Attentional modulation of intrinsic timescales in visual cortex and spatial networks

Roxana Zeraati[1,2], Yan-Liang Shi[3], Nicholas A. Steinmetz[4], Marc A. Gieselmann[5], Alexander Thiele[5], Tirin Moore[6], Anna Levina[7,2,8,*,†], Tatiana A. Engel[3,*,†]

[1] *International Max Planck Research School for the Mechanisms of Mental Function and Dysfunction, University of Tübingen, Tübingen, Germany*

[2] *Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

[3] *Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA*

[4] *Department of Biological Structure, University of Washington, Seattle, WA, USA*

[5] *Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK*

[6] *Department of Neurobiology and Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA*

[7] *Department of Computer Science, University of Tübingen, Tübingen, Germany*

[8] *Bernstein Center for Computational Neuroscience Tübingen, Tübingen, Germany*

[*] *These authors contributed equally to this work*

[†] *Corresponding authors' e-mails: engel@cshl.edu, anna.levina@uni-tuebingen.de*

ABSTRACT

Neural activity fluctuates endogenously on timescales varying across the neocortex. The variation in these intrinsic timescales relates to the functional specialization of cortical areas and their involvement in the temporal integration of information. Yet, it is unknown whether the timescales can adjust rapidly and selectively to the demands of a cognitive task. We measured intrinsic timescales of local spiking activity within columns of area V4 while monkeys performed spatial attention tasks. The ongoing spiking activity unfolded across at least two distinct timescales—fast and slow—and the slow timescale increased when monkeys attended to the receptive fields location. A recurrent network model shows that multiple timescales in local dynamics arise from spatial connectivity mimicking vertical and horizontal interactions in visual cortex and that slow timescales increase with the efficacy of recurrent interactions. Our results reveal that targeted neural populations integrate information over variable timescales following the demands of a cognitive task and propose an underlying network mechanism.

The brain processes information and coordinates behavioral sequences over a wide range of timescales[1–3]. While sensory inputs can be processed as fast as tens of milliseconds[4–7], cognitive processes such as decision making or working memory require integrating information over slower timescales from hundreds of milliseconds to minutes[8–10]. These differences are paralleled by the timescales of intrinsic fluctuations in neural activity across the hierarchy of cortical areas. The intrinsic timescales are defined by the exponential decay rate of the autocorrelation of activity fluctuations. The intrinsic timescales are faster in sensory areas, intermediate in association cortex, and slower in prefrontal cortical areas[11]. The hierarchy of timescales is observed in both spontaneous[11–13] and task-induced neural activity[14–17], and across different recording modalities including spiking activity[11,16,17], intracranial electrocorticography (ECoG)[12,15], and functional magnetic resonance imaging (fMRI)[13,18]. The hierarchy of intrinsic timescales reflects the specialization of cortical areas for behaviorally relevant computations, such as the processing of rapidly changing sensory inputs in lower cortical areas and long-term integration of information (e.g., for evidence accumulation, planning, etc.) in higher cortical areas[19].

The mechanism underlying the diversity of timescales can be related to differences in the connectivity structure observed across cortical areas. The hierarchical organization of timescales correlates with the gradients in the strength of neural connections in different cortical areas[20,21]. These gradients exhibit an increase through the cortical hierarchy in the spine density on dendritic trees of pyramidal neurons[22,23], gray matter myelination[12,24], expression of N-methyl-D-aspartate (NMDA) and gamma-aminobutyric acid (GABA) receptor genes[12,25], strength of structural connectivity measured using diffusion MRI[18], or strength of functional connectivity[13,18,26–28].

The relation between the connectivity and timescales is further supported by computational models. Differences in timescales across cortical areas can arise in network models from differences in the strength of recurrent excitatory connections[23,26,29]. These models matched the increase in the strength of excitatory connections to the increase in spine density of pyramidal neurons[23] or to changes in structural and functional connectivity[26,29]. Moreover, models demonstrate that the topology of connections in addition to the connection strength can affect the timescales of network dynamics. For example, slower timescales emerge in networks with clustered connections compared to random networks[30], or heterogeneity in the strength of inter-node connections gives rise to diverse localized timescales in a one dimensional network[31].

