Supporting Information

Brain state limits propagation of neural signals in laminar cortical circuits
Natasha Kharas, Ariana Andrei, Samantha R. Debes, and Valentin Dragoi*

*Corresponding author. Email: valentin.dragoi@uth.tmc.edu

This file includes:

Methods

Figures S1 to S16
DETAILED METHODS

All experiments were performed in accordance with protocols approved by the US National Institutes of Health Guidelines for the Care and Use of Animals for Experimental Procedures and were approved by the Institutional Animal Care and Use Committee at the University of Texas Health Science Center at Houston.

Surgical procedures
A titanium head post was implanted in the medial frontal region with the help of multiple anchor screws. Following a recovery period of about 10 days, monkeys were trained for 3–4 months on visual fixation and contrast detection tasks. After a monkey learned the tasks, a recording chamber (inner diameter of 17 mm) for single-unit multiple electrode recording was cemented over areas V1 and V4 (according to the magnetic resonance imaging map). A few stainless steel screws were inserted into the skull around the recording chambers and a thin stainless steel wire was wrapped around the screws for additional support.

Electrophysiological recordings
A total of 75 sessions and 1407 units in V1 were recorded from three rhesus monkeys (Macaca mulatta), W, C, and T. An additional 54 sessions and 856 units were recorded in V4 (see Supplementary Figs. S3-S8). Overall, we recorded the activity of 1920 cells across the layers of V1 in 3 monkeys (n = 95 sessions across all behavioral states). We recorded 789 V1 units in the task condition, 608 V1 units in the fixation (no stimulus) condition, and 523 V1 units in the rest condition. A total of 3840 trials involving stimulus presentation were recorded across 32 sessions in the fixation condition (n = 2 monkeys); a total of 11520 trials were recorded across 32 sessions in the task condition (n = 2 monkeys); a total of 2500 pseudo-trials were recorded in the rest condition across 31 sessions in V1 (n = 3 monkeys). For brain area V4, we recorded the activity of 1712 cells across the layers of V4 in 2 monkeys (n = 108 sessions across all behavioral states). We recorded 856 V4 units in the task condition, 600 V4 units in the fixation (no stimulus) condition, and 256 V4 units in the rest condition. A total of 3507 trials involving stimulus presentation were recorded across 38 sessions in the fixation (no stimulus) condition (n = 2 monkeys); a total of 1620 trials were recorded across 54 sessions in the task condition (n = 2 monkeys); a total of 3630 pseudo-trials were recorded in the rest condition across 16 sessions in V4 (n = 2 monkeys)
Neuronal activity was recorded using a laminar probe with 16–24 equally spaced contacts at 100 µm (U-Probe, Plexon). The probe was advanced into the brain using the NAN drive system (Plexon), which is attached to the recording chamber. All penetrations were perpendicular to the surface of the cortex and initiated at the same target depth relative to surface. All x, y, and z coordinates were measured relative to a stable grid in the recording chamber. The needle with viral concoction was advanced by a precision, computer controlled micro-manipulator (NAN instruments) to a pre-established depth based on the current source density to cover all layers of the cortex. We mapped the cortical layers prior to the injections for each injected column.

Local field potentials (LFP) were recorded in all channels, multi-unit activity was recorded in most channels, and single-unit activity was recorded on average in 50% of the channels. We recorded both spiking activity and local field potential (LFP) signals by means of a Multichannel Acquisition Processor System (MAP, Plexon). Real-time neuronal signals were processed with the MAP system at a sampling rate of 40 kHz. The signals were first filtered by a preamplifier box into spike channels (150 Hz–8 kHz, one pole low-cut, three pole high-cut, with programmable referencing, 50× gain) and field potential channels (0.07, 0.7, 3–170, 300, 500 Hz user selectable, one pole low-cut, one pole high-cut, 50×). Single-unit signals were further amplified, filtered, and viewed on an oscilloscope, and heard through a speaker. The spike waveforms above threshold were saved and fine sorted after data acquisition was terminated using Plexon’s offline sorter program. After a unit was isolated, its receptive field was mapped with dynamic gratings or using reverse correlation while the animal maintained fixation. Waveforms that crossed a pre-defined threshold (~4 s.d. above the amplitude of the noise signal) were stored for offline analyses. Spike waveforms were manually processed with Plexon’s offline sorter program using waveform clustering parameters such as principle component analysis, along with spike amplitude, timing, width, valley, and peak. Units were analyzed in principal component spaces chosen based on the spike waveform shape and low signal-to-noise ratio. High- and low-amplitude noise was manually removed using an offline sorter. Single units were subsequently analyzed using custom scripts in MATLAB.

