www.sciencemag.org/content/354/6316/1140/suppl/DC1

Supplementary Material for

Selective modulation of cortical state during spatial attention

Tatiana A. Engel,* Nicholas A. Steinmetz, Marc A. Gieselmann, Alexander Thiele, Tirin Moore, Kwabena Boahen

*Corresponding author e-mail: tatiana.engel@stanford.edu

Published 2 December 2016, *Science* **354**, 1140 (2016)
DOI: 10.1126/science.aag1420

This PDF file includes:

Supplementary Materials for: Selective Modulation of Cortical State During Spatial Attention

Tatiana A. Engel*, Nicholas A. Steinmetz*, Marc A. Gieselmann, Alexander Thiele, Tirin Moore, and Kwabena Boahen

*These authors contributed equally to this work

Corresponding author e-mail: tatiana.engel@stanford.edu

Contents

1 Materials and Methods: Behavior and Electrophysiology

1.1 Subjects

Experimental procedures have been described previously (*14*). Two male monkeys (*Macaca mulatta*; $8 - 12$ kg) were used in these experiments. All experimental procedures were in accordance with NIH Guide for the Care and Use of Laboratory Animals, the Society for Neuroscience Guidelines and Policies, and Stanford University Animal Care and Use Committee. General surgical procedures have been described previously (*40*).

1.2 Behavioral task and visual stimuli

Two monkeys were trained on a cued change-detection task with a change-blindness manipulation and an antisaccade response. The monkeys performed difficult orientation change detections at peripheral locations. The discrimination was made more attentionally demanding by the simultaneous disappearance and reappearance of all peripheral stimuli (change-blindness). A central cue indicated which location would contain the change. The monkey was rewarded for reporting a successful detection with a saccade to the diametrically opposite peripheral location (antisaccade response) or for withholding a saccade if no stimulus changed.

The sequence of events for most trials was as follows. All time ranges are uniformly distributed and independently chosen unless otherwise stated. A small white dot (~ 0.15 dva

diameter) appeared on the screen, and the monkey initiated a trial by fixating it. Shortly (within 333 ms and 140 ms for monkeys G and B, respectively), four peripheral target stimuli appeared. After a brief delay of $200 - 500$ ms, the attention cue appeared: a white line < 0.5 dva in length and one pixel (< 0.1 dva) in width, originating at the fixation dot and extending in the direction of one of the four stimuli (randomly chosen on each trial with equal probability). The cue indicated with 90% or 93% validity which of the four stimuli, if any, would change on each trial. After a postcue period of $600 - 2,300$ ms with the display static as described, the four peripheral stimuli synchronously disappeared for a brief (< 270 ms) interval ("blank period") and then reappeared. Upon reappearance, one of the four stimuli changed its orientation (i.e. rotated in place) on 50% of trials. On these change trials, the monkey could earn a reward by executing a saccadic eye movement within 800 ms to the stimulus opposite the changed stimulus. On the other 50% of trials, all four stimuli appeared at identical orientations to those they had before disappearing. In this case, the monkey was rewarded for maintaining fixation on the central dot for 800 ms. The trial was terminated without reward if, at any time prior to the stimulus reappearance, the eye position left a small square box (~ 1.5 dva width) around the fixation dot.

The target stimuli were static Gabor patches: oriented black and white gratings in a circular Gaussian aperture. For monkey G, the gratings were square wave; for monkey B, they were sine wave modulated. Both types elicited robust responses from the neurons in this study. In both monkeys, the gratings were at maximal contrast for the monitor. The dimensions of the gratings varied somewhat from session to session but were typically ~ 4 dva in diameter and approximately one cycle/dva in spatial frequency. The location of the gratings also varied depending on the receptive field locations of the neurons being recorded, but the centers were always between 5 and 8 dva eccentricity. All four gratings had equal eccentricity and were spaced evenly at 90° intervals around a circle. The screen background was dark gray for monkey G and middle gray for monkey B, but for neither monkey were the mean luminances of the gratings matched to the background color.

The orientation of the grating took one of 8 possible values, evenly spaced from $0°$ to $180°$. The grating orientations were chosen independently for each of the four stimuli and for each trial. The difficulty of the task was varied by the amount of grating rotation: $45°$, $67°$, or $90°$. Trials with these different rotation magnitudes were interleaved randomly. The rotation was clockwise or counterclockwise with equal probability, independently chosen for each trial.

Visual stimuli were displayed on a monitor at 120 Hz and $1,680 \times 1,050$ resolution (17.8 pixels/degree). The monitor was positioned 28.5 cm from the monkeys' eyes. Presentation of stimuli was controlled by Cortex software (http://dally.nimh.nih.gov). Eye position was monitored in some sessions for monkey G with a scleral search coil. In the remaining sessions for monkey G as well as all sessions for monkey B, eye position was monitored with an EyeLink 1000 video eye-tracking system. Microsaccades were defined as eye movements that exceeded 0.1 degrees amplitude and had maximum velocity > 10 degrees/second for $\geqslant 10$ ms (*41*).

1.3 Neural Recordings

1.3.1 Linear array recordings

Recordings were made with 16-channel U-Probes. These electrodes are cylindrical in shape (constant diameter of 185 μm) and have a row of 16 circular platinum/iridium electrical contacts (15 μm diameter) with 150 μm center-to-center spacing. The total length of the array is 2.25 mm. Data were amplified and recorded using the Omniplex system. Wide-band data, filtered only in hardware at 0.5 Hz high pass and 8 kHz low pass, were recorded to disk at 40 kHz. Spikes were detected from this signal and sorted as described in detail previously (*14*).

Recordings were targeted with MRI (*42*) in order to be as perpendicular to V4 cortical layers as possible so as to maximize the overlap of RFs of recorded neurons (fig. S1). In this way, the single grating stimulus presented in the lower-left visual field during the task could be positioned to drive all recorded V4 neurons. MRI images were collected with a 3T scanner and were T1-weighted. The position of the recording chamber was visualized by filling it with copper sulfate, a contrast agent (*42*). Area V4 was identified as the prelunate gyrus. We visually identified both the surface of V4 as well as the border of V4 grey matter with the white matter below, and computed vectors normal to both surfaces. These approach vectors were followed by the electrodes using a custom tiltable microdrive.

1.3.2 Detection of multi- and single-unit activity spikes

Multi- and single-unit activity (MUA and SUA) detection procedures have been described in detail previously (*14*). Briefly, the raw 40 kHz-sampled voltage trace was match-filtered (i.e., convolved with a spike-shaped filter (*43*)) and peaks were detected. Then, a moving-threshold algorithm was employed with a very high threshold to detect the largest peaks and to exclude any small spikes that occurred within the exclusion window around these largest peaks. Next, progressively lower thresholds were used until approximately 100 Hz of spiking activity was detected on each channel. Out of 736 total recorded channels in the 46 sessions, we were able to isolate 285 single neurons. Multi-unit activity exhibited response dynamics very similar to isolated single neurons: e.g., in On/Off firing rate modulation (fig. S4C,F), and auto- and cross-correlations (fig. S2).

1.3.3 Receptive field measurement

RFs were measured by recording spiking responses to briefly flashed stimuli. The stimuli were 3×3 dva squares on an evenly spaced 6×6 grid, each flashed for 200 ms with a random length $200 - 300$ ms blank screen interval between flashes. The grid coordinates were $-15, -12, -9, -6, -3$ and 0 dva relative to the fixation point in both horizontal and vertical dimensions. The grid covered the lower left visual field with a square 15 dva on each side. Spikes were counted for each stimulus presentation in the window 0 to 200 ms relative to stimulus onset, and were averaged across all presentations of each stimulus. To compute receptive field

contours and centers, the minimum response was subtracted from the map, then the map was upsampled by cubic interpolation from a grid with 3 dva spacing to one with 0.1 dva spacing. The RF center was the location of the peak of the upsampled map, and the visualization contour was the contour at 75% peak height.

2 Materials and Methods: Data Analysis and Modeling

Our dataset consisted of a total of 46 recording sessions (25 in monkey G and 21 in monkey B) during the attention task, and 5 recording sessions in monkey G during the fixation task. All analyses were performed using a custom code written in Matlab. Code is available upon request via email.

2.1 Auto- and cross-correlation analysis

In order to measure the extent of the On-Off transitions across the whole dataset, we computed auto- and cross-correlations of spiking activity for all units and unit pairs, respectively. Correlations were computed using a jitter correction method, which corrects for slow temporal correlations and for stimulus-locked correlations (*44, 45*). The jitter-corrected correlations are computed by subtracting the expected value of correlations produced from a resampled version of original spike trains, with spike-times randomly shuffled (jittered) within a specified temporal window (the jitter window). The correction term is the average over all possible resamples of the original spike trains (the expected value), and is subtracted from the raw correlation.

Empirical correlations are computed using small discrete time bins Δt, and spikes are jittered within larger jitter bins T. The jittering technique preserves two marginals: the total spike count in each time bin Δt summed across all trials (the peristimulus time histogram, PSTH), and the spike count within each jitter bin T on each trial (the instantaneous firing rate computed in bins T). For each spike on each trial, a new spike is chosen randomly from the set of all spikes within the same jitter bin on all of the trials. Jitter correction removes correlations on timescales greater than the jitter window T. Because it preserves the PSTH shape, jitter correction also removes any correlations due to stimulus-locked firing rate modulation.

To assess the statistical significance of empirical correlations in each time bin Δt, the measured correlation value needs to be compared to the distribution of correlations produced by all possible resamples. For sufficiently large numbers of spikes and experimental trials, the resampled distribution of correlations in each time bin Δt is well approximated by a Gaussian. Therefore, we derived analytical formulas for the mean and variance of the resampled distribution, and transformed jitter-corrected correlations into standard z-scores, normalizing them by the standard deviation of the resampled distribution. Accordingly, in each time bin the z-scored correlation represents the deviation of the correlation from the mean of the resampled distribution, measured in units of its standard deviation.

We computed auto- and cross-correlations for stimulus-driven activity during the sustained

response to the grating stimuli (400 to 800 ms window after the attention cue onset) and for spontaneous activity during the initial fixation period of the attention task (-333 to 30 ms window in monkey G, and -140 to 30 ms window in monkey B relative to the visual stimulus onset; these two monkeys were required to fixate for different durations prior to the stimulus onset). Activity was analyzed on successfully completed trials on which no other behavioral event happened during the specified time window. Units were only included if their average firing rate during the specified time window was greater than 5 Hz. Correlations were computed in $\Delta t = 1$ ms bins, using jitter windows of $T = 400$ ms and $T = 170$ ms for stimulus-driven and spontaneous activity, respectively. Only cross-correlations on channels at least one channel apart (300 μm) were included and cross-correlations on adjacent channels were not included in the analysis, to eliminate the possibility that spikes from the same neuron could be registered on both channels. Population averages were expressed as standard z-scores by normalizing the sum of auto- and cross-correlations by the square root of the number of units and unit-pairs, respectively.

We also computed auto-correlations for a separate dataset from ref. (*46*), where activity of 212 single neurons was recorded in area V4 with conventional sharp electrodes in two monkeys under similar behavioral conditions (fig. S2C). All details of behavioral task and recording techniques were published previously (*46*). In brief, two monkeys performed a visually guided, delayed saccade task which was initiated by fixation of the central spot. Immediately following fixation, an oriented bar stimulus appeared in the RF of the neuron under study and remained there until the end of the trial. The monkeys maintained fixation for a variable delay ($0.5 - 1$ s) and, when the fixation spot was extinguished, performed a saccade to one of 3 possible locations. Spiking activity was analyzed within the -400 to 0 ms window relative to the disappearance of the fixation spot. Neurons were only included in the analysis if their average firing rate during this window was greater than 5 Hz ($n = 165$). Auto-correlations were computed in $\Delta t = 1$ ms bins, using a jitter window of $T = 400$ ms.

2.2 Hidden Markov Model analysis

We hypothesized that firing rates in our recordings switched randomly between the On and Off phases. However, these On-Off firing rate dynamics are not observed directly, but only through spikes emitted stochastically with the corresponding instantaneous firing rates. This random point process adds another layer of stochasticity: the underlying firing rate must be inferred from the measured spike counts. The spike-count fluctuates from time-bin to time-bin even when the underlying rate is constant. We employed the Hidden Markov Model (HMM) to segment spiking data into the On and Off episodes in a statistically principled way.

The HMM (*47, 48*) was fitted to the stimulus-driven MUA on 16 channels in the attention task within the window, which started 400 ms after the attention cue onset and ended at the end of the post-attention-cue period (i.e. start of the blank period, see Fig. 1D). The duration of this time-window ranged from 200 to $1,900$ ms across trials. These data were analyzed separately for 32 conditions: 4 behavioral conditions—covert attention, overt attention and 2

control conditions—times 8 stimulus orientations. During each recording session, monkeys performed 1500 trials on average, which amounts to 46 trials per condition on average. This number was sufficient for a good estimate of the model parameters, which is evident from tight confidence intervals on the model parameters obtained by bootstrapping (see Fig. 2B,E and supplementary methods 2.2.1). The summary statistics for each unit or recording were obtained by averaging across these conditions. We also fitted the HMM to the spontaneous activity during the initial fixation period of the attention task within the time window starting immediately following fixation (333 ms and 140 ms prior to the stimulus onset in monkeys G and B, respectively) and ending 30 ms after the stimulus onset; and during the fixation task within the 0 to $3,000$ ms time window following fixation (fig. S4).

2.2.1 Fitting the Hidden Markov model

The HMM assumes that MUA on 16 simultaneously recorded channels is influenced by a common binary-valued latent variable s, switching between the Off ($s = 0$) and On ($s = 1$) phases. Dynamics of the latent variable are modeled as a first-order Markov process, meaning that at each time step t, the probabilities of transitioning between phases depend only on the value of s at time t: $P(s_{t+1}|s_t, \ldots, s_0) = P(s_{t+1}|s_t)$. Dynamics of the latent variable evolve in discrete time steps. Accordingly, MUA was converted into spike-counts using 10 ms bins, such that n_t^j is the spike-count recorded on channel j in a 10 ms bin following time t.

The latent variable s cannot be observed directly, but only through spikes. The HMM assumes that at each phase—On and Off—spikes on each channel are generated according to Poisson point-processes, whose mean firing rates depend on the current phase as defined by the emission matrix $\mathbf{\Lambda}$, which specifies λ_j^s, the expected spike-count of channel j in phase s within a time bin, i.e. the rate of a Poisson process. Particularly, the conditional probability of observing n spikes on channel j while in a phase s is given by

$$P(n|s) = \frac{(\lambda_j^s)^n}{n!} e^{-\lambda_j^s}. \tag{1}$$

Thus, the parameters of the HMM are the transition matrix $\mathbf{P}$ ($P_{11} = p_{\text{off}} = P(s_{t+1} = 0|s_t = 0)$, $P_{22} = p_{\text{on}} = P(s_{t+1} = 1|s_t = 1)$, $P_{12} = 1 - p_{\text{off}}$, $P_{21} = 1 - p_{\text{on}}$); the emission matrix $\mathbf{\Lambda}$, and the vector of probabilities $\boldsymbol{\pi}^{\mathbf{0}}$ governing the initial value of the latent variable s_0 ($\pi_i^0 \equiv P(s_0 = i)$).

The HMM was fitted with the Expectation Maximization (EM) algorithm (*47, 48*). To determine the model's convergence, at each EM iteration we monitored relative changes (|new—original|/|original|) in the model's log-likelihood and in the transition and emission matrices. The optimization was terminated when the relative changes dropped below specified criteria: 10^{-5} for the log-likelihood, and 10^{-3} for the transition and emission matrices.

To avoid local optima, the EM-optimization was repeated for ten random parameter initializations, and the parameter set yielding the maximal likelihood was chosen. The initial parameter values for vector $\boldsymbol{\pi}^{\mathbf{0}}$ and for each row of the transition matrix $\mathbf{P}$ were sampled from a Dirichlet distribution, and the elements of the emission matrix were initialized from a uniform distribution on $[0, 2\langle n_t^j \rangle]$, where $\langle n_t^j \rangle$ is the mean spike-count of channel j.

The fitted HMM is not identifiable: swapping the labels $s = 0$ and $s = 1$ would result in exactly the same model. Therefore, to consistently label the inferred latent states as On and Off phases, we computed the average firing rate across all channels for the two latent phases, and labeled the phase with higher average firing rate as the On-phase ($s = 1$), and the other phase as the Off-phase ($s = 0$).

To obtain confidence bounds for the estimated HMM parameters in each condition, ten bootstrap samples were generated by sampling trials randomly with replacement from the set of all trials available for that condition. For each bootstrap sample, the HMM was refitted, and the confidence bounds were estimated as 5% and 95% percentile of the bootstrapped parameter values.

Once the parameters of the HMM were estimated, the most likely latent sequence of On and Off episodes was decoded on a trial-by-trial basis using the Viterbi algorithm. To verify the validity of Markovian assumption, we tested whether the sequence of decoded On and Off episodes exhibited history effects. For each condition, we computed Pearson correlation coefficients between the durations of consecutive Off and On episodes, and of consecutive On and Off episodes, and then averaged these correlations across conditions for each recording session. The average correlations were not significantly different from zero: mean Off-On correlation $r = 0.01$, $p = 0.74$; mean On-Off correlation $r = -0.02$, $p = 0.15$ (Wilcoxon signed rank test). Thus the Markovian assumption was confirmed in the data.

To fit the HMM, we used the bin size $\Delta t = 10$ ms. The bin size can neither be too large nor too small for important reasons. If the bin size is too large (larger than a typical transition time), the spike count in a single bin would be the average over several On and Off episodes, resulting in a poor fit. On the other hand, if the bin size is too small (e.g., 1 ms), then the On or Off phase become harder to distinguish because the difference between the mean spike count in the On and Off phases decreases linearly with the bin size while the standard deviation decreases as the square root of the bin size (assuming Poisson statistics). We verified that 10 ms bin size provides a reliable fit and also a good temporal resolution. To this end, we repeated the fitting procedure multiple times with random initial conditions for the EM-algorithm, and observed that estimated transition times were unchanged between different repetitions of the fit.

2.2.2 Cross-validation procedure and variance explained

The quality of the HMM fit was assessed in a two-fold cross-validation procedure (fig. S4). To this end, the HMM parameters were estimated on a random half of trials (training trials). These estimated parameters were then used to decode the most likely sequence of the On and Off episodes on the remaining half of the trials not used for fitting the model (testing trials) and then to compute the cross-validation score. The procedure was repeated with training and testing trials swapped. The scores were averaged over the two cross-validation folds.