These observations suggest that timescales of neural activity are hardwired in the connectivity structure and thus are fixed characteristics of the dynamics in each brain area. However, cognitive tasks often require flexible changes in the dynamics, which may lead to the modulation of timescales. Indeed, neural timescales measured with ECoG exhibit a widespread increase across multiple cortical association areas during working memory maintenance, which is consistent with the emergence of persistent activity in this period[12]. However, whether the modulation of timescales can occur rapidly and selectively in local neural populations processing specific information during a task has not been tested. It is also unclear whether the timescales can flexibly change in sensory cortical areas and in cognitive processes other than memory maintenance. Moreover, mechanisms that can support the flexible modulation of intrinsic timescales according to behavioral demands are unknown.

To answer these questions, we examined how the timescales of spiking activity in visual cortex were affected by the trial-to-trial alterations in the cognitive state due to visual spatial attention. We analyzed spiking activity recorded from local neural populations within cortical columns in primate area V4 during two different spatial attention tasks and a fixation task. In all tasks, the autocorrelation of intrinsic activity fluctuations deviated from a single exponential decay, which is commonly assumed for estimating the timescale. Instead, the autocorrelation indicated a multiplicity of timescales in the local activity fluctuations. We characterized these timescales using a precise Bayesian estimation method, which confirmed that at least two distinct timescales–one fast and one slow–were present in the local population dynamics. Moreover, the slow timescale was longer on trials when monkeys attended to the receptive fields of the recorded neurons than on trials when they attended to a different location, while the fast timescale did not change.

To identify the mechanisms that can underlie the multiplicity of timescales and their flexible modulation, we developed a network model with local spatial connectivity. The model consisted of interconnected units representing cortical minicolumns, which are the basic anatomical and physiological building blocks of the neocortex[32,33]. With the model, we show analytically and in simulations how multiple timescales in the activity of each minicolumn arise from the spatial structure of recurrent connectivity. The fastest timescale is induced by the recurrent excitation within a minicolumn representing biophysical properties of constituent neurons and vertical connections across cortical layers. A range of slow timescales are induced by horizontal interactions between minicolumns. The timescales depend on the topology of connections, and the slow timescales disappear from the local dynamics in networks with random connectivity. The model indicates that modulation of timescales during attention can be explained by a slight increase in the efficacy of recurrent interactions in visual cortex. Our results suggest that multiple timescales in local population activity arise from the spatial network structure of the neocortex and the slow timescales can flexibly adapt to rapid changes in the cognitive state due to dynamic effective interactions between the neurons.

Results

Multiple timescales in fluctuations of local neural population activity. We analyzed spiking activity of local neural populations within cortical columns of visual area V4 from monkeys performing a fixation task (FT) and two different spatial attention tasks (AT1, AT2)[34,35] (Fig. 1a-c, Supplementary Fig. 1). The activity was recorded with 16-channel linear array microelectrodes from vertically aligned neurons across all cortical layers such that the receptive fields (RFs) of neurons on all channels largely overlapped. In FT, the monkey was rewarded for fixating on a blank screen for 3 s on each trial (Fig. 1a). During AT1, the monkeys were trained to detect changes in the orientation of a grating stimulus in the presence of three distractor stimuli and to report the change with a saccade to the opposite location (antisaccade, Fig. 1b). On each trial, a cue indicated the stimulus that was most likely to change, which was the target of covert attention, and the stimulus opposite to the cue was the target of overt attention due to the antisaccade preparation. During AT2, the monkey was rewarded for detecting a small luminance change in a grating stimulus in the presence of a distractor stimulus