**Local Field Potential (LFP) analysis**

LFP activity was recorded at 1 KHz. The raw signal (measured in Volts) was processed through the bandpower function in MATLAB, which outputs the amplitude of the signal in the frequency
range (for example, 20-80 Hz). We corrected the filter-induced timing delays of LFP data by using the FPAlign utility from Plexon. We then filtered the LFP channels with a band pass filter (0.5-80 Hz) and we applied an 8th order Butterworth notch filter at 60 Hz. LFP power was computed for each recording channel independently, and then was z-scored across trials. LFP power ratio (PR) was computed from the LFP power in the low-frequency bands (0.5-10 Hz) and LFP power in the high-frequency bands (30-80 Hz). To compare the LFP bandpower between behavioral states, we normalized the resultant power (arbitrary units). We relabeled the color scale as 0 to 1 to reflect the normalized power (A.u.) in various frequency bands. Power ratio, PR, was computed as follows: PR = (P10-P80)/P10, where P10 is the spectral power in the 0.5-10 Hz range, and P80 is the spectral power in the 30-80 Hz. The purpose of the local field potential analysis was to confirm the behavioral state of the animal in each session, hence this measure was used as a metric of brain state (1). Only the rest sessions in which PR was significantly different from task sessions using a Wilcoxon rank sum test (P < 0.05) were considered valid. Overall, we found that, as expected, rest was associated with a significant increase in low-frequency power (1-10 Hz) and a significant decrease in high-frequency power (30-80 Hz).

Quantification of rest period
To ensure that animals were resting during the designated rest sessions, we coordinated the start of the rest condition to occur at the time of day when monkeys naturally take daytime naps (2), with sessions starting at approximately 2 pm each day. A rest session was considered valid if (i) the monkey’s eyes were closed for > 85% of total session duration and (ii) PR was significantly higher (Wilcoxon rank sum test, P < 0.05) in rest compared to wakefulness (fixation and task). Eye position was monitored using a camera-based system sampling at 1 KHz (EyeLink II, SR Research). Eye closure was derived from eye position using custom MATLAB scripts. A total of 31 rest sessions across three monkeys, W, T and C met our criteria, in which we recorded 523 units.

Current Source Density (CSD) analysis to identify cortical layers
For each recording session, we identified layers in the visual cortices V1 and V4 using current source density (CSD) analysis (3). We recorded LFP data using laminar probes (Plexon U-probes with equally-spaced 100 μm intercontacts) while presenting a full field high contrast natural image. We averaged the filtered LFP signals across trials to obtain the evoked response potential (ERP)
for each channel. We computed the current source density by using the second spatial derivative of the LFP time-series across equally spaced laminar contacts using the iCSD toolbox for MATLAB. The granular layer was identified by finding the maximum sink, measured by nA/mm³. Channels located in the primary sink were assigned to the granular layer, while channels above the sink were assigned to the supragranular layer and channels below the sink were assigned to the infragranular layer (3). Several controls were performed to ensure that the above protocol for laminar characterization is reliable: (i) small movements of the entire laminar probe shift the CSD plot towards the performed direction of movement, (ii) shuffling contact position on the laminar probe destroys the deliverable CSD plot, and (iii) total laminar length was verified with anatomical measurements.