For MUA on each channel j, the fraction of variance explained by the HMM (R^2) was

computed as

$$R^2 = 1 - \frac{\text{Var}_{\text{res}}[n_j]}{\text{Var}_{\text{tot}}[n_j]} = 1 - \frac{\sum_{t=1}^{N}(n_t^j - \lambda_j^{s_t})^2}{\sum_{t=1}^{N}(n_t^j - \langle n_t^j \rangle)^2}. \quad (2)$$

Here $\text{Var}_{\text{tot}}[n_j] = \frac{1}{N}\sum_{t=1}^{N}(n_t^j - \langle n_t^j \rangle)^2$ is the total variance of the spike-count data, $\text{Var}_{\text{res}}[n_j] = \frac{1}{N}\sum_{t=1}^{N}(n_t^j - \lambda_j^{s_t})^2$ is the residual variance not accounted for by the HMM, and N is the total number of time bins.

The cross-validation procedure for SUA was slightly modified, because the HMM was fitted to the MUA only. Accordingly, SU firing rates in the On and Off phases were not the model parameters and had to be estimated separately. Specifically, the most likely latent sequence of On and Off episodes was decoded from the MUA on training trials using the HMM parameters estimated on training trials. Then the SU firing rates during the On and Off phases were estimated from training trials as the mean firing rates during these decoded On and Off episodes, respectively. Finally, the estimated On and Off firing rates of SUA were used in equation 2 in place of λ_j^s to compute R^2 on test trials. In this analysis, we only included SUs, whose trial-average firing rate during the analyzed time window was greater than 1 Hz.

We also computed R^2 across timescales (fig. S4B,E). To this end, spike-counts of MUA and SUA were evaluated for 10 different bin-sizes (integration times) ranging from 50 to 500 ms in 50 ms steps. R^2 was computed following the same procedures, but the predicted spike-count in equation 2 was the average of the firing rates predicted by the HMM within each integration-time bin.

2.2.3 The maximal explainable variance

In order to interpret results of the cross-validation analysis, we derived an analytical expression for the maximal explainable variance R^2_{max} given that the HMM assumptions hold true. This calculation assumes that spikes are generated by a Poisson process with the mean firing rate alternating between two levels corresponding to the On and Off phases, and that the latent sequence of the On and Off episodes and the firing rates in each phase are known precisely (no decoding error).

Intuitively, all variance due to firing rate fluctuations is eliminated in this idealized scenario by the precise knowledge of the mean firing rate in each time bin. Therefore, the residual variance of spike-count $\text{Var}_{\text{res}}[n]$ is just the variance of the Poisson point-process, which is equal to its mean $E[n]$. Thus equation 2 transforms into

$$R^2_{\text{max}} = 1 - \frac{E[n]}{\text{Var}_{\text{tot}}[n]} = 1 - \frac{1}{\text{FF}}, \quad (3)$$

where the Fano factor (FF) is a standard measure of the firing-rate variability, defined as the variance over the mean of the spike-count. More formally, if λ^{on} and λ^{off} are the mean firing rates during the On and Off phases, N is the number of time steps (i.e. 10 ms bins) within the

integration time, and $\bar{n}$ is the number of time steps spent in the On phase on average, then the residual variance can be calculated as

$$\mathrm{Var}_{\mathrm{res}}[n] = \bar{n}\langle(n-\lambda^{\mathrm{on}})^2\rangle_{\mathrm{on}} + (N-\bar{n})\langle(n-\lambda^{\mathrm{off}})^2\rangle_{\mathrm{off}} = \bar{n}\lambda^{\mathrm{on}} + (N-\bar{n})\lambda^{\mathrm{off}} = E[n], \quad (4)$$

where we again used the fact that the variance of a Poisson process is equal to its mean. Equation 3 follows immediately.

In fig. S4A,D, the maximal explainable variance curve simply depicts equation 3. In fig. S4B,E, in order to compute the average maximal explainable variance curves, we first computed R^2_{max} for each SUA and MUA using their FF evaluated using the specified integration-time, and then averaged across the population.

2.2.4 Firing rate modulation by the On-Off transitions

We tested whether firing rates of SUA and MUA were significantly different between the On and Off phases. First, the most likely sequence of On and Off phases was decoded using the HMM fitted on all trials. Then, for each decoded On and Off episode, the firing rate was calculated as the spike-count during that episode divided by its duration. The resulting distributions of firing rates across all episodes were compared with Wilcoxon sign-rank test for each behavioral condition. If significance rate across 32 behavioral conditions was higher than chance at 0.05 significance level (i.e. if firing rates were significantly different in 2 or more conditions), then the unit firing rate was declared to be significantly modulated by On-Off transitions.

2.2.5 Modulation of On-Off dynamics by attention

SUA and MUA were segmented into the On and Off episodes using the HMM fitted on all trials. For each unit, the average firing rates during the On and Off episodes were computed, as well as the average durations of On and Off episodes. For each behavioral condition, the results were averaged across 8 stimulus orientations, and for the control condition they were also averaged across two directions of the attention cue.

2.2.6 Relationship between the On-Off dynamics and behavioral performance

To determine how the On-Off dynamics in V4 population activity are related to animals' behavioral performance on a trial-by-trial basis, we fitted the HMM to the 16-channel MUA during the blank period of attention task (lasting from the initial-stimulus offset until 30 ms after the test-stimulus onset, see Fig. 1D) using the same procedures as described in supplementary methods section 2.2.1. For each recording, the HMM was fitted separately for each of 8 stimulus orientations and 4 attention conditions, including all (correct and error) trials. All trials were then divided into two groups—On-phase and Off-phase trials—according to the decoded population state at the last time bin (i.e. 30 ms after the changed stimuli onset). The probability to detect an orientation change was computed as a proportion of change trials with the correct antisaccade response, separately for On-phase and Off-phase trials and separately for covert attention

and control conditions. We then computed the difference in the detection probability between the On-phase and Off-phase trials and tested whether this difference was significantly different from zero across recording sessions (Wilcoxon signed rank test).

To investigate how the On-Off dynamics are related to behavioral performance across time, we computed the detection probability separately for trials that were in the On phase and in the Off phase at each time bin for the whole duration of the blank period in the covert attention condition. To rule out any possible effects of stimulus-offset transients on our results, only trials with the blank-period duration of at least 190 ms were included in this analysis. To test whether the detection probability was significantly different between the On-phase and Off-phase trials across all times, we employed a non-parametric permutation approach that controls for multiple comparisons across all time bins (*49*). First, we calculated t-values for the difference in detection probability between the On-phase and Off-phase trials across recordings, separately for all time bins (non-randomized t-values). Next, the following permutation procedure was repeated 10,000 times: for each recording, the On-phase and Off-phase trials were exchanged (i.e. relabeled) with probability 0.5 or left unchanged otherwise. Subsequently, paired t-tests were performed across recordings between relabeled data for all time bins. The maximal and the minimal t-values across all time bins were saved, resulting in 10,000 maximal and 10,000 minimal t-values, and the 97.5 and 2.5 percentiles were determined for these two empirical distributions, respectively. For each time bin, the non-randomized t-value was considered significant if it was larger than 97.5 percentile of maximal t-values or less than 2.5 percentile of minimal t-values. This procedure corresponds to a two-sided test with a global false positive rate of 0.05 and corrects for the multiple comparisons across the whole time interval.

2.3 Analysis of local field potentials

The local field potential (LFP) was defined as the continuous voltage signal filtered in hardware with a high-pass filter at 0.5 Hz, then filtered in software with a 4-pole Bessel low-pass filter at 200 Hz and second-order notch filter at 60 Hz, and finally downsampled at 1 kHz. Ten out of 31 recording sessions with On-Off transitions had to be excluded from the LFP analyses due to the presence of low frequency artifacts, likely resulting from amplifier saturation. Current source density (CSD) was calculated using the discrete double spatial derivative (*50*) with "Vaknin" electrodes added to yield the same number of CSD channels as LFP channels (*51*).

LFP power during the On and Off phases was computed for On and Off episodes of at least 250 ms duration using a multitaper method with three tapers and a spectral bandwidth of 5 Hz (*52*). Average LFP and CSD aligned to the On-to-Off and Off-to-On transitions included individual traces only for as long as the On and Off episodes on either side of the transition persisted, such that transition-aligned averages included fewer episodes for greater times relative to the transition time. Latency of transitions in firing-rates and CSD was determined as the first time when the average traces aligned to On-to-Off and Off-to-On transitions switched (i.e. the sign of their difference changed). The latency of transitions in firing-rates was averaged across all channels within each recording. Because the sign of CSD signal changes across cortical

depth (sources/sinks), only one CSD channel with a large amplitude was selected per recording for latency calculation.

3 Supplementary Information

3.1 Signature of On-Off transitions in auto- and cross-correlations

In order to measure the extent to which the On-Off transitions permeated cortical dynamics, we computed auto- and cross-correlations of spiking activity across all datasets (supplementary methods 2.1). Auto- and cross-correlations were computed using a jitter-correction method (*44,45*) that eliminates stimulus-locked correlations and slow firing-rate correlations on timescales longer than a specified time window (400 ms in figs. S2A,B,E and S7B and 170 ms in fig. S2C,D; see supplementary methods 2.1). At each time lag, the correlation value was expressed as a z-score relative to the distribution of correlations expected by chance.

The On-Off transitions observed during the attention task manifested in a triphasic shape in the spike auto- and cross-correlations of MUA (fig. S2A). A typical auto-correlation of multi-unit activity (MUA) showed a highly significant central peak flanked by two significant negative lobes, indicating that spikes on individual channels arrived in temporal clusters separated by intervals of low spiking activity. A typical cross-correlation of MUA showed a similar triphasic shape, indicating that periods of excessive spiking and quiescence were nearly synchronous even on distant channels. Auto- and cross-correlations of isolated single-unit activity (SUA) exhibited the same temporal clustering (fig. S2B, lower z-scores are at least in part due to lower firing rates of SUs), suggesting that SU firing was also modulated by On-Off transitions. Similar to the stimulus-driven activity, the On-Off transitions in the spontaneous SUA and MUA during the fixation period also exhibited a triphasic shape in the auto- and cross-correlations (fig. S2C,D). Thus triphasic shape of spike auto- and cross-correlations suggests that On-Off transitions were prevalent across the whole dataset.

3.2 Relationship between On-Off transitions and local field potentials

3.2.1 Timing relationship between firing-rate transitions and LFP

We investigated the relationship between the On-Off firing-rate transitions and local field potentials (LFPs). To this end, we computed the average LFPs aligned to the times of On-to-Off and Off-to-On transitions (fig. S3A,C). The transition-aligned LFPs showed large deflections around the times of transitions. To confirm that these LFP deflections reflected local activity, we also computed the current source density (CSD). The transition-aligned CSD exhibited a characteristic pattern of sources and sinks alternating through the cortical depth, which switched polarity around the times of On-Off transitions (fig. S3B), thus confirming that the LFP deflections indeed reflected local activity. In summary, the low frequency LFP/CSD fluctuations

were strongly phase-locked to the On-Off transitions, consistent with the previous reports that low-frequency LFP is coupled to spiking activity, particularly in the synchronized state (*11*).

This phase-locking relationship between the On-Off transitions and LFP raises the question of whether changes in the On-Off spiking dynamics observed during attention could be a consequence of the changes in the low-frequency LFP power during attention (*14, 21*). The conventional interpretation—that the LFP represents primarily synaptic currents—leaves open the question of whether these synaptic currents arise from distant sources, and therefore constitute inputs that may drive local spiking activity, or whether they arise from local sources, and thus are a consequence of the local spiking that we measured. In addition, the conventional interpretation remains in question (e.g., the recent suggestion that LFP is driven in majority by active dendritic currents (*53*)), and the exact origin of the LFP signal remains an area of active research.

Nonetheless, we examined whether transitions in spiking activity lead or lag the phase-locked LFP fluctuations. To quantify the timing relationship between the On-Off firing-rate transitions and CSD, we computed the latency to transition for both signals. The latency was defined as the time from the moment of transition (as determined by the HMM) until the time when the average activity aligned to the On-to-Off transitions crossed the average activity aligned to the Off-to-On transitions. The CSD signal did not lead the firing-rate signal for any recording, and on average lagged the firing-rate signal by 10.8 ms ($p < 10^{-4}$, Wilcoxon signed rank test) across recordings. This result supports the conclusion that—assuming the conventional interpretation of the LFP is correct—the spikes we measured were the dominant source of synaptic input reflected in LFPs. If this is indeed the case, the changes in duration of On and Off episodes we observed explain the changes in LFP power with attention, rather than vice versa.

3.2.2 LFP power shifts associated with On-Off transitions

We examined whether LFP power-spectra were different between the On and Off phases. We computed LFP power-spectra separately during the On and Off episodes, and found that gamma power ($40 - 80$ Hz) was enhanced and low frequency power ($0 - 10$ Hz) was suppressed in the On phase relative to the Off phase (fig. S3D,E). This observation links our finding that On-episode durations increase during attention with the earlier findings that the low-frequency LFP power is suppressed and gamma-range LFP power is enhanced during attention, which has been replicated in our dataset (*14*) and those of others (*21*).

3.3 Cross-validation and variance explained by the HMM

3.3.1 Variance explained by the HMM

We quantified the amount of spiking variability captured by the HMM in a cross-validation procedure by computing the fraction of variance explained (R^2) on a separate subset of data not used for fitting the model (fig. S4, supplementary methods 2.2.2). To interpret the results of the cross-validation analysis, we also computed the maximal explainable variance R^2_{max}, given that

the HMM assumptions hold true (i.e. spikes are produced by switching between two Poisson processes with different mean rates). Since Poisson point-processes represent an irreducible source of variability, the HMM accounts only for variability arising from the difference in their mean rates. Accordingly, we took advantage of a simple relationship between R^2_{max} and the *Fano factor* (FF), a measure of firing-rate variability (i.e. the variance divided by the mean spike-count in a specified time interval). Namely, $R^2_{\text{max}} = 1 - 1/\text{FF}$ (see supplementary methods 2.2.3 for derivation). Note that if spikes on a channel are from a single Poisson process (i.e. $\text{FF} = 1$), then R^2_{max} equals zero for that channel even if its mean firing rate is estimated very accurately by the HMM. The actual variance explained can deviate from R^2_{max} if the model assumptions are not perfectly met in the data as well as due to inaccuracies in estimating model parameters and in decoding the On and Off episodes from finite data.

The HMM performed strikingly well: the variance explained approached R^2_{max} for many multi-units (fig. S4A). We computed the variance explained for spike-counts in time-windows of different lengths (fig. S4B). On short timescales, the variance explained approaches zero, because the number of spikes falling in a short window becomes small (0 or 1 in the limit), hence, any spiking process becomes indistinguishable from a Poisson process. On the other hand, the variance explained also approaches zero on timescales exceeding the characteristic timescale of On-Off transitions, because, over a very long time-window, the On-Off rate fluctuations average out, hence, the expected number of spikes is determined just by the overall mean firing rate. Accordingly, the variance explained exhibits a maximum at the time corresponding to the characteristic timescale of On-Off transitions, which was approximately 200 ms in our data (fig. S4B). On average, the HMM captured about half of the maximal explainable variance.

We also computed the variance explained for isolated single-units. Similar to the result for MUA, the variance explained approached R^2_{max} for many SUs (fig. S4A). On average, the variance explained was lower for SUA than for MUA. This trend was partially due to lower FF values of SUA (fig. S4A) and partially due to SUA variability within each phase being different from Poisson (i.e. SUA during On and Off phases deviated more strongly from a Poisson process). For a few single units that spiked more regularly than a Poisson process in the On and Off phases, the variance explained exceeded R^2_{max}. For a small number of single units the variance explained was negative, indicating that these units did not follow the population On-Off dynamics. This finding is consistent with the observation that cortical neurons differ in their coupling to the overall population firing, ranging from strongly coupled "choristers" to weakly coupled "soloists" (*54*). Overall, the HMM with a binary-valued latent variable provided an excellent description of the population spiking activity, accounting on average for nearly half of the maximal explainable variance.

To quantify the extent to which spiking of individual units was modulated by On-Off transitions, we computed the On-Off firing rate modulation index $\text{MI}_{\text{on-off}} = [r_{\text{on}} - r_{\text{off}}]/[r_{\text{on}} + r_{\text{off}}]$, where r_{on} and r_{off} are the average firing rates during the On and Off phases, respectively. This modulation index varies from -1 to 1, where positive values indicate higher firing rate during the On phase, negative values indicate higher firing rate during the Off phase, and $\text{MI}_{\text{on-off}} = 0$ indicates that firing rate was not modulated by the On-Off transitions. Firing rates during the

On and Off phases were significantly different (Wilcoxon signed rank test, $p < 0.05$) in all MUs and in the overwhelming majority of SUs (91.1%), and $MI_{on\text{-}off}$ was positive in nearly all MUs (99.6%) and in the majority of SUs (90.9%, fig. S4C). For several SUs, the $MI_{on\text{-}off}$ value approached 1, indicating that these units spiked exclusively during the On episodes and almost never spiked during the Off episodes. Very similar results were obtained for spontaneous activity during the fixation period of the attention task (fig. S4D-F). In summary, our analyses demonstrate that spiking of individual neurons throughout the cortical depth is strongly modulated by the On-Off transitions during spontaneous and stimulus-driven activity.

3.3.2 Selecting the number of latent phases in the HMM

To verify whether HMM with two latent phases was a sufficiently good description of our data, we performed cross-validation analyses fitting HMMs with number of phases n ranging from 1 (single Poisson process for each channel) to 8. We implemented a 4-fold cross-validation procedure, where the model was fitted to a randomly selected subset of 3/4 trails, and then the cross-validation error Var_{res} was computed for each channel on the remaining 1/4 trials using 200 ms window (as described in supplementary methods 2.2.2). The procedure was repeated 4 times (folds) using each of four 1/4 trials as a test-set for computing the cross-validation error. The cross-validation error for HMMs with $n = 1, \dots 8$ phases was normalized by the cross-validation error of the 1-phase HMM and then averaged across all channels, conditions (4 attention conditions $\times$ 8 stimulus orientations) and cross-validation folds.

Plotting this normalized cross-validation error against the number n of latent HMM phases shows how model fit improves with every additional phase for each recording session (fig. S5A), which allows for selection of the most parsimonious model based on the elbow (kink) in this error plot (so called elbow method). For most recording sessions, addition of the second HMM-phase greatly reduced the cross-validation error compared to the 1-phase HMM, whereas adding more phases resulted in only marginal improvements. The error curves for these recordings display an elbow at $n = 2$. For some recording sessions, HMMs with $n > 1$ phases did not perform better than the 1-phase HMM and the error curves did not exhibit a kink for these recordings.