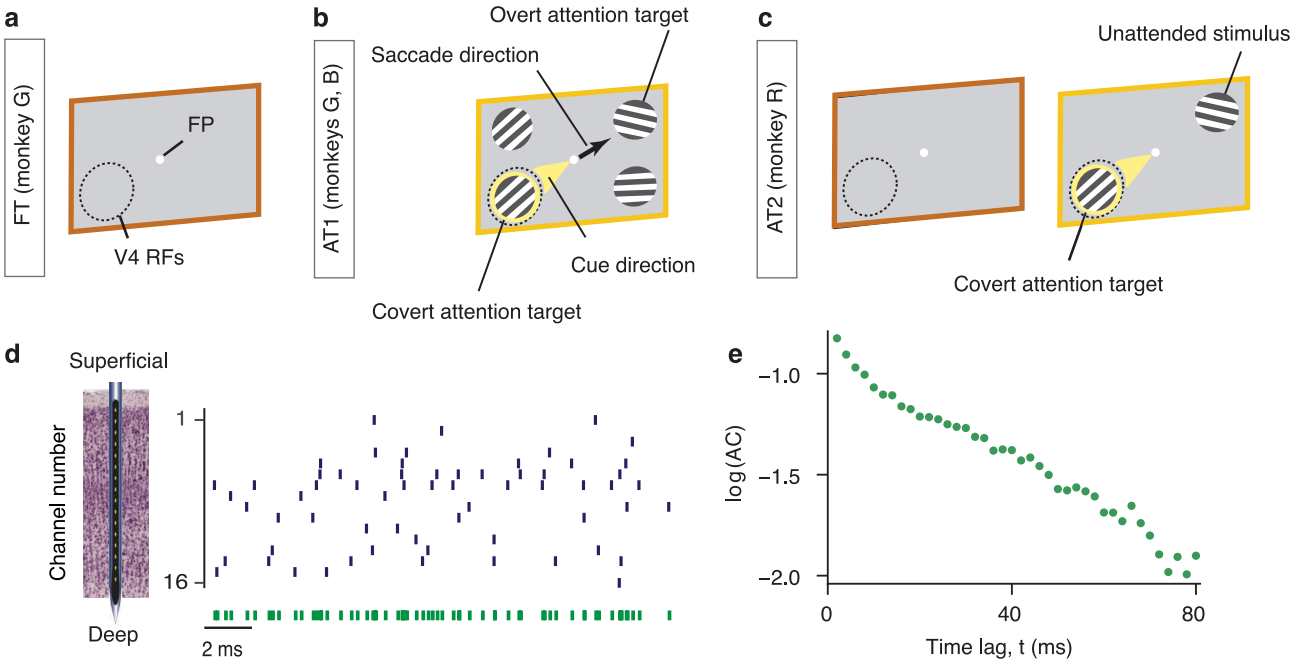

Fig. 1. Computing autocorrelations of spiking activity in V4 columns during fixation and attention tasks. **(a)** In the fixation task (FT), the monkey was rewarded for fixating a central fixation point (FP) on a blank screen for 3 s on each trial. **(b)** In the attention task 1 (AT1), monkeys were trained to detect an orientation change in one of four peripheral grating stimuli, while an attention cue indicated which stimulus was likely to change (yellow spotlight). Monkeys reported the change with a saccade to the stimulus opposite to the change (black arrow). The cued stimulus was the target of covert attention, while the stimulus opposite to the cue was the target of overt attention. **(c)** In the attention task 2 (AT2), the monkey was rewarded for detecting a small luminance change in one of two grating stimuli, directed by an attention cue. The monkey responded by releasing a bar. The brown frame shows the blank screen in the pre-stimulus period. In all tasks, epochs marked with brown frames were used for analyses of spontaneous activity and epochs marked with orange frames were used for the analyses of stimulus-driven activity. The cue was either a vertical line (AT1) or two small dots (AT2). The dashed circle denotes the receptive field locations of recorded neurons (V4 RFs) and was not visible to the monkeys (see Supplementary Fig. 1 for details). **(d)** Multi-unit spiking activity (black vertical ticks) was simultaneously recorded across all cortical layers with a 16-channel linear array microelectrode. The autocorrelation of spike-counts in 2 ms bins was computed from the spikes pooled across all channels (green ticks). **(e)** The autocorrelation (AC) computed from the pooled spikes on an example recording session. Multiple slopes visible in the autocorrelation in the logarithmic-linear coordinates indicate multiple timescales in neural dynamics.

placed in the opposite hemifield. The monkey reported the change by releasing a bar. An attentional cue on each trial indicated the stimulus where the change should be detected, which was the target of covert attention (Fig. 1c).