Eye movement control

Eye position was calibrated at the beginning of each session using a five-point calibration procedure. On each trial, monkeys were trained to fixate on a dot (0.2 deg in size) at the center of the screen within a small rectangular 1-2 deg window. To ensure fixation, eye position was constantly monitored by using an eye tracker operating at 1 KHz (EyeLink II; SR Research). The eye-tracker gains were adjusted so that they were linear for horizontal and vertical eye deflections. Microsaccades were analyzed every 10 ms by using a vector velocity threshold of 10 deg/s (this corresponds to a 0.1 deg eye movement between consecutive 10-ms intervals). If a detected microsaccade exceeded 0.25 deg (fixation instability), the trial was automatically aborted. Thus, if at any point during the trial, eye position exceeded 0.25 deg outside the boundaries of the rectangular box, the trials described in the fixation and detection tasks below were automatically aborted. For analysis in rest, eye closure was calculated based on eye X and Y traces using custom MATLAB scripts. Only sessions in which eye closure exceeded 85% of the total session time were considered valid rest sessions. Pupil size was continuously measured monocularly at 1 kHz sampling frequency (EyeLink II; SR Research). We included in the analysis only the time periods in which luminance was constant and the fixation point was the only stimulus displayed on the screen (during the delay period, see Figures 1A and 1B). For the pupil analysis (see Figures 1H and 2I), pupil size was downsampled at 100 Hz and the median value was taken over the entire delay period in each trial. The pupil size trace was z-scored to compare across sessions. The average pupil size was calculated per trial and correlated with state transitions/s using the Pearson correlation coefficient R. P values < 0.05 were considered significant.
Dynamics of laminar population activity in different brain states

Neuronal activity was recorded during wakefulness and rest conditions in three rhesus monkeys, W, C, and T. During wakefulness, monkeys were trained to perform fixation (no stimulus) and contrast detection. During rest, monkeys were in a dark room for 20-45 minutes. During wakefulness and rest sessions, we recorded single and multi-unit activity from area V1 (57 sessions; similar experiments were performed in area V4 in 54 sessions, e.g., Figs. S1, S3B, S4, S5, S6B, S7 and S8B). Importantly, we measured the dynamics of laminar population activity for a period of 1-s across trials during wakefulness and rest. The main objective was to directly compare the dynamics of neuronal responses during rest and wakefulness in the absence of visual stimulation (both during passive fixation and during a demanding task, as described below).

Passive fixation. Monkeys fixated on a central point (0.1 deg in size) presented on a 19” CRT video monitor (Dell, 60 Hz refresh rate) on a dark background in a dark room. Fixation was maintained within a 1-deg window for 1000 ms in the absence of sensory stimulation. Successful fixation was rewarded with 5 drops of juice. If fixation was broken, trials are aborted. Each session consisted of 120-480 trials. We recorded 3,840 trials across 32 sessions (608 units).

Task. Monkeys performed a contrast detection task using gray-scale sinusoidal gratings of luminance-varying contrasts generated using MATLAB with Psychophysics Toolbox 75 and presented binocularly on a 19” CRT video monitor (Dell, 60 Hz refresh rate) on a dark background in a dark room. While monkeys fixated, stimuli with a diameter of 2-3 deg were displayed at 2-4 deg eccentricity. The location and size of the stimuli covered the multiple receptive fields of the cells recorded (Figure S1). Stimulus location and size were optimized in each session such as to stimulate the largest number of simultaneously recorded cells. Stimulus presentation was recorded and synchronized with the neural data using the Experiment Control Module programmable device (ECM, FHC Inc.). Stimuli had a fixed spatial frequency (1.7 cycles per degree) and were displayed for 300 ms (starting 450-1000 ms after fixation onset), and were followed by a 1.1-s delay (‘blank’ screen). Luminance contrast was defined as the change in luminance (peak to trough), divided by the mean luminance for each sinusoidal grating. Peak luminance values for each stimulus were 0.107, 0.120, 0.133, 0.280 cd/m2 and 0.08 cd/m2 in the no stimulus condition (Tektronix, J17). In each experiment, stimuli could have one of four different luminance contrasts and were present on 50% of the trials. At the end of the delay interval, monkeys were required to signal the presence
of the stimulus by releasing the lever or maintaining contact if no stimulus was displayed. Correct behavioral responses were rewarded with 5 drops of juice. Importantly, we measured the dynamics of laminar population activity for a period of 1000 ms in which the screen was ‘blank’ that started 100 ms after stimulus was turned off (no-stimulus trials were randomly interspersed with stimulus trials). Each session consisted of 360-720 total trials. A total of 11,520 trials across 32 sessions were recorded.

Rest. In the rest condition, either performed before or after the passive fixation or contrast detection tasks, monkeys remained in the same experimental set-up in a dark room and stimulus screen with a dark background for 20–45 minutes. Monkeys began to rest at approximately 2 p.m., which is around the time monkeys naturally take daytime naps (2). A total of 31 rest sessions with 523 units were recorded across three monkeys, W, T and C (see ‘Quantification of rest sessions’ above).