We selected the most parsimonious model for each recording session, by considering how adding each additional phase to the HMM reduced the cross-validation error (i.e. improved the model fit). We computed the relative change in cross-validation error as the difference in normalized cross-validation error for HMMs with $n + 1$ and n phases (fig. S5B) and defined a criterion for determining the number of phases in the most parsimonious HMM: the number of phases was increased only if an additional phase reduced the cross-validation error by more than 10%. For all recordings, HMMs with more than 2 phases were never selected. For 15 (33%) recordings, a one-phase HMM was the most parsimonious model. For most recordings (31 total, 67%), the two-phase HMM was the most parsimonious model. Across all recordings, the variance explained by the 2-phase HMM (R^2) was tightly correlated with the average spike-count correlation (i.e. noise correlation, a pairwise measure that quantifies the level of

coordinated activity across ensemble), which indicates that On-Off dynamics largely contribute to measured spike-count correlations.

3.4 Replication of On-Off dynamics across different datasets

We ascertained that our observation of On-Off transitions in the visual cortex of behaving monkeys was not contingent on specific experimental procedures employed in our laboratory nor on details of our behavioral tasks. To this end, we analyzed two additional datasets recorded from area V4 of four different behaving monkeys, in two different laboratories, and with two different types of electrodes. The first additional dataset consisted of 20 recording sessions from area V4 of two fixating monkeys obtained with a different type of 16-channel linear array electrodes. All experimental procedures were in accordance with the European Communities Council Directive RL 2010/63/EC, and Use of Animals for Experimental Procedures, and the UK Animals Scientific Procedures Act. Recordings were made with 16-channel V-Probes (Plexon). These electrodes are conical in shape (tapering diameter of maximum 185 μm) and have a row of 16 circular platinum/iridium electrical contacts (15 μm diameter) with 150 μm center-to-center spacing. Visual stimuli were square-wave gratings centered on the receptive field location. We analyzed multi-unit activity of visually responsive channels. The number of visually-responsive channels per recording ranged between 8 and 12 (median 9). On-Off transitions synchronous across cortical layers were clearly observed in raw spike rasters during both spontaneous and stimulus-driven activity (fig. S7A) and manifested in a tri-phasic shape of spike auto- and cross-correlations (fig. S7B). The second additional dataset consisted of 212 single units recorded from area V4 with conventional single electrodes in two different monkeys during fixation with a static oriented-grating stimulus presented in neurons' receptive field (see supplementary methods 2.1 for details). In this additional single-unit dataset, spike auto-correlations also exhibited a very similar triphasic shape (fig. S2E), indicating that such correlations are not an artifact of the linear array microelectrodes.

We replicated all our HMM-based results on the additional laminar dataset. First, HMM captured on average half of the explainable variance (R^2) during spontaneous and stimulus-driven activity in this additional laminar dataset (fig. S7C,D; cf. fig. S4A,B,D,E). Second, spiking of individual units throughout the cortical depth was strongly modulated by the On-Off transitions: firing rates during the On and Off phases were significantly different (Wilcoxon signed rank test, $p < 0.05$) in the overwhelming majority of MUs (97% during stimulus-driven activity, 95.4% during spontaneous activity), and $\text{MI}_{\text{on-off}}$ was positive in nearly all MUs (99% during stimulus-driven activity, 98% during spontaneous activity), (fig. S7E; cf. fig. S4C). Third, for most recording sessions (17 total, 85%), the two-phase HMM was the most parsimonious model (fig. S7F,G; cf. fig. S5A,B). For all recordings, HMMs with more than 3 phases were never selected. The three-phase HMM was selected only for one recording with the relative decrease in error relative to the two-phase HMM of 10.12%, just hovering below the 10% acceptance criterion. For 2 (10%) recording sessions, one-phase HMM was the most parsimonious model. Fourth, across all recording sessions, the variance explained by the two-phase HMM (R^2) was

tightly correlated with the average spike-count correlation (fig. S7H, cf. fig. S5C). Thus results of HMM analyses were quantitatively very similar between the two laminar datasets.

3.5 Rate-matching control

We examined whether the observed increase in the On-episode duration could be confounded by higher trial-averaged firing rates in the attention conditions. Longer durations of On episodes contribute to higher trial-averaged firing rates, and thus the HMM fit might account for higher trial-averaged firing rates by estimating longer On episodes. This possibility is unlikely, however, because firing rates and transition probabilities (controlling episode durations) are all free parameters in the HMM and can be adjusted independently to fit the data. Nonetheless, we performed a rate-matching analysis to control for any potential biases.

For each unit, we equated the trial-average firing rates across four behavioral conditions (corresponding to four directions of the attention cue) by randomly deleting subsets of spikes from conditions with higher firing rates, and then fitted the HMM to the rate-matched data and estimated the durations of On and Off episodes. As expected, even when enhancement of the trial-average firing rate was eliminated (fig. S9B,C) the average duration of On episodes was still significantly longer during covert and overt attention than in control conditions (fig. S9D, covert: median change in duration 13 ms, $p < 10^{-3}$; overt: median change in duration 19 ms, $p < 10^{-5}$). Therefore, the increase in On-episode durations was not an artifact of the concurrent increase in the trial-averaged firing rates.

3.6 Modeling On-Off dynamics as a continuously varying latent state

3.6.1 Gaussian Process Factor Analysis (GPFA) model

While HMM describes the On-Off dynamics as instantaneous transitions between discrete latent phases, we also considered an alternative approach. We fitted our data by the GPFA model, where the latent state is assumed to be continuous and evolve smoothly in time (*26, 55*). Specifically, the one-dimensional latent state $x(t)$ is modeled as a Gaussian process with a smooth auto-correlation function parametrized by a single timescale τ:

$$\mathrm{Cov}[x(t_i), x(t_j)] = \exp\left(-\frac{(t_i - t_j)^2}{2\tau^2}\right).$$

A linear-Gaussian relationship is then defined between the latent state x and neural observations $y_k(t)$ ($k = 1, \ldots 16$):

$$y_k(t)|x(t) \sim \mathcal{N}(c_k x(t) + d_k, \sigma_k).$$

The time-series of neural observations $y_k(t)$ for channel k is obtained by measuring spike-counts in 10 ms bins and then performing the square-root transformation of these spike-counts, which is known to stabilize the variance. The parameters c_k, d_k, σ_k for each channel and the timescale parameter τ for the entire ensemble (49 parameters total) are estimated from data using the

expectation-maximization algorithm. Similar to the HMM, these estimated parameters are then used to decode on a trial-by-trial basis the most likely continuous latent trajectories that underlie the observed spike trains (*55*).

3.6.2 Correspondence between the On and Off episodes decoded by HMM and GPFA

We examined the correspondence between continuous trajectories decoded by GPFA and the sequences of discrete On and Off episodes decoded by HMM. By visual inspection, GPFA's latent trajectories were also closely aligned to the periods of vigorous and faint spiking (fig. S10A). During the On and Off episodes in spike rasters, GPFA's trajectories were flatter and did not change sign, while around the transition times, the trajectories switched sign and had a steeper slope. Accordingly, the distributions of GPFA's decoded latent state deviated from the *a priori* assumed Gaussian distribution and were clearly bimodal for many recordings (fig. S10B). For each empirical distribution of the latent state, we quantified deviation from the unimodal Gaussian distribution by computing Sarle's bimodality coefficient: BC $= (\gamma^2 + 1)/\kappa$, where γ is the skewness and κ is the kurtosis of the empirical distribution. BC varies between 0 and 1, where larger values indicate stronger bimodality and BC $= 1/3$ for the unimodal Gaussian distribution. Across recording sessions, BC of the GPFA's latent state correlated with the variance explained (R^2) by the two-phase HMM (fig. S10C). Moreover, the subset of recordings with BC $\approx 1/3$ nearly perfectly overlapped with the subset of recordings for which the one-phase HMM (i.e. single-Poisson process for each channel) was the most parsimonious model (fig. S10C, c.f. fig. S5). Thus, both discrete-state HMM and continuous-state GPFA models identified similar features in the spiking activity and were comparable in their ability to detect On-Off transitions.

Next, we examined how dynamics of the GPFA-decoded continuous trajectories corresponded to the On-Off transition dynamics determined by HMM. Times of zero-crossings, when GPFA's latent trajectories switched sign, closely coincided with the On-Off transition times detected by the HMM (fig. S10A). We therefore defined the On and Off episodes in GPFA's continuous latent trajectories as periods between two subsequent zero-crossings during which the trajectory stayed above and below zero, respectively. Such segmentation based on zero-crossings of the GPFA's latent trajectories resulted in On and Off episodes that closely matched the On and Off episodes determined by HMM (fig. S10B). We verified that the empirical distributions of the GPFA-decoded On- and Off-episode durations tightly overlapped with those decoded by HMM (fig. S10D). However, there was a difference in how these empirical distributions of episode durations were related to GPFA's and HMM's model parameters. HMM has two distinct timescale parameters τ_{on} and τ_{off}, and the empirical distributions of On and Off episode durations (decoded by either HMM or GPFA) tightly overlapped with the theoretical exponential distributions with the decay time-constants calculated from the fitted HMM transition probabilities (fig. S10D). GPFA, on the other hand, has only a single timescale parameter τ, which is optimized for maximal smoothness of latent trajectories that still preserves essential features in the data. This fitted GPFA timescale τ was systematically smaller than the time-

constants of empirical exponential distributions of On and Off episode durations (fig. S10D). This discrepancy is explained by the fact that transitions between the On and Off phases in the data occur much faster than the average dwelling times in these On and Off phases. Since GPFA has only a single timescale parameter τ, its value is estimated closer to the shortest timescale in the data, because otherwise excessive smoothing around transition times would reduce the fit accuracy. Accordingly, the average durations of On and Off episodes determined by HMM and GPFA models were very similar across recording sessions (fig. S10E, lower panels). These average episode durations were accurately captured by the τ_{on} and τ_{off} timescales of the HMM (fig. S10E, upper panels), whereas the GPFA's timescale parameter τ was systematically smaller than the average episode durations across recording sessions (fig. S10E, middle panels).

Therefore both HMM and GPFA represent adequate models capable to capture On-Off transitions in ensemble spiking activity. However, the dynamics decoded by HMM from the data agree with the HMM's *a priori* assumptions (exponential distributions of episode durations with time-constants matching HMM's parameters), whereas dynamics decoded by GPFA from data differ from the GPFA's *a priori* assumptions (GPFA's timescale parameter was systematically smaller than the empirical episode durations and GPFA's empirical latent-state distributions deviated from the unimodal Gaussian prior). These discrepancies together indicate that our data exhibit a timescale separation between transition and dwelling times, which is typical for metastable systems.

3.6.3 Attentional modulation of On-Off dynamics decoded by GPFA

We considered whether the observed increase in the On-episode duration could be confounded by HMM's *a priori* assumption of discrete On and Off phases in spiking activity. In a control analysis, we compared the average duration of On and Off episodes decoded from data by the GPFA model in attention and control conditions. The average duration of On episodes decoded by GPFA was significantly longer during covert and overt attention than in control condition (fig. S10F, upper panel, covert: median change in duration 12 ms, $p = 0.001$; overt: median change in duration 15 ms, $p < 10^{-4}$). Therefore, the increase in On episode durations was confirmed by the GPFA model, which assumes continuous unimodal fluctuations of the latent state instead of the discrete On and Off phases.

3.7 Relationship between On-Off transitions and microsaccades

Activity of neurons in the visual cortex is profoundly influenced by small fixational eye movements, called microsaccades (*56, 57*). In particular, V4 neurons exhibit strong excitatory responses within ~ 100 ms following a microsaccade (*56*). Therefore, we can expect that On-Off dynamics are also influenced by microsaccades. We quantified the relationship between On-Off transitions and microsaccades in order to assess whether microsaccades could account for the changes in the On-Off dynamics observed during attention.

We quantified the relationship between the On-Off transitions and microsaccades by computing cross-correlations between the times of microsaccades and times of On-to-Off and Off-to-On transitions. First, we detected microsaccades and found that, the median microsaccade amplitude was 12.7 and 14.7 arcmin in monkeys G and B, respectively (fig. S11A), similar to previous studies (*56, 57*). Then, we computed cross-correlations using the jitter-correction method (the same procedure that was used to compute spike auto- and cross-correlations, with 10 ms bin size and 100 ms jitter bin size). Cross-correlation between the times of microsaccades and Off-to-On transitions exhibited a small negative deflection around zero time-lag and a large positive peak at ~ 60 ms time-lag (fig. S11B, upper panel), indicating that the probability of transition from Off to On phase was enhanced ~ 60 ms following a microsaccade. On the other hand, cross-correlation between the times of microsaccades and On-to-Off transitions exhibited a small negative peak at ~ 60 ms time-lag (fig. S11B, lower panel), indicating that transitions from On to Off phase were less likely to occur ~ 60 ms following a microsaccade. This relationship between the microsaccades and On-Off transitions was not deterministic: state transitions occurred during epochs of highly stable gaze (no microsaccades, fig. S11C, lower panel), and conversely, not every microsaccade was followed by a transition (fig. S11C, upper panel). On average, only 11% of all Off-to-On transitions were preceded by a microsaccade and only 28% of microsaccades occurring during the Off phase were followed by an Off-to-On transition within a $30 - 75$ ms window relative to the time of microsaccade (fig. S11D).

While the relationship between the On-Off transitions and microsaccades was not deterministic, the existing (small) probabilistic relationship between them raises the question of whether the spatially selective changes in cortical state dynamics we observed could be due to microsaccades. Indeed, since the probability of Off-to-On transitions is enhanced and the probability of On-to-Off transitions is suppressed following a microsaccade, an increase in the On-episode durations could, in principle, result from an increase in the rate of microsaccades. However, to account for the spatially selective increase in the On-episode durations, the microsaccade rate should selectively increase only on trials when monkeys attended (covertly or overtly) to the receptive field (RF) location of recorded V4 neurons, which is unlikely since monkeys attended to one of four spatial locations on every trial. We compared the microsaccade rate between trials when monkeys attended to the V4 RF location and control trials (fig. S11E), and found no significant increase of microsaccade rate in all cases except the covert attention condition in monkey B, where the microsaccade rate was significantly higher than in control condition ($p = 0.002$, Wilcoxon signed rank test). However, a significant increase in the On-episode durations during attention was observed in all attention conditions in both monkeys, except for the covert attention condition in monkey B (supplementary table 1). Thus the only instance in which the microsaccade rate was higher in the attention than in the control condition was the one in which no increase in the On-episode durations was observed. In contrast, an increase in the On-episode durations was observed in all other conditions for which the microsaccade rate was the same.

We also considered a possibility that increases in On-episode durations could result from changes in the direction of microsaccades between attention and control conditions. We first

compared the microsaccade frequency between attention and control conditions, across microsaccade directions, aligned to the RF location (fig. S11F, the RF location corresponds to 0°, which aligns with the covertly attended stimulus in the covert condition, and is opposite to covertly attended stimulus in the overt condition). Indeed, the relative frequency of microsaccades between the attention and control conditions differed across directions (chi-squared test, monkey G: covert $p < 10^{-10}$, overt $p < 10^{-10}$; monkey B: covert $p < 10^{-10}$, overt $p < 10^{-10}$). However, the pattern of this effect differed between the two monkeys. Although there was an overall tendency for microsaccades to be more frequently directed towards the approximate location of the covertly attended stimulus, microsaccades were more frequently directed towards the location opposite the covertly attended stimulus in the covert condition for monkey B (chi-squared residuals test at 0.05 significance level with Bonferroni correction, 20° bins, centers of significant bins: monkey G: covert 20°, 40°, 60°, 80°; overt 220°, 260°; monkey B: covert 160°, 180°, 200°, 220°; overt 160°, 180°, 200°).

We then considered whether microsaccades towards the RF had a greater impact on On-Off transitions. Given the lack of an overall increase of microsaccade rate in attention above control conditions (with the exception of the covert condition for monkey B, fig. S11E, supplementary table 1), biases in the microsaccade frequency across directions should only be expected to contribute to the observed increase in On-episode durations if those biases coincide with increases in the rate of Off-to-On transitions following microsaccades in the bias directions, or with decreases in the rate of On-to-Off transitions following microsaccades in the bias directions (see fig. S11B). However, the fraction of microsaccades followed by Off-to-On transitions was not significantly different across microsaccade directions in either covert or overt conditions (fig. S11G, chi-squared test, covert $p = 0.367$, overt $p = 0.311$). The fraction of microsaccades followed by On-to-Off transitions was also not significantly different across microsaccade directions in the covert condition (chi-squared test, $p = 0.122$, histogram not shown since the overall fraction of microsaccades followed by On-to-Off transitions was low, 0.085). In the overt condition, the fraction of microsaccades followed by On-to-Off transitions was significantly lower only for one direction (-20° bin, chi-squared residuals test at 0.05 significance level with Bonferroni correction) that, however, was opposite the directions towards which the microsaccade frequency was increased in this condition (fig. S11F). Thus the rate of On-Off transitions was largely independent of microsaccade directions. Furthermore, the one direction for which there was a subtle change in the On-Off transition rates was not aligned with those directions for which the microsaccade frequency changed. Therefore changes in the microsaccade rate and direction cannot explain the increase of On-episode durations during attention.

To further rule out the possibility that the increase in On-episode durations was due to microsaccades, we also repeated all analyses for only the trials in which no microsaccades occurred during the analyzed time period. Although this procedure left us with less data, the pattern of results was unchanged (fig. S11H, cf. Fig. 3D). On-episode durations were significantly longer in the overt attention condition than in control (Wilcoxon signed rank test, overt: median change in duration 10 ms, $p = 0.005$), and On-episode durations were also longer in the covert attention condition relative to control and approached significance (Wilcoxon signed

rank test, covert: median change in duration 9 ms, $p = 0.122$). When the two attention conditions were combined, the increases in On-episode durations were highly significant (Wilcoxon signed rank test, median change in duration 10 ms, $p < 10^{-3}$). Thus, the increase of On-episode durations during attention was not due to microsaccades.

3.8 Relationship of On-Off dynamics to spike-count variability and correlations

It has been long known that responses of cortical neurons vary substantially across repeated measurements under identical experimental conditions (usually called trials). This response variability is classically quantified by counting spikes within some time-window T, and then calculating the mean and variance of these spike-counts N across trials for individual neurons, and correlations between spike-counts for pairs of simultaneously recorded neurons. Spike-count variability and correlations have been experimentally scrutinized using different time-windows T, typically ranging from 50 to $1,000$ ms across studies (*22, 23*). Many characteristic features of this correlated spike-count variability have been documented, yet the origin of these correlated fluctuations remained unknown. Since our work identifies On-Off dynamics as a major source of such correlated fluctuations in neural responses, it is important to establish how different features of spike-count statistics can be explained by knowing underlying On-Off dynamics.