We analyzed the timescales of fluctuations in local spiking activity by computing the autocorrelations (ACs) of spike counts in 2 ms bins. Previous laminar recordings showed that the neural activity is synchronized across cortical layers alternating spontaneously between synchronous phases of high

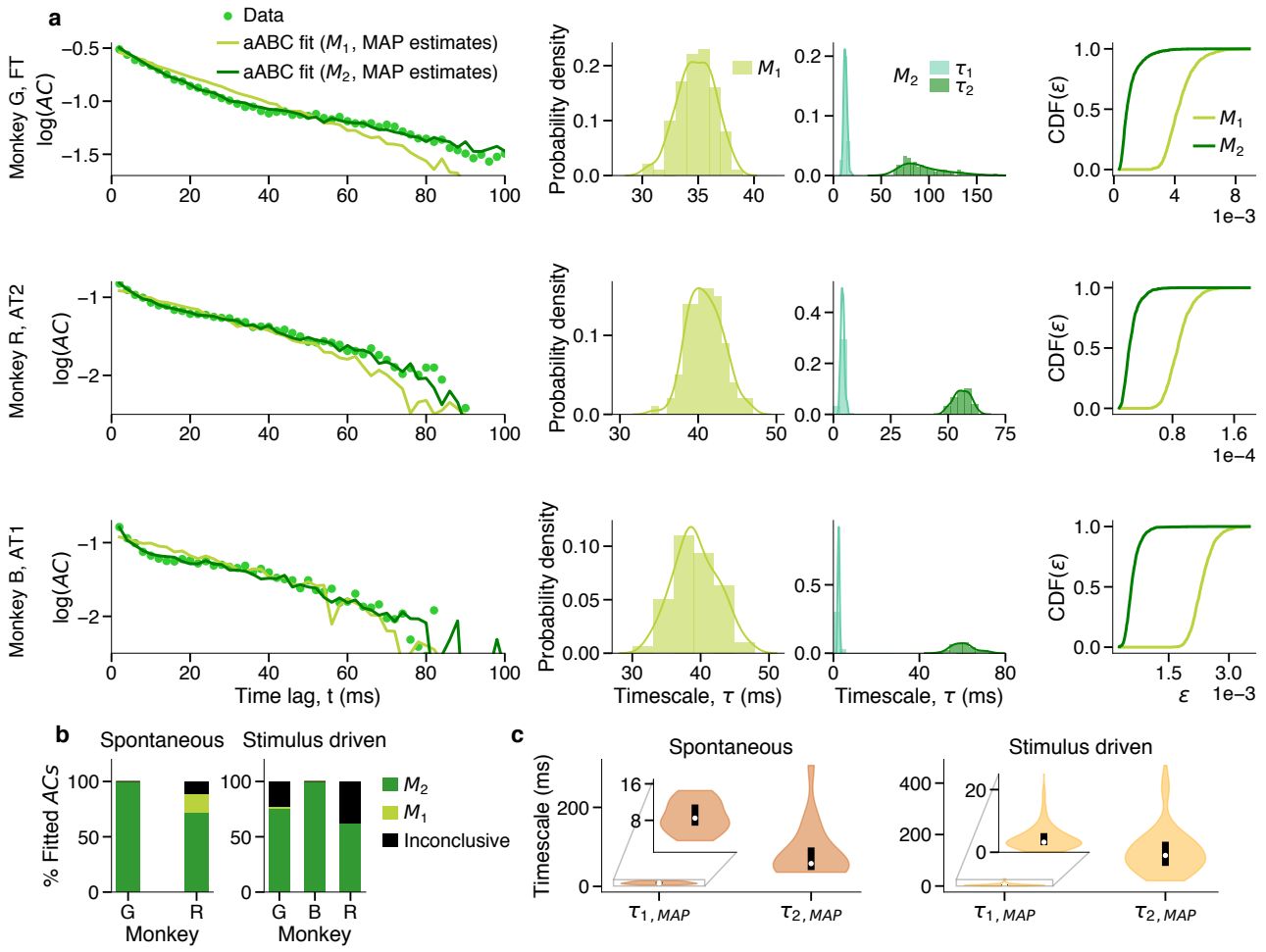

Fig. 2. Two timescales in ongoing spiking activity within V4 columns. **(a)** Comparison between the two-timescale (M_2) and one-timescale (M_1) generative models for three example recording sessions (rows). The models were fitted to autocorrelations of V4 spiking activity using the adaptive Approximate Bayesian Computations (aABC). The shape of the neural autocorrelation (AC) is reproduced by the autocorrelation of synthetic data from the two-timescale model with the maximum *a posteriori* (MAP) parameters, but not by the one-timescale model (left panels). Autocorrelations are plotted from the first time-lag ($t = 2$ ms). Marginal posterior distribution of the timescale estimated by fitting M_1 is in between the posterior distributions of timescales estimated by fitting M_2 (middle panels). Cumulative distribution of errors $\mathrm{CDF}_{M_i}(\varepsilon)$ between the autocorrelations of V4 data and synthetic data generated with parameters sampled from the M_1 or M_2 posteriors (right panels). M_2 is a better fit since it produces smaller errors (i.e. Bayes factor $= \mathrm{CDF}_{M_2}(\varepsilon)/\mathrm{CDF}_{M_1}(\varepsilon) > 1$, Methods). **(b)** In most recording sessions, the autocorrelations during spontaneous and stimulus-driven activity were better described with two distinct timescales (M_2) than a single timescale (M_1). For a few fits the model comparison was inconclusive as the observed statistics were insufficient to distinguish between the models. The total number of fitted autocorrelations for each monkey (G, R, B) was $N_G = 5,\ N_R = 18$ for spontaneous, and $N_G = 57,\ N_R = 24,\ N_B = 39$ for stimulus-driven activity. **(c)** MAP estimates for the fast and slow timescales were heterogeneous across recording sessions during spontaneous and stimulus-driven activity. Violin plots show the distributions of timescales for the autocorrelations that were better fitted with two timescales. The distributions were smoothed with Gaussian kernel densities. The white dot indicates the median, the black box is the first to third quartiles. Inset shows a zoomed range for the fast timescale.