In conditions where we performed laser stimulation (passive fixation/wakefulness and rest), experimental conditions were well controlled. That is, animals were in the same experimental setup in a dark room, and in both conditions, the monitor had a dark background and there was no visual stimulus in either condition. Furthermore, neurons’ firing rates were not significantly different between passive fixation and task conditions (Figure S12). Thus, experimental conditions during wakefulness and rest were identical: dark room, dark monitor, and no visual stimulation. These identical experimental conditions allowed us to evaluate the effect of laser stimulation during wakefulness (passive fixation) and rest.

Hidden Markov Model analysis
To avoid stimulus-related confounds we examined neuronal responses for 1-s in the absence of external stimulation during visual fixation and during the contrast detection task (in the ‘blank’ trials; Figure 1A-B), and during rest when monkeys had their eyes closed (Figure 1C). We fit a hidden Markov model (HMM) to our spiking data as described previously in detail (4). Briefly, multi-unit and single-unit summed spiking data were binned every 10 ms, and the HMM model, which assumes that spikes were generated via a Poisson process, was fitted to spiking data to identify two states (On and Off state) and the transition probabilities between these states using the hmmtrain function in MATLAB R2017b. Two states were selected by analyzing the average
cross-validation error for n states (where n=1-4). The crucial point in error reduction (as denoted by the arrow) occurred at two states for the dataset (Figure S13). Although increasing the number of states reduces the cross-validation error, it is also prone to over clustering the data. Thus, the arrow denotes the optimal number of states with the lowest error without overfitting the model indicating that two states was the most parsimonious model for the dataset (Figure S13; combined rest, task, and fixation data). The model was trained on 50% of the data set and tested on 50% of the data set. The model output was analyzed using the hmmviterbi function in MATLAB R2017b, which utilizes the Viterbi algorithm to classify the spiking data into the two states for all time points. Because the HMM is an unsupervised learning model, only the number of states and the convergence threshold were set. The higher firing rate state was designated as On state, and the lower firing rate state was designated as Off state, with minimal overlap between On and Off states indicating appropriate clustering by the model (Figure S3).

The proportion of time spent in the On or Off states in the wakefulness condition is relatively equal (Figure S15, left panel). The relative amount of time spent in the On (higher firing rate) state is largest in the task condition (5, 6). This is expected as visually responsive neurons increase firing rates in response to the stimuli presented in the tasks. Whereas, the rest condition is dominated by the Off (low firing rate) state (Figure S15, left panel), which is expected as the “DOWN” states in rest are a distinct feature of this brain state (7, 8). Further, the proportion of time spent in the On and Off states does not vary based on the arousal of the animal (Figure S15, right panel). The proportion of time spent in these states does not indicate the probability of switching between states, which is reflected in the transition rate as described below.

As a control, the HMM model was trained on shuffled data (by shuffling the timing of spikes for each unit). The final output was calculated as the output of the model trained on the original data subtracted from the output of the model trained on shuffled data. The final output was analyzed to determine the average firing rates in the On and Off states and the duration of the On and Off states. Lastly, for the time period of 1 second, the switch between On and Off states was calculated and labeled as ‘transition/s’. Transitions measured during laser duration of 300 ms were labeled as ‘transitions’. This measure was used in further analysis, as it is a derivative of not only the presence of On/Off states but also the fluctuation between them. The fluctuation between states was examined during passive fixation, during task, and during rest. All transition rate differences
between behavioral states (fixation, task and rest) were significant within each monkey (P < 0.05, Wilcoxon rank sum). Furthermore, there were no statistically significant differences between monkeys (P > 0.05, Chi-squared test).

The Hidden Markov Model (HMM) was fitted to the population activity of single-unit and multi-unit activity recorded across a cortical column, and the model was fitted to each session. To test for robustness, we repeated the analysis by fitting the model across all sessions and the results did not significantly differ. Thus, whether we ran the HMM per session or across sessions, our results did not change. For example, the mean transition rate in the fixation condition was 0.3, and 0.2 respectively, when the model was trained across vs. per session (P > 0.05, Wilcoxon rank sum test). When we attempted to train the model across all behavioral states, the model did not converge due to the high variability across conditions. Thus, we trained the unsupervised HMM on each behavioral state separately and presented the data accordingly.