3.8.1 Spike-count statistics in the model of On-Off dynamics

We used our two-state model of the On-Off dynamics to predict spike-count statistics measured over some arbitrary time-window T (fig. S12A). In the model, dwelling times in the On and Off phases are exponentially distributed with the average durations τ_{on} and τ_{off}, respectively. Neurons emit spikes as inhomogeneous Poisson processes with different mean rates r_{on} and r_{off} during the On and Off phases, respectively. Since the On-Off dynamics are the same for all neurons, we introduce a two-state process $r(t)$ switching between 0 (Off-phase) and 1 (On-phase). Firing rates $\lambda(t)$ of different neurons are then parametrized as

$$\lambda(t) = r_{\text{off}} + r(t)(r_{\text{on}} - r_{\text{off}}), \tag{5}$$

where the On and Off firing rates (r_{on} and r_{off}) can be different across neurons, but the switching process $r(t)$ is the same. On each trial, the number of spikes generated by a neuron during the time-window T is described by a Poisson random variable with the rate given by the integral

$$\Lambda = \int_{t}^{t+T} \lambda(t)dt. \tag{6}$$

The Poisson rate Λ fluctuates from trial-to-trial, because the proportion of On and Off phases within the time-window T varies from trial-to-trail (fig. S12A). Since the On-Off switching

process is the same for all neurons, it is convenient to separate the time-parameters $\tau_{\text{on}}, \tau_{\text{off}}$ and T (that are common for all neurons) from the On and Off firing-rate parameters (that vary across neurons) by introducing a normalized rate R:

$$R = \int_t^{t+T} r(t)dt. \tag{7}$$

The Poisson rate Λ for each neuron on every trial is then simply expressed as

$$\Lambda = r_{\text{off}}T + R\Delta r, \tag{8}$$

where we introduced the On-Off firing-rate difference

$$\Delta r = r_{\text{on}} - r_{\text{off}}. \tag{9}$$

In this On-Off model, all spike-count statistics of individual neurons are analytically calculated from just four model parameters: the timescales τ_{on}, τ_{off} and firing rates r_{on}, r_{off}, and spike-count correlations between pairs of neurons are analytically calculated from just six parameters (On and Off firing rates are needed for both neurons) for arbitrary time-window T. For each neuron, the mean and variance of the spike-count N are expressed through the mean and variance of the normalized rate R as:

$$\begin{aligned} \text{E}[N] &= r_{\text{off}}T + \Delta r\text{E}[R], & (10)\\ \text{Var}[N] &= \Delta r^2\text{Var}[R] + r_{\text{off}}T + \Delta r\text{E}[R]; & (11) \end{aligned}$$

and for a pair for neurons i, j, the spike-count correlation is expressed as:

$$r_{\text{sc}} = \text{Corr}[N_i, N_j] = \frac{\Delta r_i \Delta r_j \text{Var}[R]}{\sqrt{\text{Var}[N_i]\text{Var}[N_j]}}. \tag{12}$$

The mean $\text{E}[R]$ and variance $\text{Var}[R]$ of the normalized rate are analytically calculated as functions of τ_{on}, τ_{off} and T:

$$\text{E}[R] = \frac{\tau_{\text{on}}}{\tau_{\text{off}} + \tau_{\text{on}}}T, \tag{13}$$

$$\text{Var}[R] = \frac{2\tau_{\text{on}}^2\tau_{\text{off}}^2}{(\tau_{\text{off}} + \tau_{\text{on}})^3}\left[T - \frac{\tau_{\text{on}}\tau_{\text{off}}}{\tau_{\text{off}} + \tau_{\text{on}}}\left(1 - \exp\left(-\frac{\tau_{\text{on}} + \tau_{\text{off}}}{\tau_{\text{off}}\tau_{\text{on}}}T\right)\right)\right]. \tag{14}$$

3.8.2 On-Off dynamics predict multiplicative gain fluctuations

Previous studies have found that spike-count variability of cortical neurons can be well described by "gain-modulated Poisson" models, where neuron's firing rate on every experimental trial is modulated (multiplied) by a multiplicative gain factor that fluctuates across trials (*31,32*). The multiplicative noise (gain) in these models is an assumption based solely on the experimental observation that the variance-to-mean relation of spike-counts is accelerating, i.e. deviates

more and more from the Poisson expectation (unity line) as spike-count grows (see e.g., figs. 1,2 in Ref. (*31*) and fig. 1-supplement 2 in Ref. (*32*)). These models, designed to capture spike-count statistics over some fixed time-window T (e.g., $T = 1,000$ ms in Ref. (*31*) and $T = 200$ ms in Ref. (*32*)), explicitly assume that gain fluctuations occur on timescales slower than T and neurons' firing-rates are constant over the duration of each time-window T. This assumption entails that within-trial spike-time correlations should be exactly zero on timescales shorter than T, which, however, is not supported by the temporal profile of spike-time correlations measured on these short timescales (see figs. S2,S7B).

Gain-modulated Poisson models are functional models of neural response (such as those based on Generalized Linear Models) and their components do not have clear biophysical interpretation. Accordingly, these models do not provide insight into the sources or causes of apparent fluctuations in the multiplicative amplification signals. In contrast, our work identifies a new phenomenon, synchronous On-Off transitions, which appear to represent exactly such a source of correlated variability in neural responses. It is therefore natural to ask whether spike-count variability arising from the On-Off dynamics is consistent with the multiplicative gain fluctuations. Note that our HMM (and GPFA) models track the On-Off firing-rate fluctuations with much finer temporal resolution (10 ms bins) than the time-windows over which apparent gain-fluctuations have been observed ($200 - 1,000$ ms). Hence, our two-state On-Off model can predict spike-count statistics over these longer time-windows, but not the other way around: multiplicative gain-fluctuations *per se* do not predict existence of On-Off transitions.

We calculated analytically the spike-count variance and mean for the two-state On-Off model and found that the model predicts the accelerating variance-to-mean relation that has been interpreted as evidence for multiplicative gain fluctuations (fig. S12B). We considered two scenarios. In the first scenario, we computed the mean and variance for the same neuron over variable duration time-windows (ranging from 1 to 316 ms). We used the two example neurons from fig. S12A, which follow through exactly the same sequence of On-Off episodes on every trial, but have different On and Off firing-rates. Variance-to-mean relation is accelerating for both neurons, but it is more rapidly accelerating for the neuron with greater difference $\Delta r = r_{\text{on}} - r_{\text{off}}$ (red and blue lines in fig. S12B). In this scenario, where the spike-count mean and variance are changing because the time-window T is changing, the On-Off model predicts that the mean $\mathrm{E}[R]$ increases linearly with T, while the variance $\mathrm{Var}[R]$ increases quadratically with T, for small T (it saturates to a linear dependence for large T, see equations 13,14). Hence the model predicts a quadratic departure from the Poisson prediction (unity line) that is faster for neurons with greater Δr (see equation 11).

In the second scenario, we computed the mean and variance of spike-count over a fixed time-window across a hypothetical population of 2,000 neurons, which are all modulated by the same On-Off dynamics, but whose On and Off firing rates vary across the population (gray dots in fig. S12B). In this scenario, $\mathrm{E}[R]$ and $\mathrm{Var}[R]$ are the same for all neurons, but the mean and variance of spike-count are varying because of the variation in r_{off} and Δr across the population. The variance-to-mean relation is accelerating, because the mean spike-count increases linearly with Δr while the spike-count variance increases quadratically with Δr (see equations 10,11).

Note that in actual physiological measurements the variance-to-mean relation is expected to be more scattered than the trend predicted by the On-Off model due to variation in On and Off timescales across recordings and also because empirical estimates of variance from just a few hundred trials are very noisy. In summary, our two-state model predicts the signature of multiplicative gain fluctuations in spike-count statistics and explains how it emerges from the underlying On-Off dynamics.

3.8.3 On-Off dynamics predict spike-count correlations

Our two-state On-Off model predicts that spike-counts of pairs of neurons are correlated across trials, because the neurons follow through the same sequence of On and Off episodes on every trial (fig. S12C). The value of spike-count correlation r_{sc} for each pair can be analytically calculated from six model parameters (equation 12). As the equation shows, the value of r_{sc} depends not only on the variance of the common On-Off switching process (Var$[R]$), but also on the amplitude of modulation by these On-Off dynamics for each neuron (i.e. Δr_i and Δr_j in the pair). Thus even when all neurons in the population are synchronized to the same On-Off dynamics, the spike-count correlations vary across pairs due to variation in neurons' On and Off firing rates.

Our On-Off model predicts, that when spike-count correlations are studied across multiple conditions (e.g., in response to different sensory stimuli), the value of r_{sc} should change as a function of condition-dependent firing-rate changes even if the underlying On-Off dynamics would not change across conditions (although we find that On-Off dynamics change as a function of stimulus condition as well, fig. S6). Typically, the spike-count correlations are studied as a function of the geometric mean of mean firing rates of two neurons across conditions, and it has been observed that there is no consistent relationship between changes in the geometric mean rate and spike-count correlation: for different pairs of neurons r_{sc} can decrease, increase or not change at all as a function of the geometric mean rate (*31*). These different trends have been fitted by the gain-modulated Poisson model augmented with two additional phenomenological parameters: the point-process correlation and the gain correlation parameters (*31*). These two extra parameters, fitted separately for each pair, are key for capturing the different spike-count correlation trends, but their physiological origin is unclear (as it is usually the case with functional models).

In contrast, our On-Off model can analytically explain different spike-count correlation trends simply from how On and Off firing rates of neurons change across conditions (equation 12). If changes in mean firing rate across conditions are mainly driven by changes in r_{off} and r_{on}, but without substantial change in the On-Off firing-rate differences Δr, the model predicts decrease in spike-count correlations with increase in geometric-mean rate (fig. S12D). On the other hand, if changes in mean firing-rate across conditions are mainly driven by changes in Δr, a positive trend is predicted (fig. S12E). In neurophysiological data, we expect to see a heterogenous mixture of On and Off firing rate changes across conditions for different neurons, as well as changes in the On-Off dynamics themselves across conditions, which would imply a

broad distribution of how spike-count correlations depend on mean firing rates.

3.8.4 On-Off dynamics account for attention-related changes in spike-count correlations

Several recent studies have shown how spike-count variability and correlations can be affected by cognitive factors such as attention, learning and task structure (*24, 25, 58*). The most widely known example is the reduction in spike-count correlations between visual cortical neurons during attention, first reported by Mitchell et al. (2009) (*27*) and Cohen and Maunsell (2009) (*28*). Both these studies found that attention-related reduction in spike-count correlations improved signal-to-noise ratio of pooled neural signals substantially more than concurrent increase in mean firing rates, suggesting that decorrelation may be a universal signature of attention. However, subsequent studies showed that changes in spike-count correlations during attention can be much more versatile. In particular, spike-count correlations were found to flexibly increase or decrease during attention depending on whether the neurons provided evidence for the same or opposite choices (*29*). Moreover, spike-count correlations were found to increase between pairs of neurons with overlapping receptive fields situated in cortical areas V1 and MT when attention was directed within their common receptive field, while, at the same time, spike-count correlations decreased between neurons within each area (*30*). In combination these results show that the effects of spatial attention on spike-count correlations are more complex than previously thought and can follow opposing trends (increase vs. decrease) depending on the structure of a behavioral task and on the relative positioning of neurons in the cortex. Since these different trends in spike-count correlations could in principle arise from appropriate changes in On-Off dynamics (fig. S12D,E), we wondered (i) how spike-count correlations changed during attention in our dataset, and (ii) whether these changes in spike-count correlations can be accounted for by the On-Off dynamics model.

Changes in spike-count correlations during covert and overt attention. We examined attention-related changes of spike-count correlations in our dataset. Compared to the original work by Mitchell et al. (2009) and Cohen and Maunsell (2009), our experimental design incorporated two notably different approaches (*14*). First, we recorded across cortical layers from neurons largely within a single column, and with tightly overlapping receptive fields. Second, in our task, the focus of covert attention was behaviorally dissociated from the target of an upcoming saccade (overt attention). Thus attentional demands, as well as relative positioning of neuronal receptive fields in our dataset differed from previous studies. Since both factors are likely to affect attention-related changes in spike-count correlations, we could have not predicted how spike-count correlations should change during covert and overt attention in our experiment.

We measured changes in spike-count correlations of MUA during covert and overt attention relative to control conditions using $T = 200$ ms window (400 to 600 ms after the attention-cue onset) and found a broad distribution of changes across pairs (fig. S13A). Changes in spike-count correlations varied a great deal from pair to pair (standard deviation of changes: 0.21

for both types of attention). However, their distribution across the population was nearly symmetrical around zero and the net change in spike-count correlations was small, yet significant: on average spike-count correlations slightly increased during covert attention and slightly decreased during overt attention (covert: mean change 0.011, $p < 10^{-4}$; overt: mean change -0.005, $p < 10^{-4}$; permutation test with 10^4 shuffles). Thus in our data, changes in spike-count correlations during covert and overt attention were broadly distributed and on average had small opposing trends.

On-Off model accounts for changes in spike-count correlations. The observation of opposing average changes in spike-count correlations during covert and overt attention may seem difficult to reconcile with the increase in On-episode durations, which was observed in both attention conditions. Intuitively, an increase in On-episode durations is expected to reduce the variance of the On-Off switching process (Var$[R]$) and, as a consequence, to reduce spike-count correlations (equations 12,14). Indeed, when all other parameters are held constant, the average measured increase in On-episode durations predicts a very small reduction in spike-count correlations (fig. S13B).

However, the On-Off model also makes a very specific prediction that the magnitude of spike-count correlation should increase with the magnitudes of On-Off firing-rate difference ($\Delta r = r_{\text{on}} - r_{\text{off}}$). Specifically, the magnitude of spike-count correlation between neurons i and j is directly proportional to the product $\Delta r_i \Delta r_j$ (equation 12). This is intuitive, because the only source of spike-count correlations in our On-Off model is the common switching process, hence the stronger a neuron is modulated by the On-Off dynamics (the greater is its Δr), the stronger it will be correlated with other neurons in the population. As a consequence, the change in spike-count correlation is predicted to be proportional to the change in $\Delta r_i \Delta r_j$ during attention. This prediction was clearly borne out by the data: we found that the measured change in spike-count correlations had a strong trend as a function of the change in the pair's On-Off firing-rate difference defined as $\sqrt{\Delta r_i \Delta r_j}$ (y-axis vs. color-axis in fig. S13C). Moreover, changes in spike-count correlations measured from the data were accurately matched by the changes in spike-count correlations predicted by the On-Off model, i.e. computed on a pair-by-pair basis using the fitted HMM parameters: τ_{on}, τ_{off}, r_{on}, and r_{off} (y vs. x-axis in fig. S13C). We observed that attention-related changes in spike-count correlations ranged widely across the population and that this broad distribution of changes was accurately matched by predictions of the On-Off model.

Thus, although the increase in On-episode durations predicts a slight reduction in spike-count correlations during attention, another important factor determining changes in spike-count correlations is change in pair's On-Off firing-rate difference (defined as $\sqrt{\Delta r_i \Delta r_j}$). In our data, changes in a pair's On-Off firing-rate difference were broadly distributed but on average they were slightly, yet significantly, more positive for covert than for overt attention (inset in fig. S13C, covert: median change 1.5 Hz; overt: median change 0.3 Hz, $p < 10^{-4}$, permutation test with 10^4 shuffles). This slight difference was sufficient to explain small opposing trends observed on average in spike-count correlation changes: the On-Off model predicted on average

slight increase in spike-count correlations during covert attention and slight decrease during overt attention (covert: mean predicted change 0.011, overt: mean predicted change -0.004). Interestingly, the dependence of changes in spike-count correlations on changes in pair's On-Off firing-rate difference was very similar for the covert and overt attention conditions (fig. S13C), and the difference observed on average between these two attention conditions was merely a result of slightly different changes in firing rates, rather than of a more fundamental difference in structure of population activity.

In addition, note that changes in spike-count correlations did not systematically depend on changes in geometric mean firing rate of the pair (fig. S13D). When the data were grouped according to changes in geometric mean firing rates, the average change in spike-count correlations for each group was very small (slightly positive and slightly negative for covert and overt attention, respectively). The broad distribution of spike-count correlation changes was masked in this representation, although excellent agreement between the data and model predictions was still observed.

3.8.5 Relationship to previous work

Attention-related changes of spike-count correlations in our laminar recordings are accurately accounted for by the On-Off dynamics model, but differ from the previously observed reductions in spike-count correlations (*27, 28*) (though changes during overt attention are similar). Among the experimental differences that may account for why our observations differ from those of Mitchell et al. (2009) and Cohen and Maunsell (2009), the most parsimonious may be the difference in the lateral separation between neurons in their recordings and ours. Spike-count correlations are known to decrease sharply with lateral separation, which explains why the overall magnitude of spike-count correlations in our data (median MUA correlation 0.318) is greater than typically reported for laterally separated neurons in V4 ($\sim 0.05 - 0.2$) (*23*). Moreover, the decrease in spike-count correlations with lateral separation may be itself explained by a limited spatial coherence of the On-Off states, which is expected given the local modulations of On-Off dynamics during attention that we observe. Indeed, the On-Off model predicts three factors that could contribute to attention-related reduction in spike-count correlations: 1) increase in On-episode durations, 2) changes in the On and Off firing rates across the population, such that $\Delta r_i \Delta r_j$ is reduced on average, 3) changes in how the On-Off fluctuations are coordinated laterally across cortex, such that their spatial coherence is reduced. The last factor is likely to contribute substantially to reduction in spike-count correlations among laterally separated neurons. Testing this hypothesis and identifying relative contributions of these three factors remains an important area for future study, which would require understanding how attention interacts with spatial coordination of On-Off dynamics, potentially through dense recordings at multiple laterally separated locations in the cortex.

A recent model by Rabinowitz et al. (2015) (*32*) has parsimoniously attributed reduction in spike-count correlations during attention to reduction in fluctuations of shared modulatory signals. This functional model of population activity, based on a gain-modulated Poisson

model (*31*) with a set of modulatory gain signals shared across neurons, was fitted to Cohen and Maunsell (2009) dataset. The variance of the shared gain in the fitted model was found to decrease during attention, which accounted for some (but not all) of the reduction in spike-count correlations in the data. The model thus provides a quantitative framework encapsulating correlated spike-count variability, but it leaves unsettled what causes apparent fluctuations in shared gain and what mechanisms reduce their variance during attention. Similar to the discussion above, our model of On-Off dynamics suggests three possible factors that could underlie the apparent reduction in gain variance: changes in the timescales, firing rates and spatial coherence of the On-Off states. Determining the relative contributions of these factors remains an important area for future study.

