, 2, paired permutation test) when comparing decoding accuracy using the raw activity and the slow drift-subtracted activity, unlike the increase in decoding accuracy observed in responses from false alarm trials (Fig. 6**f**). Dots indicate means across sessions, and error bars indicate ± 1 s.e.m. These results indicate that the results in Figure 6**f** are not due to adaptation-like effects.



Supplementary Figure 7: The slow drift axis overlapped with the top stimulus-encoding axes for V4 responses to natural images and to sinusoidal gratings. Related to Figure 7.

One possible reason that the slow drift is removed by a downstream readout of V4 activity (Fig. 6) is that otherwise the slow drift could corrupt sensory information, impacting perception and non-impulsive decisions. To see if this were the case, we asked to what extent the slow drift axis aligns with axes that encode sensory information, where these stimulus-encoding axes were estimated based on V4 responses to natural images (Fig. 7).

a. Same convention as Figure 7**d**, shown for each of the five sessions individually. We asked whether the slow drift axis overlapped with any of the stimulus-encoding axes (i.e., at least one orange bar above gray) versus no overlap (i.e., no orange bar above gray). We computed the fraction of the slow drift variance captured by the top 12 stimulus-encoding axes (variance was summed across the top 12 axes), and found that these top 12 stimulus-encoding axes more closely overlapped with the slow drift axis than with a randomly-oriented axis for each session ($p < 0.002$, proportion of random runs with fractions above the fraction of the slow drift). The sum of these fractions (orange bars) equal the fractions reported in Fig. 7**e**. These results indicate that the slow drift tends to lie along stimulus-encoding axes, and thus the slow drift could corrupt the fidelity to which V4 activity encodes relevant image statistics relevant to downstream areas. This potential corruption of sensory encoding motivates the removal of the slow drift in order for downstream readout areas to better recover the relevant stimulus information.

(continued on next page...)

Supplementary Figure 7: (...continued from previous page)

b. In **a**, the slow drift axis overlapped with the top stimulus-encoding axes for V4 responses to natural images. Here, we perform the same analysis for the change-detection task (Fig. 1**a**), where we estimated the stimulus-encoding axes based on V4 responses to sinusoidal gratings (16 different orientations). We computed the stimulus-encoding axes by applying PCA to repeat-averaged V4 responses to 16 different orientations of sinusoidal gratings (30°, 39°, 42°, 44°, 46°, 48°, 51°, 60°, 120°, 129°, 132°, 134°, 136°, 138°, 141°, 150°), which represented rotations of the two base gratings (45° and 135°) either clockwise or counterclockwise by 4 different angles. These repeats occurred in the final flash of correct trials for which the stimulus changed inside the RFs of the V4 neurons. Because the animal could make a saccade any time after a stimulus change, we took V4 spike counts over the first 175 ms after stimulus onset, before any saccades were made. We did not consider sessions for which we recorded fewer than 16 neurons (i.e., less than the number of stimuli). We took the top 4 stimulus-encoding axes identified by PCA, which explained 85% of the repeat-averaged response variance (median across sessions and monkeys). We found that the top 4 stimulus-encoding axes captured a significantly larger amount of slow drift variance than if the slow drift lied along a random axis (left panel: monkey 1, right panel: monkey 2; asterisk indicates $p < 0.002$, permutation test).

This result, together with the results in **a** and Figure 7, suggest that the slow drift lies along stimulus-encoding axes of V4 activity, and thus may impair downstream readout if not removed.