**Optogenetics experiments**

*Viral vector injections.* ChR2 was expressed specifically in V1 excitatory cells using the same lentiviral vector as used previously in monkeys (9). High titer (> 10^9 IU/ml) purified lentivirus was obtained from the University of North Carolina Gene Therapy Center Vector Core. The virus was injected through a 29 gauge needle connected via mineral oil filled tubing to a Hamilton syringe mounted on a perfusion pump (KD Scientific). The needle was advanced by a precision, computer controlled micro-manipulator (NAN instruments) to a pre-established depth (corresponding to the lowest depth at which unit activity was found in preliminary experiments). After a 15-minute wait (to allow for stabilization), 1 µl of virus suspension was delivered over a 10-minute period. The needle was then retracted slowly upwards (0.1 mm/min) in 200-300 µm steps, and an additional 1 µl of virus suspension was delivered at 3-4 additional depths. Five-minute-wait periods were interleaved before and after each virus delivery and retraction steps. Multiple injections were performed in each V1 chamber (8 for Monkey C, 11 for Monkey W) closely grouped together and forming a rectangular pattern. This injection method ensured that the virus was injected evenly across the entire depth of cortical column in all layers (based on CSD analysis performed in preliminary experiments).

*Layer-specific optogenetic stimulation.* Optogenetic stimulation was achieved using a 100 mW, TTL controlled, DPSS blue (473 nm) laser (RGBLase) coupled to a 200 µm optical fiber. The end
of the fiber was inserted into a 356 µm, stainless steel cannula for stability and mounted on the NAN Microdrive. The light intensity at the tip of the cannula was kept to ≤ 10mW/mm² (Coherent Lasermate power meter) such as to target approximately 400 µm of brain area (equivalent to the anatomical measurement of a cortical layer (3)). The cannula-encased fiber was then slowly lowered into the brain at one of the injection sites. Laminar recording electrodes (U-probe, Plexon) were advanced transdurally through adjacent sites on the recording grid at a 0.6 mm (center-to-center) distance from the fiber. The optic fiber and electrodes were mounted separately and could be manipulated independently; they were positioned to minimize the distance between the optical fiber tip and the probe, with the devices often touching at the target depth. The target depth was set to be the center of a given cortical layer (see Methods section: CSD analysis to identify cortical layers). The combination of titrating the light intensity to target 400 µm of tissue and targeting the optical fiber to the center of a cortical layer (cortical layers in V1 macaque span ~400 µm) allowed us to activate neurons in a layer specific manner (Figure 3). After advancing the optic fiber and a recording electrode into the cortex and reaching the injection depth, optical stimulation of the neurons was achieved by delivering 10 ms light pulses at 1 Hz (1 cycle, 2 sessions), 20 Hz (10 or 15 cycles; 4 sessions) or 35 Hz (10 cycles; 16 sessions). The laser output was regulated via TTL pulses driven by a waveform generator (Model 3220A, Agilent Technologies), controlled by the experiment control module (FHC Inc). Increasing light intensity above the constant level used in our study prevented us from distinguishing between action potentials across electrode contacts and spike artifacts due to light onset (visible in saline tests). Conversely, decreasing light intensity made it difficult to activate a sufficient number of cells to examine signal propagation across the cortical column. For these reasons, we titrated light intensity until we reached a satisfactory level that yielded significant activation (9) across the width of four channels on the laminar array (approximately the average width of cortical layers macaque V1).

Optogenetic stimulation in wakefulness and rest. To examine signal propagation in different brain states, a total of 18 sessions (200-720 total trials) were performed with 342 V1 units recorded in monkeys W and C. An experimental day consisted of wakefulness and rest sessions either arranged in rest-awake-rest or awake-rest-awake blocks on a given recording day. The reason for repeating the 1st block (either rest or awake) was to ensure that the effects observed are stable and robust. On average, the ‘awake’ lasted 45 min while a ‘rest’ session lasted 20 min (there was a 15-min inter-block interval). We optogenetically stimulated a subset of units (described in detail in ‘Layer-
specific optogenetic stimulation”) in the laminar column and simultaneously recorded from the entire depth of cortex using laminar U-probes. Optogenetic stimulation was triggered on trials during passive fixation while no visual stimulus was presented on the screen. On each stimulation trial, laser was triggered 300 ms after the monkey acquired fixation (the average ITI was 14 s). Optogenetic stimulation and control (no laser) trials were matched in number and they were randomly and evenly distributed across all trials. During rest, the optogenetic stimulation protocol performed in wakefulness was repeated while lights were turned off (stimulation and no-stimulation trials were randomly interleaved). The duration and number of light stimulation and control trials were identical for the wakefulness and rest conditions. As a control, we varied the frequency of optogenetic stimulation while making sure that the total duration of laser stimulation (100 ms) was held constant throughout a recording session: 35 Hz, 10 cycles, 10 ms width; 20 Hz, 10 cycles, 10 ms width; 1 Hz, 1 cycle, 100 ms width. The frequency of stimulation was held constant across brain state conditions.