The shared-gain model and our On-Off model may appear to engage different neural mechanisms. This is because we model neural phenomena occurring on much shorter timescales than phenomena modeled by Rabinowitz et al. (2015). Over longer timescales (trial-to-trial), both models indeed capture the same neural phenomenon (e.g., see fig. S12B), but our On-Off model also captures phenomena on shorter timescales (within single trials) that are not modeled by Rabinowitz et al. (2015). On the other hand, Rabinowitz et al. (2015) modeled data recorded from neurons more broadly distributed across the cortical surface, and across hemispheres, and therefore their model may capture features of variability not present in our recordings from single cortical columns. In that way, the two studies are complementary. Whereas ours provides finer temporal resolution, theirs covers a broader spatial scale. Moreover, our data provide two substantial insights into mechanisms of selective attention, beyond what could have been inferred from an apparent reduction in shared-gain variability invoked by Rabinowitz et al. (2015) study. First, the main contribution of our work is the observation that cortical-state dynamics are locally modulated during attention, which could not have been predicted from reduction in shared-gain variability. Reduction in gain variance could arise from mere changes in firing rates during attention without any change in the underlying On-Off dynamics (consistent with the first scenario in Fig. 3B). Second, trial-to-trial fluctuations of shared-gain, which are predictive of behavioral performance, have been interpreted by previous studies as arising mainly from fluctuations in attentional amplification signals (*32, 33*). Our results show that, in addition, behavioral performance is influenced by endogenous fluctuations on faster timescales, which exist independently of sensory or cognitive events, but can interact with these events.

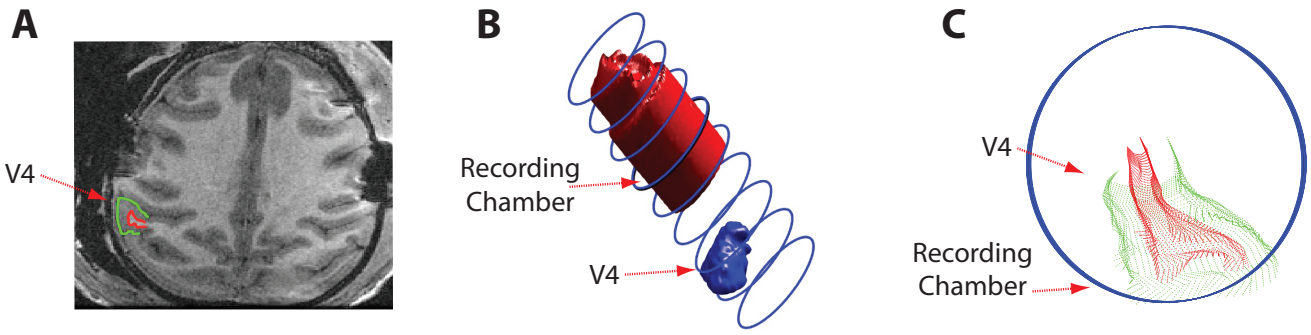

Fig. S1. MRI targeting of V4 recordings. MRI targeting was used to calculate electrode approach vectors perpendicular to V4 cortical layers. (**A**) Coronal section of MRI image from monkey G. The border of V4 with white matter (red) and pial surface (green) were manually identified. The apparent gaps in the MRI image (left and right side, anterior to V4) are "shadows" caused by the titanium skull screws used to secure the implanted headpost. (**B**) The relative position of the recording chamber and area V4 in three dimensional space. (**C**) The location of V4 within the recording chamber, viewing along the axis of the recording chamber as if looking down through it.

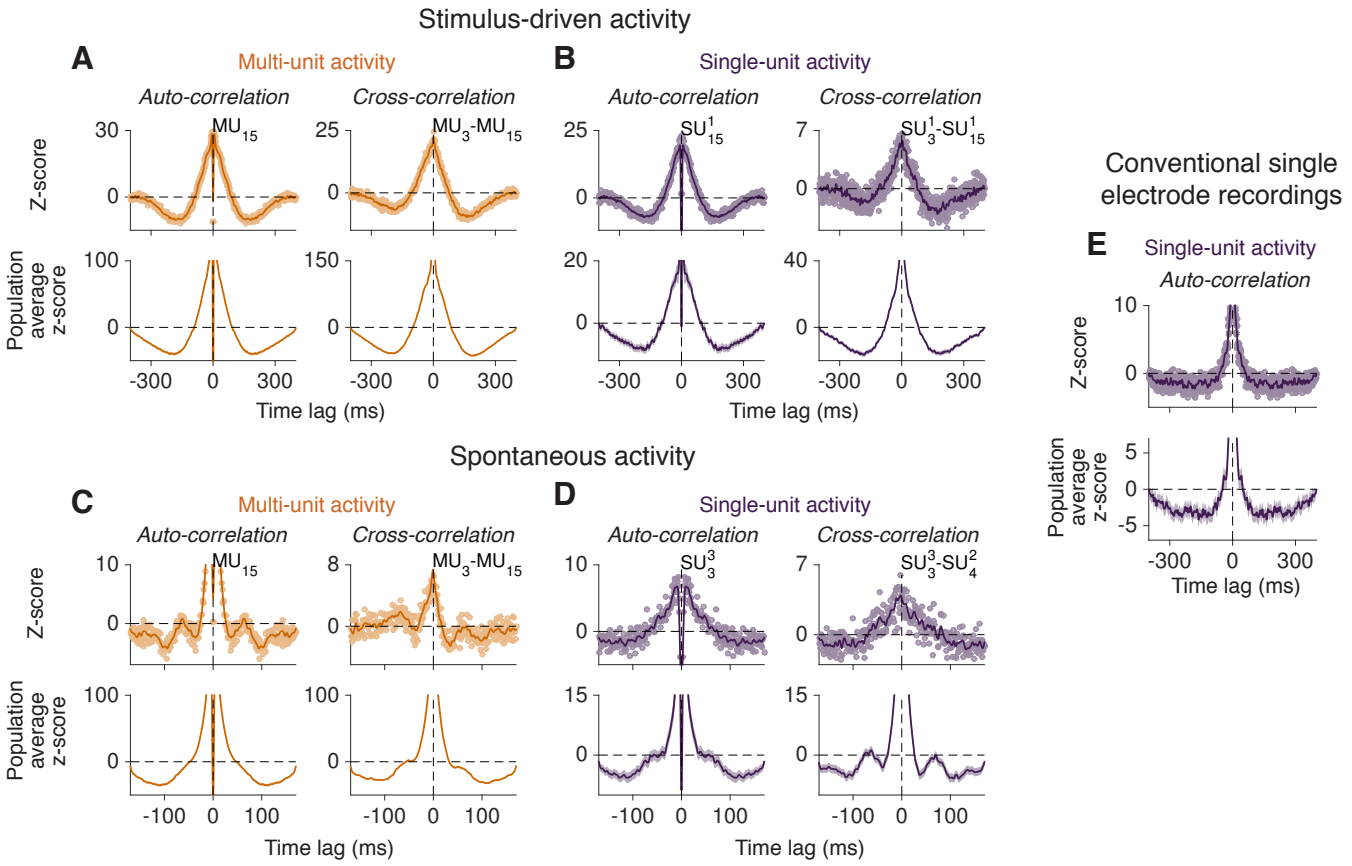

Fig. S2. Synchronous On-Off transitions are manifested in a tri-phasic shape of spike auto- and cross correlations. (**A**) Auto- and cross-correlations for stimulus-driven activity during the attention task. Upper panels: example auto- (left) and cross-correlation (right) of multi-unit activity on channel 3 and channels 3 and 15 (1.8 mm apart in depth), respectively, from the recording shown in Fig. 1c, during the time epoch when both stimulus and attention cue were present. Dots are raw correlation values, and solid line is a smoothed correlation drawn to guide the eye. Lower panels: population-average auto- and cross-correlations across all units and unit-pairs, respectively; shaded area depicts ±sem. (**B**) Same as A for single-unit activity. (**C**) Same as A, but for multi-unit activity during the fixation period of the attention task before stimulus was presented. (**D**) Same as C, but for single-unit activity. (**E**) Same as left panel in B, but for a different dataset of 165 single units recorded with conventional single electrodes in area V4 under similar behavioral conditions (dataset from ref. (*46*)). Auto-correlations are computed during fixation with a static oriented-grating stimulus presented in the neurons' receptive field.

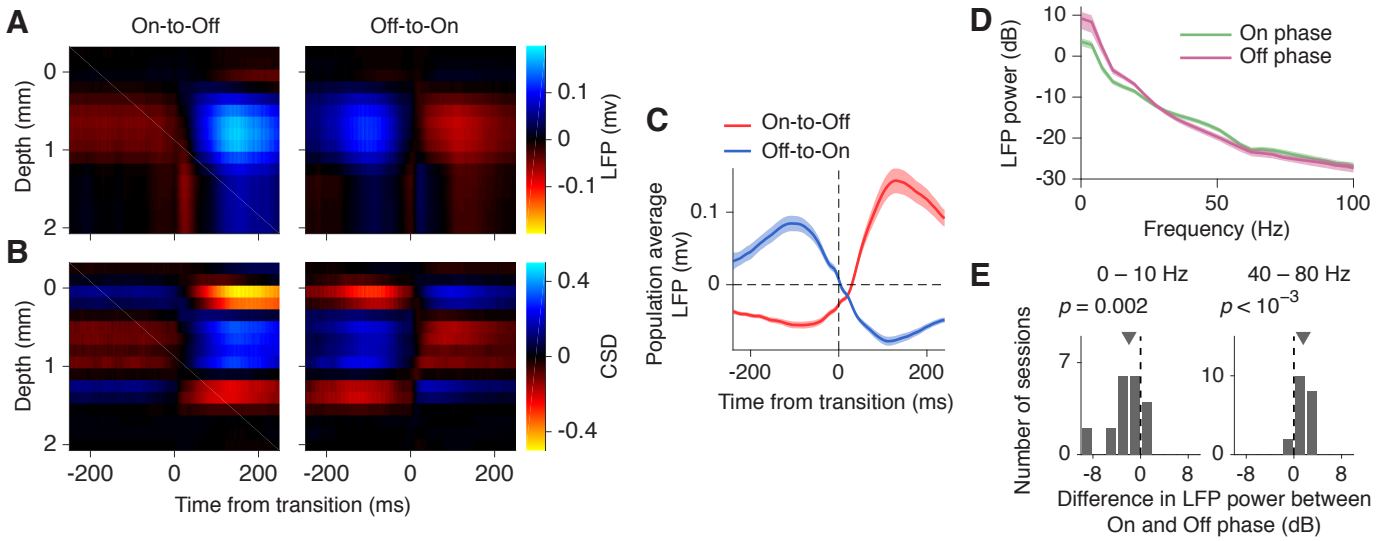

Fig. S3. Relationship between local field potentials (LFPs) and On-Off transitions. (**A**) Average local field potentials aligned to the times of On-to-Off (left panel) and Off-to-On (right panel) transitions for an example recording session. LFPs for 16 recording channels are ordered in depth from superficial (top) to deep (bottom). (**B**) Average current source densities (CSDs) aligned to the times of On-to-Off (left panel) and Off-to-On (right panel) transitions for the example session from panel A. CSD exhibits a characteristic pattern of sources and sinks alternating through the depth, which flips polarity around the time of On-to-Off and Off-to-On transitions. (**C**) LFPs aligned to the times of On-to-Off (red line) and Off-to-On (blue line) transitions averaged across all channels and all recordings. LFP exhibits a clear switch of polarity around the times of On-to-Off and Off-to-On transitions in spiking activity. Shaded area represents s.e.m. across recordings. (**D**) LFP power shifts between the On and Off phases. LFP power is suppressed at low and enhanced at gamma-range frequencies during the On relative to Off phases. Power spectra were computed for all On and Off episodes of at least 250 ms duration, and then averaged first across all episodes and then across recordings. Shaded area indicates s.e.m. across recordings. (**E**) Distribution of the difference in LFP power between the On and Off phases for low ($0 - 10$ Hz, left panel) and gamma ($40 - 80$ Hz, right panel) frequency ranges across recordings. Low-frequency LFP power is higher during Off phases, whereas gamma-range LFP power is higher during On phases. p-values are for the Wilcoxon signed rank test.

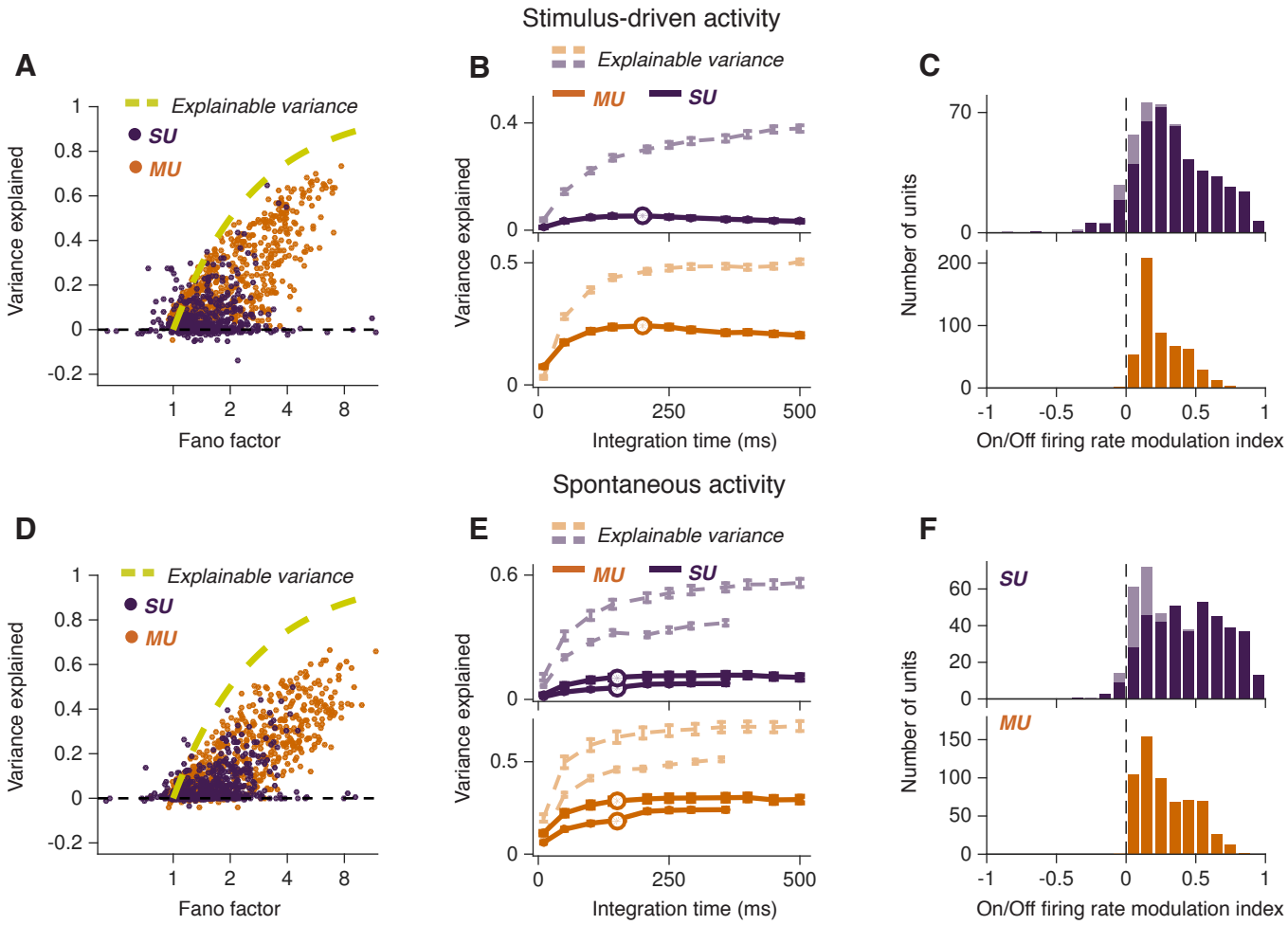

Fig. S4. Activity modulation by On-Off transitions. (**A**) Scatter plot of the variance explained by the HMM versus spike-count variability measured by the Fano factor for single- (SU, purple dots) and multi-units (MU, orange dots). Each dot represents the average across all trials of one recording. Dashed gold line depicts the maximal explainable variance for each Fano factor value. (**B**) Population-average of the variance explained by the HMM as a function of the integration time window for single- (SU, purple line) and multi-units (MU, orange line). Pale-colored lines depict the corresponding maximal explainable variance. White circles indicate 200 ms integration time window that was used in panel A. Error bars are $\pm$sem. (**C**) Distribution of the On/Off firing rate modulation index ($[r_{\text{on}} - r_{\text{off}}]/[r_{\text{on}} + r_{\text{off}}]$) for single- (SU, purple) and multi-units (MU, orange). Dark shading in the histograms corresponds to units whose firing rate was significantly modulated by On-Off transitions (all multi-units were significantly modulated). Single-units with modulation index ≈ 1 almost never fire spikes during the Off phase. (**D-F**) Same as in A-C, but for spontaneous activity recorded during fixation period of the attention task and during fixation task (five additional recording sessions), corresponding to two sets of lines in panel E. White circles indicate 150 ms integration time window that was used in panel d. Due to shorter fixation period in the attention task, integration time windows were tested only up to 350 ms and 150 ms for monkeys G and B, respectively.