**Population coupling**

Population coupling is computed by calculating the cross-correlation between the smoothed firing rate of a neuron and the spike-triggered average of the population (stPR) (10). Population rate used for stPR computation for an individual unit excluded the spikes of that unit. Population rate was computed by accumulating spiking activity of all (multi and single) units with 1 ms resolution and smoothing the resulting vector with a Gaussian of half-width 12 ms (10). The baseline level of each stPR was subtracted. For individual units, the firing rate of a unit was computed with 1 ms resolution while smoothing the resulting vector with a Gaussian of half width of $12/\sqrt{2}$ ms. The population coupling of unit $i$ is calculated as (where $f$ represents the smoothed firing rate of a single unit, $\mu$ represents the firing rate of a single unit and $\|f_i\|$ represents its norm):

$$c_i = \frac{1}{\|f_i\|} \int f_i(t) \sum_{j \neq i} (f_j(t) - \mu_j) \, dt$$

stPRs were normalized by the median size of the stPR of the shuffled data in each recording such that stPRs could be compared across recordings. Spikes were shuffled according to the raster marginals model. This shuffling procedure produces a uniform sample that preserves the mean firing rate and population rate distribution to that of the original data (10). In brief, the data was divided into non-overlapping 1 ms bins, and a binary matrix was constructed with one column for
each time bin and one row for each recorded unit. To shuffle, random 2-by-2 submatrices were repeatedly chosen with each row and column of the submatrix containing a 0 and 1. 1 indicates that a unit spiked at the corresponding time bin, 0 indicates no spikes. The positions of 0s and 1s were then exchanged in the submatrix, which leaves the summed values of each row and column identical.

**Putative excitatory/inhibitory (E/I) ratio**

Spike waveforms of well-isolated single units were spline interpolated to a 2.5 μs resolution. We then computed the time from the trough to the peak of the average waveform and set a threshold of 200 μs to classify the single units as broad-spiking (putative excitatory) and narrow-spiking (putative inhibitory) based on previous studies (5, 11). We used the peak-to-trough time to classify units as this measure provides the most reliable separation of excitatory and inhibitory neurons (12). The mean waveform width of the putative inhibitory neurons was 183.4 ± 7.9 μs and putative excitatory was 400.6 ± 9.6 μs (5, 11, 12) (excitatory vs. inhibitory, Wilcoxon rank sum test, P < 0.0001). Of the 208 single units analyzed, 15% were inhibitory. The E/I ratio was calculated by dividing the mean firing rate of putative excitatory cells by the mean firing rate of putative inhibitory cells for each session.

**Statistical analysis**

We used non-parametric statistical tests throughout the analysis. For rest and wakefulness comparisons, Wilcoxon rank sum test was used. In other cases, we used Chi squared test and Pearson’s correlation for significance analysis. We applied the Holm-Bonferroni correction wherever multiple comparisons were performed. All differences reported in spontaneous and optogenetically induced state transitions were significant in each individual monkey (P < 0.05, Wilcoxon rank sum). Furthermore, there were no significant differences between monkeys (P > 0.05, Chi-squared test).
1. C. B. Beaman, S. L. Eagleman, V. Dragoi, Sensory coding accuracy and perceptual performance are improved during the desynchronized cortical state. *Nat. Commun.* **8**, 1308 (2017).

2. J. T. Daley, R. S. Turner, A. F. Bs, D. L. Bliwise, D. B. Rye, Prolonged Assessment of Sleep and Daytime Sleepiness in Unrestrained Macaca Mulatta. *Sleep* **29**, 221–231 (2017).

3. B. J. Hansen, M. I. Chelaru, V. Dragoi, Correlated Variability in Laminar Cortical Circuits. *Neuron* **76**, 590–602 (2012).