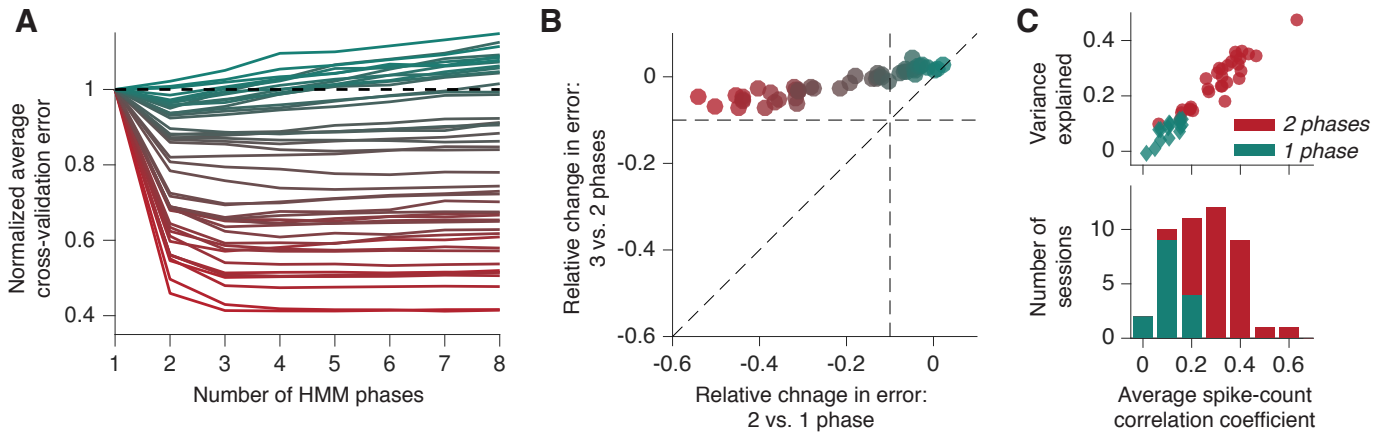

Fig. S5. Selecting the number of latent phases in the HMM. (**A**) Average cross-validation error for HMMs with n phases ($n = 1, \ldots 8$), normalized by the average cross-validation error of the HMM with 1 phase (equivalent to a single Poisson process). For each recording session, the 4-fold cross-validation error was computed in 200 ms windows for each condition (4 attention conditions $\times$ 8 stimulus orientations), and then averaged across all channels, conditions and cross-validation folds. Each line represents one recording session. For most recordings (red lines), addition of the second phase greatly reduced cross-validation error compared to the single-phase HMM, whereas adding more phases resulted in only marginal improvements: the error curves display an elbow at $n = 2$, suggesting that 2-phase HMM is the most parsimonious model for our data. For some recordings (teal lines), HMMs with $n > 1$ phases did not perform better than a single-phase HMM. (**B**) Relative change in error from adding an additional phase to the HMM. Relative change in error is defined as the difference in normalized cross-validation error for HMMs with $n + 1$ and n phases. Large negative differences indicate large reduction in the cross-validation error (improvement of the model fit). The scatter plot shows the change in error for 2- relative to 1-phase HMM (x-axis) versus 3- relative to 2-phase HMM (y-axis). Each dot represents one recording session, colors are the same as in panel a. Dashed vertical and horizontal lines indicate the criterion, which was used to determine the number of phases in the most parsimonious HMM: the number of phases was increased only if adding an additional phase reduced the cross-validation error by more than 10%. For most recordings (red circles), the 2-phase HMM was the most parsimonious model. For some recordings (teal circles), 1-phase HMM (single Poisson process) was the most parsimonious model. These recordings did not show clear evidence for On-Off transitions and therefore were excluded from the subsequent analyses of On and Off episodes durations. (**C**) Upper panel: Scatter plot of the variance explained by the 2-phase HMM versus average spike-count correlation (i.e. noise correlation) across recording sessions (red circles and teal diamonds: 2-phase and 1-phase recordings, respectively). Spike-count correlations were computed as Pearson correlation coefficients between spike counts of two nonadjacent MUA channels within the last 200 ms of the attention period (i.e. before stimulus offset) for trials of the same condition, and then averaged across all pairs for each recording. Lower panel: distribution of spike-count correlations across recordings (red and teal bars: stacked histograms for 2-phase and 1-phase recordings, respectively). Recordings for which 1-phase HMM was the most parsimonious model, consistently showed low spike-count correlations.

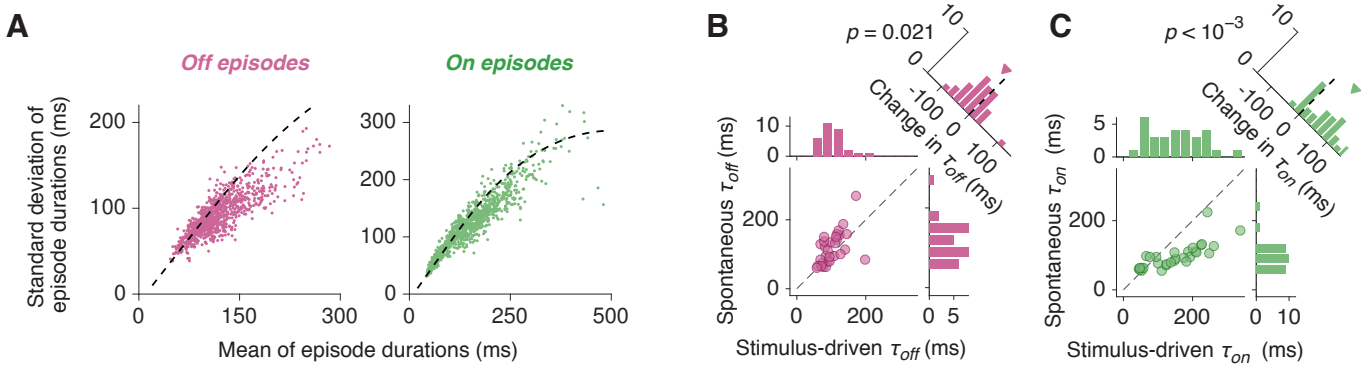

Fig. S6. On episode durations increase during visual stimulation. (**A**) Relationship between the measured mean (x-axis) and standard deviation (y-axis) of On and Off episode durations (green and pink dots, respectively) during stimulus-driven activity. Each dot represents one condition for one recording. The measured mean-to-standard-deviation relationship agrees well with the theoretical expectation for an exponential distribution truncated between 10 ms and 1,000 ms (black dashed line), which correspond to the minimal (HMM bin-size) and maximal (typical duration of the analysis time-window) possible measurement of an episode duration, respectively. (**B**) Scatter plot of the Off timescales τ_{off} during stimulus-driven (x-axis) versus spontaneous activity (y-axis). Each dot corresponds to a single recording. Flanking histograms show the marginal distributions. The Off timescale τ_{off} is slightly shorter during visual stimulation than during spontaneous activity (upper right histogram obtained by computing the difference in τ_{off} between the stimulus-driven and spontaneous activity for each recording; triangle indicates median of the distribution; p-value is for Wilcoxon signed rank test). (**C**) Same as in A, but for On timescales τ_{on}. The On timescale is approximately twice as long during visual stimulation than during spontaneous activity.

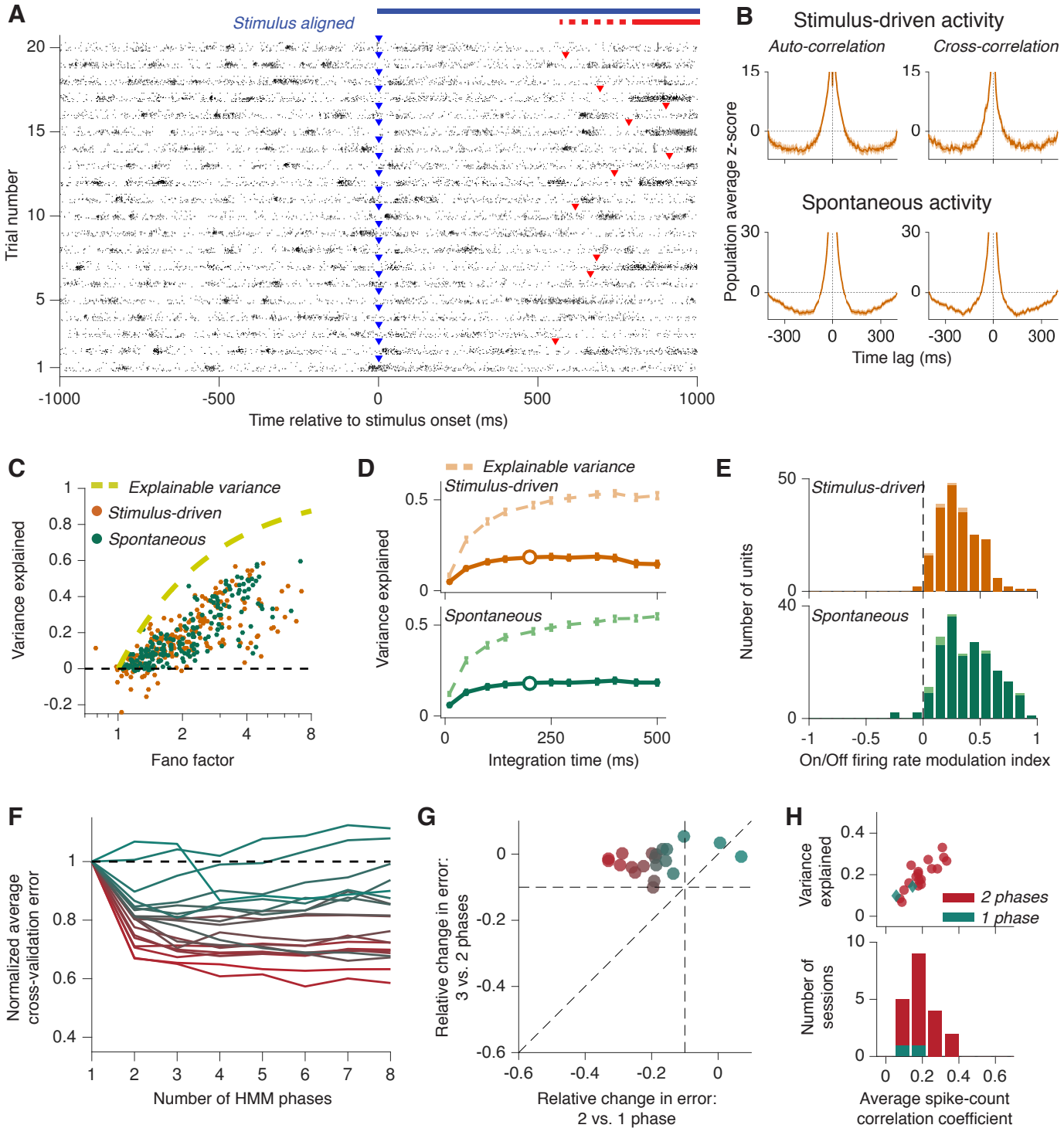

Fig. S7. Replication of On-Off transitions in laminar recordings from area V4 performed with a different type of linear array electrodes, in two different behaving monkeys, and in a different laboratory. (**A**) An example recording showing spontaneous On-Off transitions in spiking activity, which occur synchronously across cortical layers while monkey is fixating (cf. Fig. 1e). Multi-unit activity is shown for

twenty example trials (horizontal bands), with spike times on each of 12 simultaneously recorded channels indicated by dots in the corresponding row within each band. Activity is aligned to the stimulus' onset time (indicated by blue triangles). Time period prior to stimulus onset corresponds to spontaneous activity. Red triangles indicate onset time of stimulus dimming event. The durations of stimulus and dimming event are indicated above the plot by blue and red lines, respectively; red dashed line indicates randomized dimming onset time. On-Off transitions were not locked to these behavioral events. (**B**) Same as Figs. S2A,C. (**C**) Same as Fig. S4A for spontaneous (green dots) and stimulus-driven (orange dots) MUA. (**D**) Same as Fig. S4B for spontaneous (green) and stimulus-driven (orange) MUA. (**E**) Same as Fig. S4C for spontaneous (green) and stimulus-driven (orange) MUA. (**F**) Same as Fig. S5A. (**G**) Same as Fig. S5B. (**H**) Same as Fig. S5C.

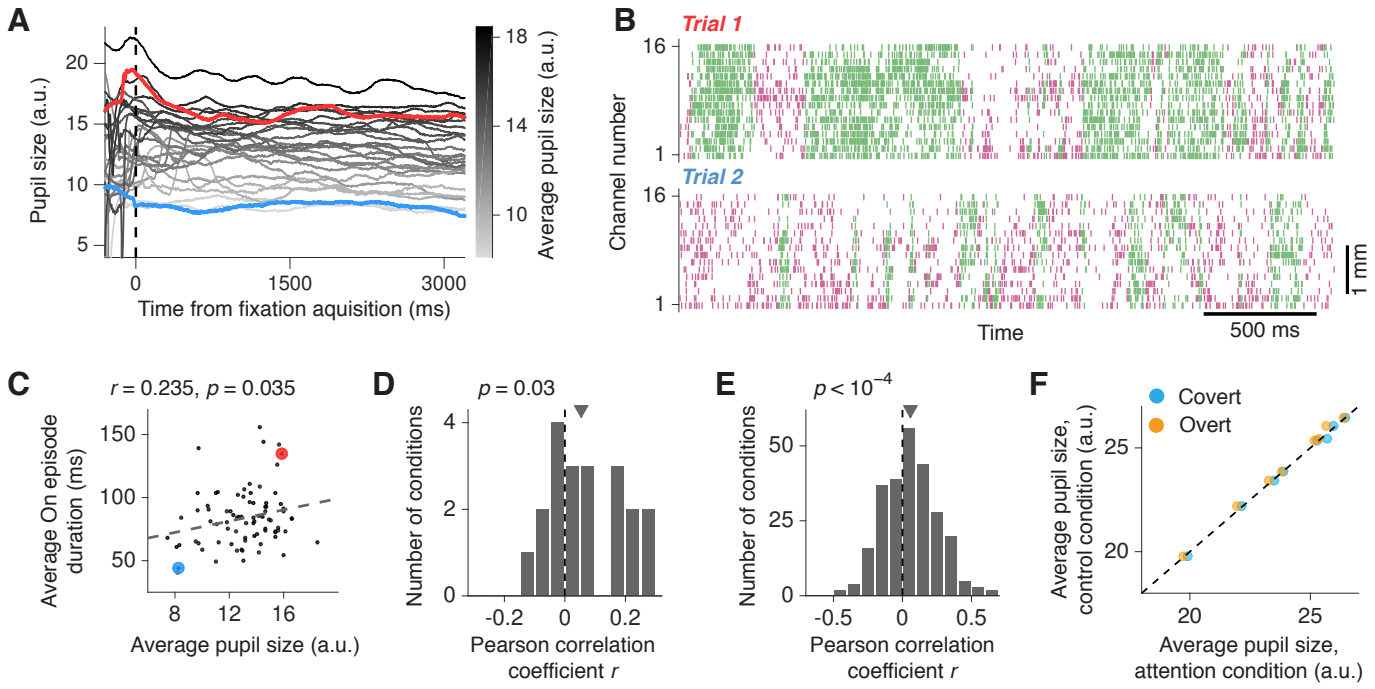

Fig. S8. On-Off dynamics correlate with pupil size during fixation and attention tasks. (**A**) Example traces of pupil size measurements during fixation task in monkey G on trials without visual stimulation (the same task as in Fig. 1b and the same example session as in Fig. 1c). Light-to-dark color gradient indicates the average pupil size during 3 seconds of fixation on each example trail. Pupil size is relatively stable over the 3 s trial duration, but varies substantially on a slower timescale across trials. Red and blue traces indicate two example trials with relatively large and small pupil size, respectively, which correspond to the example trials in panels B and C. (**B**) Average duration of On episodes is greater on the example trial with larger pupil size than on the example trial with smaller pupil size. Green and pink ticks indicate spikes occurring during the On and Off episodes, respectively, as decoded by the HMM. (**C**) Average duration of On episodes is positively correlated with the average pupil size for the example session from panel a. Each dot in the scatter plot represents a single fixation trial without visual stimulation, r is the Pearson correlation coefficient. (**D**) Across sessions, average duration of On episodes is positively correlated with the average pupil size during fixation task in monkey G. Pearson correlation coefficient r was calculated separately for each of 4 stimulus conditions (no visual stimulation and grating stimulus at three contrast levels) in 5 recording sessions during fixation task, and the histogram shows the distribution across resulting 20 correlation coefficients. Triangle indicates the median of the distribution; p-value is for Wilcoxon signed rank test. (**E**) Average duration of On episodes is positively correlated with the average pupil size during attention task in monkey B. Pearson correlation coefficient r was calculated during the same time window that was used to fit the HMM (400 ms after attention cue onset until the end of the attention period), separately for each condition (32 conditions: 8 stimulus orientations times 4 attention conditions) in 8 recording sessions where the pupil size was monitored, and the histogram shows the distribution of these 256 correlation coefficients. Triangle indicates the median of the distribution; p-value is for Wilcoxon signed rank test. (**F**) Scatter plot of the average pupil size in attention (x-axis; covert - blue circles, overt - orange circles) versus control (y-axis) conditions. Each dot represents a recording session. The average pupil size was not significantly different between covert attention

and control conditions ($p = 0.55$, Wilcoxon signed rank test), and was smaller in overt attention than in control condition (median difference -10 a.u., $p = 0.008$, Wilcoxon signed rank test). Therefore, differences in the pupil size cannot explain the increase of On-episode durations during attention.

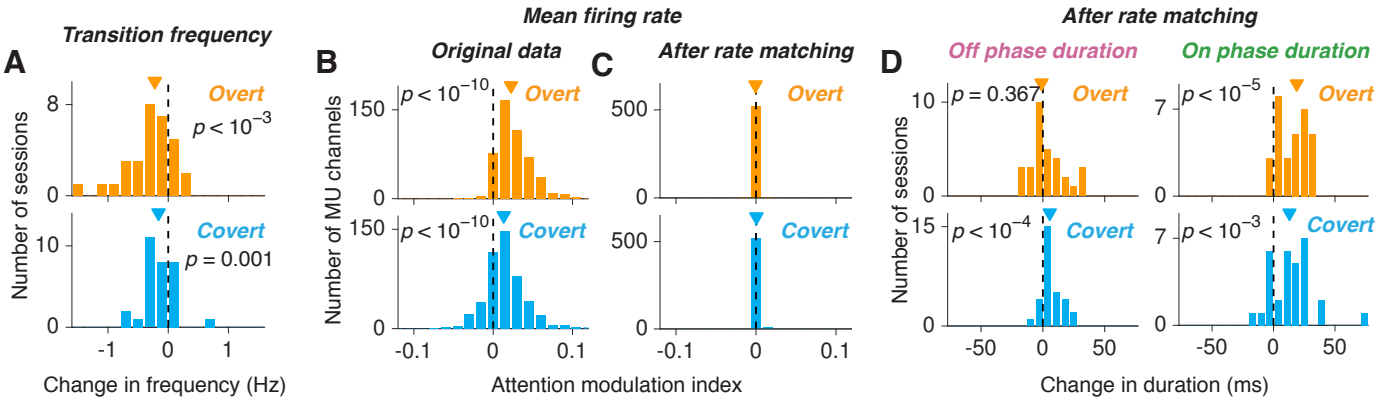

Fig. S9. Control analyses for the observed increase in On-episode durations during attention. (**A**) Distribution of the difference in average On-Off transition frequency between covert attention and control (blue), and overt attention and control (orange) conditions. On-Off transition frequency was significantly lower in attention conditions than in control conditions. (**B-D**) Increase in On-episode durations during attention was not an artifact of the increase in mean firing rate. (**B**) Distribution of the attentional modulation index ($[r_A - r_C]/[r_A + r_C]$) of the mean firing rate of multi-units by covert (blue) and overt attention (orange). Trial-average mean firing rate was higher in attention and saccade preparation conditions than in control condition. (**C**) Same as in B, but after rate matching. Rate-matching procedure equated mean firing rate across all conditions, largely eliminating the firing-rate modulation by covert and overt attention. (**D**) Distribution of the difference ($\tau_A - \tau_C$) in average duration of the On (right panel) and Off (left panel) episodes between covert attention and control (blue), and overt attention and control (orange) conditions after rate matching. The increase in On episode durations is preserved in the rate-matched data, thus this effect was not an artifact of higher mean firing rates during attention conditions. In all panels, triangles indicate medians of the distributions; p-values are for Wilcoxon signed rank test.

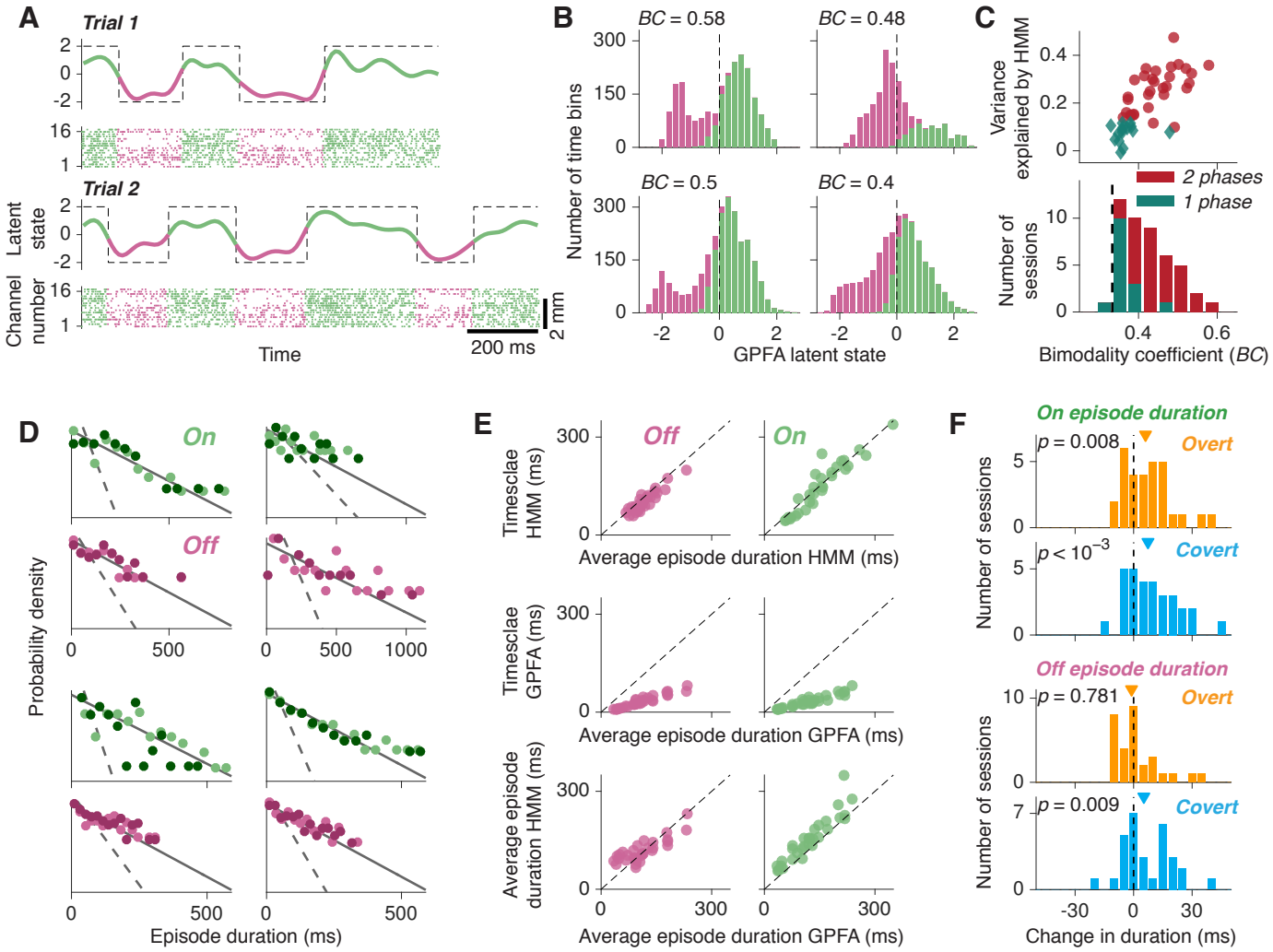

Fig. S10. Gaussian Process Factor Analysis (GPFA) model of On-Off dynamics. (**A**) Two example trials fitted by GPFA and HMM models. Upper panels: the most likely latent trajectory decoded by GPFA (solid line) and by HMM (dashed line). Lower panels: corresponding spike rasters. Coloring of spike rasters and of GPFA trajectories is based on the HMM-decoded phase (green and pink: On and Off phase, respectively). (**B**) Histogram of GPFA latent phase colored according to the corresponding HMM-decoded phase, for four example recordings (BC - bimodality coefficient, upper-left histogram corresponds to example recording in panel A). (**C**) Upper panel: scatter plot of the variance explained by HMM versus average bimodality coefficient of the GPFA latent-phase (each symbol represents a recording session; red circles and teal diamonds: 2-phase and 1-phase recordings, based on the HMM analysis in Fig. S5). Lower panel: histogram of the average bimodality coefficient of the GPFA latent phase across recording sessions (red and teal bars: stacked histograms for 2-phase and 1-phase recordings, respectively, based on the HMM analysis in Fig. S5). Dashed vertical line indicates the value of BC=1/3 for a unimodal Gaussian distribution. (**D**) Empirical distributions of On and Off episode durations for four example recordings shown in panel b (same arrangement). For the GPFA model, the On and Off episodes were determined based on zero-crossings of the latent trajectory. Empirical distributions of episode durations determined by HMM (light green and light pink) and by GPFA (dark green and dark pink) are overlaid by exponential distributions with the decay time-constants

calculated from the fitted HMM's transition probabilities (solid grey line) and from the GPFA's timescale parameter (dashed grey line). The distributions of episode durations decoded by HMM and GPFA largely overlap. GPFA's timescale parameter is systematically smaller than the average episode durations, indicating that transition events are faster than average On and Off episode durations. (**E**) HMM's timescale parameters closely match the average HMM-decoded episode durations (upper panels), whereas GPFA's timescale parameter is systematically smaller than the average GPFA-decoded episode durations (middle panels). Average episode durations decoded by HMM and by GPFA were similar (lower panel). Each dot in the scatter plots represents a recording session. (**F**) Same as Fig. 3D, but for the On and Off episode durations determined based on zero-crossings of the GPFA latent trajectory. On episodes determined based on GPFA were significantly longer in attention conditions than in control conditions, replicating the results of the HMM-based analysis.

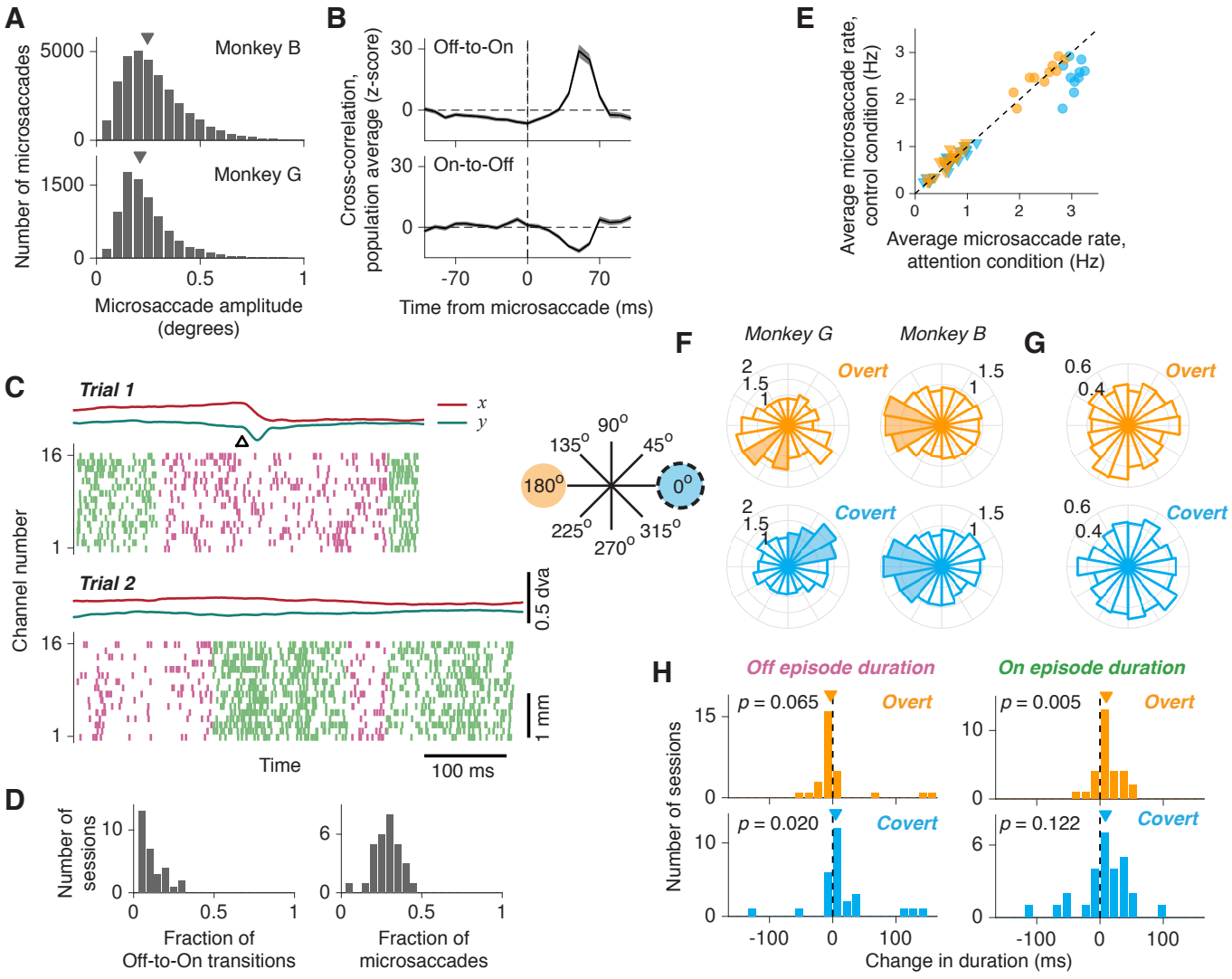

Fig. S11. Increase of On-episode durations during attention is not due to microsaccades. (**A**) Distribution of microsaccade amplitudes for monkey B (upper panel) and monkey G (lower panel) in attention task, during the time period that was used to fit the HMM (400 ms post cue until the end of the attention period). Triangles indicate the median microsaccade amplitude (12.7 and 14.7 arcmin in monkeys G and B, respectively). (**B**) Population-average cross-correlations between the times of microsaccades and times of Off-to-On (upper panel) and On-to-Off transitions (lower panel); shaded area indicates ±sem. The probability of Off-to-On transition was significantly enhanced, and the probability of On-to-Off transition was significantly suppressed ∼60 ms after a microsaccade. (**C**) Example of a microsaccade that occurs during the Off-phase and is not followed by an Off-to-On transition within the next $30-75$ ms (Trial 1, triangle indicates the time of the microsaccade with the amplitude 10.4 arcmin; red and teal lines show simultaneously recorded x and y eye position). Example of On-Off transitions observed during a period of highly stable gaze (Trial 2). Green and pink ticks in the raster plots indicate spikes occurring during the On and Off episodes, respectively, as decoded by the HMM. The data are from the same example session as in Fig. 1. (**D**) Histograms across recordings of the fraction of Off-to-On transitions that were preceded by a microsaccade within a $30-75$ ms

window (left panel) and of the fraction of microsaccades occurring during the Off-phase that were followed by an Off-to-On transition within the next $30 - 75$ ms (right panel). On average, only 12% of Off-to-On transitions were preceded by a microsaccade and only 28% of microsaccades were followed by an Off-to-On transition. (**E**) Scatter plot of the average microsaccade rate in attention (x-axis; covert - blue symbols, overt - orange symbols) versus control (y-axis) conditions. Each dot represents a recording session (monkey G - triangles, monkey B - circles). The average microsaccade rate was not significantly higher during attention compared to control conditions, except for the covert attention condition in monkey B, where the microsaccade rate was significantly higher than in control condition (blue circles, $p = 0.002$, Wilcoxon signed rank test). (**F**) Relative microsaccade frequency between covert attention and control (blue) and overt attention and control (orange) conditions across microsaccade directions, aligned to the RF location (monkey G - left panel, monkey B - right panel). The relative frequency is computed for each direction (in 20° bins) as the ratio of the number of microsaccades towards that direction in attention over control condition. Legend: 0° corresponds to the RF location (dashed circle), which aligns with the covertly attended stimulus in the covert condition (blue circle), and is opposite to covertly attended stimulus in the overt condition (orange circle). There was an overall tendency for microsaccades to be more frequently directed towards the approximate location of the covertly attended stimulus, but microsaccades were more frequently directed towards the location opposite the covertly attended stimulus in the covert condition for monkey B. Directions with significantly higher relative microsaccade frequency are highlighted with color-fill (chi-squared residuals test at 0.05 significance level with Bonferroni correction). (**G**) Fraction of microsaccades that were followed by Off-to-On transition across microsaccade directions, aligned to the RF location, during covert (blue) and overt (orange) conditions. The fraction of microsaccades followed by Off-to-On transitions did not significantly vary across microsaccade directions (chi-squared test, covert: $p = 0.367$; overt: $p = 0.311$). (**H**) Analysis of trials without microsaccades: distribution of the difference ($\tau_{\mathrm{A}} - \tau_{\mathrm{C}}$) in average duration of the On (right panel) and Off (left panel) episodes between covert attention and control (blue), and overt attention and control (orange) conditions. On episodes were significantly longer in attention conditions than in control conditions on trials without microsaccades (cf. Fig. 3D, but note the difference in the range of x-axis).

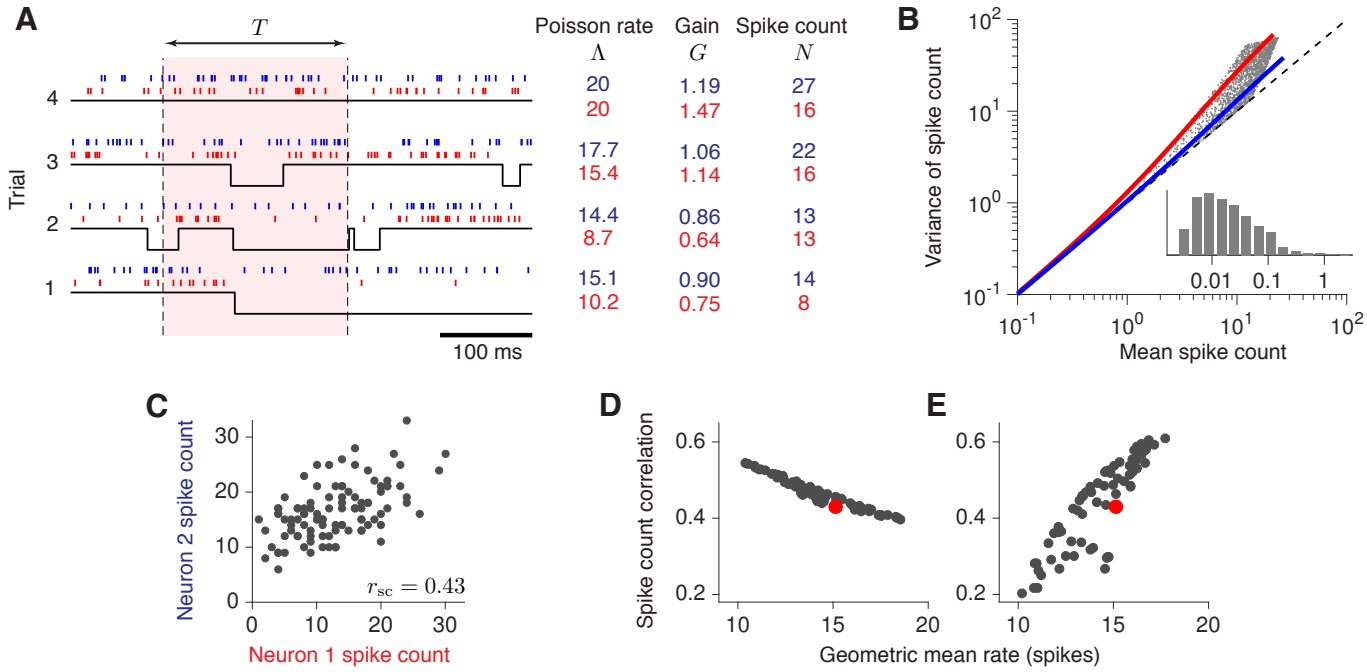

Fig. S12. On-Off dynamics explain known features of correlated spike-count variability. (**A**) Illustration of how correlated spike-count variability and gain fluctuations arise in the model of On-Off dynamics. Spike times of two example neurons (red and blue vertical ticks) are generated as inhomogeneous Poisson processes with different mean rates during the On and Off phases (black lines, four example trials are shown). The On-Off dynamics are the same for both neurons, but their firing rates during the On and Off phases are different ($r_{\text{off},1} = 20$ Hz, $r_{\text{off},2} = 60$ Hz, $r_{\text{on},1} = r_{\text{on},2} = 100$ Hz). Spike-counts N are measured across trials in a time-window T (200 ms, red shading). For each neuron, the integral of its instantaneous firing rate over the time-window T determines the rate Λ of the Poisson process on each trial. Since the proportion of On and Off phases within the time-window T varies from trial-to-trail, Λ fluctuates, which can be described as fluctuations in the gain $G = \Lambda/\mathrm{E}[\Lambda]$. The table shows the values of Λ, G and N for the two example neurons on each trial. Parameters of the On-Off dynamics are set to their experimentally measured averages: $\tau_{\text{off}} = 100$ ms, $\tau_{\text{on}} = 150$ ms. (**B**) Variance-to-mean relation of spike-counts predicted by the On-Off model exhibits signatures of multiplicative noise, which have been observed experimentally and modeled as fluctuations of a multiplicative gain. Red and blue lines show the analytical prediction from the On-Off model for the two example neurons from panel A, where the variance and mean are calculated over variable time-windows T ranging from 1 to 316 ms. Black dashed line shows prediction for a Poisson model. The grey dots show the analytical prediction of the variance-to-mean relation calculated using $T = 200$ ms window for a hypothetical population of 2,000 neurons, where r_{off} is uniformly distributed between 0 and 60 Hz and $(r_{\text{on}} - r_{\text{off}})$ is uniformly distributed between 0 and 100 Hz across the population. The inset shows histogram of the gain fluctuations for this hypothetical population, summarized as the coefficient of variation of the gain. (**C**) Spike-count correlations (noise-correlations) arise in the On-Off model, because pairs of neurons follow the same sequence of On and Off episodes on each trail. As a result, their rates Λ and spike-counts N are correlated. Scatter plot shows spike-counts for two example neurons from panel A on 100 trials. Spike-count correlation predicted analytically from the On-Off model is $r_{\text{sc}} = 0.43$. (**D**) The On-Off model predicts, that spike-

count correlation between two neurons decreases as their geometric mean rate increases, when the increase in mean rate is mainly driven by an increase in r_{off} and r_{on} without change in Δr. Spike-count correlation is shown for the two example neurons from panel A across 72 conditions, where mean-rate changes arise from variation in r_{off} of the two neurons, which is uniformly distributed between 15 and 60 Hz across conditions, while all other parameters are held constant as in panel A. Red dot corresponds to spike-count correlation from panel C. (**E**) The On-Off model predicts, that spike-count correlation between two neurons increases as their geometric mean rate increases, when the increase in mean rate is mainly driven by an increase in Δr. Spike-count correlation is shown for two example neurons from panel A across 72 conditions, where mean-rate changes arise from variation in Δr of two neurons, which is uniformly distributed between 20 and 90 Hz across conditions, while all other parameters are held constant as in panel A. Red dot corresponds to spike-count correlation from panel C.