4. T. A. Engel, et al., Selective modulation of cortical state during spatial attention. *Science* **354**, 1140–1144 (2016).

5. J. F. Mitchell, K. A. Sundberg, J. H. Reynolds, Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron* **55**, 131–141 (2007).

6. J. J. Nassi, E. M. Callaway, Parallel processing strategies of the primate visual system. *Nat. Rev. Neurosci.* **10**, 360–372 (2009).

7. N. Niethard, et al., Sleep-Stage-Specific Regulation of Cortical Excitation and Inhibition. *Curr. Biol.* **26**, 2739–2749 (2016).

8. G. Tononi, C. Cirelli, Sleep function and synaptic homeostasis. *Sleep Med. Rev.* **10**, 49–62 (2006).

9. A. R. Andrei, S. Pojoga, R. Janz, V. Dragoi, Integration of cortical population signals for visual perception. *Nat. Commun.* **10**, 1–13 (2019).

10. M. Okun, et al., Diverse coupling of neurons to populations in sensory cortex. *Nature* **521**, 511–515 (2015).

11. M. Rudolph, M. Pospischil, I. Timofeev, A. Destexhe, Inhibition Determines Membrane Potential Dynamics and Controls Action Potential Generation in Awake and Sleeping Cat Cortex. *J. Neurosci.* **27**, 5280–5290 (2007).