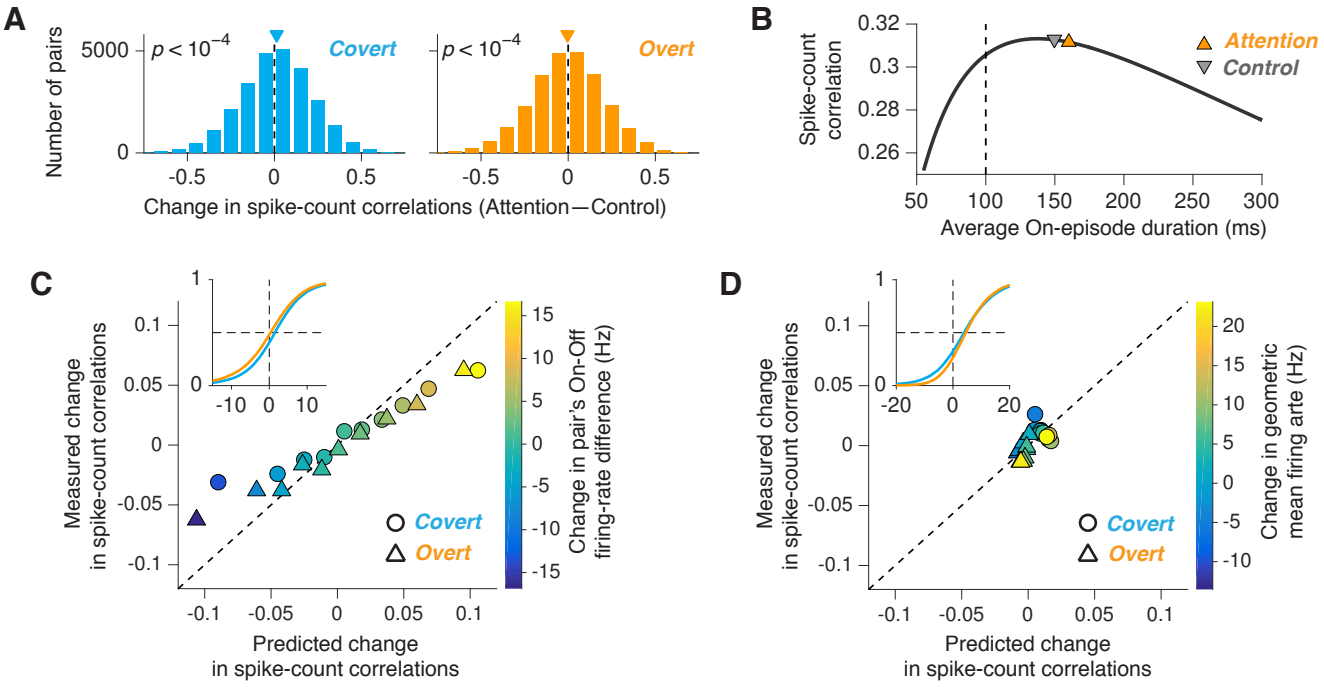

Fig. S13. On-Off dynamics explain changes in spike-count correlations during attention. (**A**) Distribution across all MU-pairs of the difference in spike-count correlations between covert attention and control (blue), and overt attention and control (orange) conditions. Triangles indicate means of the distributions; p-values are for permutation test (10^4 shuffles). (**B**) Theoretical prediction for spike-count correlation as a function of the average On-episode duration, when all other model parameters are held constant (solid line). Off-episode duration is set to its average across sessions $\tau_{\text{off}} = 100$ ms (dashed vertical line), $r_{\text{off}} = 70$ Hz, and $r_{\text{on}} = 117$ Hz. Orange and grey triangles indicate spike-count correlation predicted for attention condition ($\tau_{\text{on}} = 160$ ms) and control condition ($\tau_{\text{on}} = 150$ ms), respectively. (**C**) Comparison between change in spike-count correlations predicted by the On-Off model (x-axis) and measured from the data (y-axis) during covert (circles) and overt (triangles) attention. All pairs are divided in ten equally-sized groups based on the change in their On-Off firing-rate difference between attention and control conditions (color axis). Inset shows cumulative distribution of this change in pairs' On-Off firing-rate difference for covert (blue) and overt (orange) attention. (**D**) Same as in panel C, but pairs are divided in ten equally-sized groups based on the difference in their geometric-mean firing rate between attention and control conditions (color axis). Inset shows cumulative distribution of this difference in geometric-mean firing rates for covert (blue) and overt (orange) attention.

Monkey condition	**G** covert	**G** overt	**B** covert	**B** overt
Δ On-duration attention−control	16 ms $p < 10^{-3}$	14 ms $p < 10^{-3}$	-0.5 ms $p = 0.695$	8 ms $p = 0.002$
Δ microsaccade rate attention−control	0.01 Hz $p = 0.97$	-0.06 Hz $p = 0.03$	0.6 Hz $p = 0.002$	-0.005 Hz $p = 0.275$

Supplementary Table 1. Average change of On-episode durations and of microsaccade rate in attention relative to control conditions, separately for covert and overt attention in each monkey. The microsaccade rate was significantly higher for covert attention relative to control condition in monkey B, however, no increase of On-episode durations was observed in this condition. On the other hand, significant increase of On-episode durations, but no difference in microsaccade rate was observed in all remaining cases. P-values are for the Wilcoxon signed rank test.

References and Notes

1. K. D. Harris, A. Thiele, Cortical state and attention. *Nat. Rev. Neurosci.* **12**, 509–523 (2011). doi:10.1038/nrn3084 Medline

2. S.-H. Lee, Y. Dan, Neuromodulation of brain states. *Neuron* **76**, 209–222 (2012). doi:10.1016/j.neuron.2012.09.012 Medline

3. M. Steriade, D. A. McCormick, T. J. Sejnowski, Thalamocortical oscillations in the sleeping and aroused brain. *Science* **262**, 679–685 (1993). doi:10.1126/science.8235588 Medline

4. M. Steriade, I. Timofeev, F. Grenier, Natural waking and sleep states: A view from inside neocortical neurons. *J. Neurophysiol.* **85**, 1969–1985 (2001). Medline

5. B. Haider, A. Duque, A. R. Hasenstaub, Y. Yu, D. A. McCormick, Enhancement of visual responsiveness by spontaneous local network activity in vivo. *J. Neurophysiol.* **97**, 4186–4202 (2007). doi:10.1152/jn.01114.2006 Medline

6. A. Hasenstaub, R. N. S. Sachdev, D. A. McCormick, State changes rapidly modulate cortical neuronal responsiveness. *J. Neurosci.* **27**, 9607–9622 (2007).

7. A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, K. D. Harris, The asynchronous state in cortical circuits. *Science* **327**, 587–590 (2010). doi:10.1126/science.1179850 Medline

8. S. Crochet, C. C. H. Petersen, Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nat. Neurosci.* **9**, 608–610 (2006). doi:10.1038/nn1690 Medline

9. J. F. A. Poulet, C. C. H. Petersen, Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature* **454**, 881–885 (2008). doi:10.1038/nature07150 Medline

10. C. M. Niell, M. P. Stryker, Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**, 472–479 (2010). doi:10.1016/j.neuron.2010.01.033 Medline

11. M. Okun, A. Naim, I. Lampl, The subthreshold relation between cortical local field potential and neuronal firing unveiled by intracellular recordings in awake rats. *J. Neurosci.* **30**, 4440–4448 (2010). doi:10.1523/JNEUROSCI.5062-09.2010 Medline

12. A. Luczak, P. Bartho, K. D. Harris, Gating of sensory input by spontaneous cortical activity. *J. Neurosci.* **33**, 1684–1695 (2013). doi:10.1523/JNEUROSCI.2928-12.2013 Medline

13. A. Y. Y. Tan, Y. Chen, B. Scholl, E. Seidemann, N. J. Priebe, Sensory stimulation shifts visual cortex from synchronous to asynchronous states. *Nature* **509**, 226–229 (2014). doi:10.1038/nature13159 Medline

14. N. A. Steinmetz, T. Moore, Eye movement preparation modulates neuronal responses in area V4 when dissociated from attentional demands. *Neuron* **83**, 496–506 (2014). doi:10.1016/j.neuron.2014.06.014 Medline

15. E. Seidemann, I. Meilijson, M. Abeles, H. Bergman, E. Vaadia, Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J. Neurosci.* **16**, 752–768 (1996). Medline

16. G. Rainer, E. K. Miller, Neural ensemble states in prefrontal cortex identified using a hidden Markov model with a modified EM algorithm. *Neurocomputing* **32-33**, 961–966 (2000). doi:10.1016/S0925-2312(00)00266-6

17. M. Vinck, R. Batista-Brito, U. Knoblich, J. A. Cardin, Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* **86**, 740–754 (2015). doi:10.1016/j.neuron.2015.03.028 Medline

18. J. Reimer, E. Froudarakis, C. R. Cadwell, D. Yatsenko, G. H. Denfield, A. S. Tolias, Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron* **84**, 355–362 (2014). doi:10.1016/j.neuron.2014.09.033 Medline

19. J. Moran, R. Desimone, Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985). doi:10.1126/science.4023713 Medline

20. S. Treue, J. H. Maunsell, Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996). doi:10.1038/382539a0 Medline

21. P. Fries, J. H. Reynolds, A. E. Rorie, R. Desimone, Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**, 1560–1563 (2001). doi:10.1126/science.1055465 Medline

22. A. Renart, C. K. Machens, Variability in neural activity and behavior. *Curr. Opin. Neurobiol.* **25**, 211–220 (2014). doi:10.1016/j.conb.2014.02.013 Medline

23. M. R. Cohen, A. Kohn, Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–819 (2011). doi:10.1038/nn.2842 Medline

24. M. R. Cohen, W. T. Newsome, Context-dependent changes in functional circuitry in visual area MT. *Neuron* **60**, 162–173 (2008). doi:10.1016/j.neuron.2008.08.007 Medline

25. Y. Gu, S. Liu, C. R. Fetsch, Y. Yang, S. Fok, A. Sunkara, G. C. DeAngelis, D. E. Angelaki, Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750–761 (2011). doi:10.1016/j.neuron.2011.06.015 Medline

26. A. S. Ecker, P. Berens, R. J. Cotton, M. Subramaniyan, G. H. Denfield, C. R. Cadwell, S. M. Smirnakis, M. Bethge, A. S. Tolias, State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248 (2014). doi:10.1016/j.neuron.2014.02.006 Medline

27. J. F. Mitchell, K. A. Sundberg, J. H. Reynolds, Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* **63**, 879–888 (2009). doi:10.1016/j.neuron.2009.09.013 Medline

28. M. R. Cohen, J. H. R. Maunsell, Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594–1600 (2009). doi:10.1038/nn.2439 Medline

29. D. A. Ruff, M. R. Cohen, Attention can either increase or decrease spike count correlations in visual cortex. *Nat. Neurosci.* **17**, 1591–1597 (2014). doi:10.1038/nn.3835 Medline

30. D. A. Ruff, M. R. Cohen, Attention increases spike count correlations between visual cortical areas. *J. Neurosci.* **36**, 7523–7534 (2016). doi:10.1523/JNEUROSCI.0610-16.2016 Medline

31. R. L. T. Goris, J. A. Movshon, E. P. Simoncelli, Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014). doi:10.1038/nn.3711 Medline

32. N. C. Rabinowitz, R. L. Goris, M. Cohen, E. P. Simoncelli, Attention stabilizes the shared gain of V4 populations. *eLife* **4**, e08998 (2015). doi:10.7554/eLife.08998 Medline

33. M. R. Cohen, J. H. R. Maunsell, A neuronal population measure of attention predicts behavioral performance on individual trials. *J. Neurosci.* **30**, 15241–15253 (2010). doi:10.1523/JNEUROSCI.2171-10.2010 Medline

34. G. Aston-Jones, J. D. Cohen, An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005). doi:10.1146/annurev.neuro.28.061604.135709 Medline

35. T. W. Robbins, A. F. T. Arnsten, The neuropsychopharmacology of fronto-executive function: Monoaminergic modulation. *Annu. Rev. Neurosci.* **32**, 267–287 (2009). doi:10.1146/annurev.neuro.051508.135535 Medline

36. J. L. Herrero, M. J. Roberts, L. S. Delicato, M. A. Gieselmann, P. Dayan, A. Thiele, Acetylcholine contributes through muscarinic receptors to attentional modulation in V1. *Nature* **454**, 1110–1114 (2008). doi:10.1038/nature07141 Medline

37. B. Noudoost, T. Moore, Control of visual cortical signals by prefrontal dopamine. *Nature* **474**, 372–375 (2011). doi:10.1038/nature09995 Medline

38. E. Zagha, A. E. Casale, R. N. S. Sachdev, M. J. McGinley, D. A. McCormick, Motor cortex feedback influences sensory processing by modulating network state. *Neuron* **79**, 567–578 (2013). doi:10.1016/j.neuron.2013.06.008 Medline

39. G. G. Gregoriou, S. J. Gotts, H. Zhou, R. Desimone, High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science* **324**, 1207–1210 (2009). doi:10.1126/science.1171402 Medline

40. K. M. Armstrong, M. H. Chang, T. Moore, Selection and maintenance of spatial information by frontal eye field neurons. *J. Neurosci.* **29**, 15621–15629 (2009). doi:10.1523/JNEUROSCI.4465-09.2009 Medline

41. K. M. Armstrong, J. K. Fitzgerald, T. Moore, Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron* **50**, 791–798 (2006). doi:10.1016/j.neuron.2006.05.010 Medline

42. R. M. Kalwani, L. Bloy, M. A. Elliott, J. I. Gold, A method for localizing microelectrode trajectories in the macaque brain using MRI. *J. Neurosci. Methods* **176**, 104–111 (2009). doi:10.1016/j.jneumeth.2008.08.034 Medline

43. D. N. Hill, S. B. Mehta, D. Kleinfeld, Quality metrics to accompany spike sorting of extracellular signals. *J. Neurosci.* **31**, 8699–8705 (2011). doi:10.1523/JNEUROSCI.0971-11.2011 Medline

44. A. Amarasingham, M. T. Harrison, N. G. Hatsopoulos, S. Geman, Conditional modeling and the jitter method of spike resampling. *J. Neurophysiol.* **107**, 517–531 (2012). doi:10.1152/jn.00633.2011 Medline

45. M. A. Smith, A. Kohn, Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.* **28**, 12591–12603 (2008). doi:10.1523/JNEUROSCI.2929-08.2008 Medline

46. N. A. Steinmetz, T. Moore, Changes in the response rate and response variability of area V4 neurons during the preparation of saccadic eye movements. *J. Neurophysiol.* **103**, 1171–1178 (2010). doi:10.1152/jn.00689.2009 Medline

47. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).

48. C. Kemere, G. Santhanam, B. M. Yu, A. Afshar, S. I. Ryu, T. H. Meng, K. V. Shenoy, Detecting neural-state transitions using hidden Markov models for motor cortical prostheses. *J. Neurophysiol.* **100**, 2441–2452 (2008). doi:10.1152/jn.00924.2007 Medline

49. T. Womelsdorf, P. Fries, P. P. Mitra, R. Desimone, Gamma-band synchronization in visual cortex predicts speed of change detection. *Nature* **439**, 733–736 (2006). doi:10.1038/nature04258 Medline

50. C. Nicholson, J. A. Freeman, Theory of current source-density analysis and determination of conductivity tensor for anuran cerebellum. *J. Neurophysiol.* **38**, 356–368 (1975). Medline

51. G. Vaknin, P. G. DiScenna, T. J. Teyler, A method for calculating current source density (CSD) analysis without resorting to recording sites outside the sampling volume. *J. Neurosci. Methods* **24**, 131–135 (1988). Medline

52. M. R. Jarvis, P. P. Mitra, Sampling properties of the spectrum and coherency of sequences of action potentials. *Neural Comput.* **13**, 717–749 (2001). doi:10.1162/089976601300014312 Medline

53. M. W. Reimann, C. A. Anastassiou, R. Perin, S. L. Hill, H. Markram, C. Koch, A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents. *Neuron* **79**, 375–390 (2013). doi:10.1016/j.neuron.2013.05.023 Medline

54. M. Okun, N. A. Steinmetz, L. Cossell, M. F. Iacaruso, H. Ko, P. Barthó, T. Moore, S. B. Hofer, T. D. Mrsic-Flogel, M. Carandini, K. D. Harris, Diverse coupling of neurons to populations in sensory cortex. *Nature* **521**, 511–515 (2015). doi:10.1038/nature14273 Medline

55. B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, M. Sahani, Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* **102**, 614–635 (2009). doi:10.1152/jn.90941.2008 Medline

56. D. A. Leopold, N. K. Logothetis, Microsaccades differentially modulate neural activity in the striate and extrastriate visual cortex. *Exp. Brain Res.* **123**, 341–345 (1998). doi:10.1007/s002210050577 Medline

57. W. Bair, L. P. O'Keefe, The influence of fixational eye movements on the response of neurons in area MT of the macaque. *Vis. Neurosci.* **15**, 779–786 (1998). doi:10.1017/S0952523898154160 Medline

58. M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G.

Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, K. V. Shenoy, Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010). doi:10.1038/nn.2501 Medline