12. P. Barthó, et al., Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J. Neurophysiol.* **92**, 600–608 (2004).

13. J. A. Cardin, et al., Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459**, 663–7 (2009).
Figure S1. Overlapping receptive fields within V1 and V4. (A) Schematic of laminar recording using 16 channel U-Probe in macaque primary visual cortex (V1) and mid-level visual cortex (V4). (B) An example session is shown with overlapping V1 and overlapping V4 receptive fields. Circles represent receptive fields of individual neurons recorded along consecutive channels in V1 (black) and V4 (brown).
Figure S2. Desynchronized neural activity during wakefulness in V1. Population spike raster in a subset of fixation trials (n=20) in an example session in monkey W.
Figure S3. Firing rates in On and Off states in V1 and V4. Box plots show firing rates in states classified as On (grey) or Off (red) by the Hidden Markov Model (HMM) in fixation, task and rest in V1 (A) and V4 (B). Horizontal black bars indicate median and box edges represents first quartile (bottom) and third quartile (top). Firing rates in On states is significantly different from Off states in all conditions of fixation, task and rest in V1 (A) and V4 (B) (On vs. Off durations in V1, *P < 0.05, Wilcoxon rank sum test, n=27 sessions; On vs. Off durations in V4, *P < 0.05, Wilcoxon rank sum test, n=26 sessions).
Figure S4. Probability of transition between On and Off states during fixation, task and rest in V1 and V4. Schematic showing the probability of staying in the same state and the probability of transitioning between states (center). Probability of staying in On and Off states during fixation, task, and rest (left and right) represented as bars. Probability of transitioning between On and Off states during fixation, task, and rest in both V1 and V4, as predicted by HMM (Poff→on (top) and Pon→off (bottom)) in rest vs. fixation and task in V1, *P < 0.01, Wilcoxon rank sum test, n=27 sessions; Poff→on (top) and Pon→off (bottom) in rest vs. fixation and task in V4, *P < 0.01, Wilcoxon rank sum test, n=26 sessions).
Figure S5. Dynamics of cortical population activity during wakefulness and rest in V4. (A) Histograms of rate of transition between On and Off states during fixation, task, and rest in V4. Arrows indicate medians and error bars represent s.e.m. (B) Transitions/s plotted as mean and standard deviation during fixation, task, and rest in V4 (transition rate in rest vs. fixation and task, *P < 0.0001, Chi-squared test, n=26 sessions). (C) Scatter plot of mean pupil size and rate of transitions on a trial-by-trial basis (V4, R = -0.14, P < 0.05, Pearson’s correlation, n=384 trials).
Figure S6. State-dependent fluctuations in population activity across cortical layers. Histogram representation of transitions between On and Off states in supragranular (SG), granular (G), and infragranular (IG) layers in V1 (A) and V4 (B). Arrows indicate median and error bars represent s.e.m.
Figure S7. State-dependent fluctuations in population activity across cortical layers in V4. (A) Rate of transition between On and Off states in each cortical layer during fixation and task within V4 (During fixation, layer pairs of SG vs. IG, SG vs G, and G vs. IG in V4, P > 0.05, Chi-squared test, n=13 sessions; During task, layer pairs compared in V4 task, P > 0.05, Chi-squarest test, n=13 sessions). Bars represent mean and error bars represent standard deviation. (B) Rate of transition between On and Off states in each cortical layer during rest within V4 (Layer pairs of SG vs. IG, SG vs G, and G vs. IG in V4, n=9 sessions, P > 0.05, Chi-squared test). Bars represent mean and error bars represent s.e.m.
Figure S8. Probability of transitioning between On and Off states across layers. (A) Mean probability of a transition in a layer relative to a transition in adjacent layers in V1 during wakefulness and rest (40 ms window, **P < 0.001, Wilcoxon rank sum test, n = 13 sessions in awake and 12 sessions in rest) Bars represent s.e.m. (B) Same as A but in V4 (30 ms window, **P < 0.001, Wilcoxon rank sum test, n = 13 sessions in awake and 9 sessions in rest). Bars represent s.e.m.
Figure S9. Optogenetic stimulation yields direct and indirect activation of neurons. (A) Example session – the histogram represents the number of neurons and their light-induced activation latency calculated as the time since laser onset when the firing rate in a given trial exceeded 50% of the peak response in that trial (n=28 units over 6 sessions). Direct responses typically occurred within 2 ms of laser onset (dashed line). Greater latencies were considered indirect(13). (B and C) Mean firing rates of two example neurons showing direct (B) and indirect (C) responses to laser onset. Firing rates were calculated by averaging responses to laser onset across all trials. Error bars show s.e.m.
Figure S10. Propagation of cortical activity during rest and wakefulness does not depend on laser stimulation frequency. (A) Schematic representation of the subset of units analyzed (gray box) with a sliding window and step size of 1 unit (100 μm) progressively moving away from the stimulated units (blue highlight, ‘stim’). (B) Mean firing rates of individual units as a function of distance from the stimulated units in the rest (green) and awake (orange) conditions. Laser stimulation is at 20 Hz (rest vs. awake, *P < 0.05, Wilcoxon rank sum test, n=2 sessions). Error bars represent s.e.m. (C) Same convention as B but for 1 Hz laser stimulation (rest vs. awake, *P < 0.05, Wilcoxon rank sum test, n=2 sessions). Error bars represent s.e.m.
Figure S11. Putative excitatory/inhibitory (E/I) ratio is higher in rest. (A) All waveform traces for an example single unit. Black line represents the mean waveform extracted. (B) Mean waveform shape of putative excitatory (blue) and inhibitory (grey) single units across all sessions. (C) Histogram of waveform widths of putative excitatory and inhibitory units (92 excitatory cells, 41 inhibitory cells across 11 sessions in wakefulness and 7 sessions in rest). Arrows indicate mean waveform widths. (D) Mean putative E/I ratio per session derived from mean firing rates in excitatory and inhibitory units (rest vs. awake, **P < 0.001, Wilcoxon rank sum test, n=23 sessions).
Figure S12. Mean firing rates of neurons recorded during fixation, task, and rest conditions. *P < 0.001, Wilcoxon rank sum test, n=27 sessions. Error bars represent s.e.m. (B) Standard deviation of firing rates of neurons recorded during fixation, task and rest conditions (firing rates show higher variability in the rest condition compared to fixation and task, n = 27 sessions). (C) Mean coefficient of variation of neurons on a trial-by-trial basis for fixation, task, and rest conditions. *P < 0.001, Wilcoxon rank sum test, n=27 sessions. Error bars represent s.e.m.
Figure S13. Cross-validation error plotted for one to four state HMM models.
Figure S14. Standard deviation (A), transition rate (B), and propagation of signals (C) across behavioral states are consistent and significant in each animal (** P < 0.05, Wilcoxon rank sum test).
Figure S15: Left panel: Fraction of time spent in On and Off states across all three behavioral conditions in V1. Right panel: Fraction of time spent in On and Off states in trials with large vs. small pupil size.
Figure S16: On/Off transition rates (top panel) and spike counts (bottom panel) plotted as mean and s.e.m. during laser stimulation for stimulated and non-stimulated layers in awake and rest conditions (transition rate in awake vs. rest non-stim layers, P < 0.001, Wilcoxon rank sum test, n=13 sessions).