and low firing rates[34,36]. Therefore, we pooled the spiking activity across all layers (Fig. 1d) to obtain more accurate estimates of the spike-count autocorrelations. The shape of spike-count autocorrelations in our data deviated from a single exponential decay. In logarithmic-linear coordinates, the exponential decay corresponds to a straight line with a constant slope. The spike-count autocorrelations exhibited more than one linear slope, with a steep initial slope followed by shallower slopes at longer lags (Fig. 1e). To verify that multiple timescales did not result from pooling activity across channels, we computed cross-correlations of spike counts between different channels. The cross-correlations displayed a similar shape with multiple slopes (Supplementary Fig. 2) confirming that spiking activity on all channels had similar temporal dynamics with the fast and slow decay rates of the correlations. The multiple decay rates in the auto- and cross-correlations indicate the presence of multiple timescales in the fluctuations of local population spiking activity.

To verify the presence of multiple timescales and to accurately estimate their values from autocorrelations, we used a method based on adaptive Approximate Bayesian Computations (aABC, Methods)[37]. This method overcomes the statistical bias in autocorrelations of finite data samples, which undermines the accuracy of conventional methods based on direct fitting of the autocorrelation with exponential decay functions. The aABC method estimates the timescales by fitting the spike-count autocorrelation with a generative model that can have a single or multiple timescales and incorporates spiking noise. The method accounts for the finite data amount, non-Poisson statistics of the spiking noise, and differences in the mean and variance of firing rates across experimental conditions. The aABC method returns a posterior distribution of timescales that quantifies the estimation uncertainty and allows us to compare alternative hypotheses about the number of timescales in the data.

We fitted each autocorrelation with a one-timescale (M_1) and a two-timescale (M_2) generative model and selected the optimal number of timescales by approximating the Bayes factor obtained from the posterior distributions of the fitted models (Fig. 2a, Supplementary Fig. 3, Methods). The majority of autocorrelations were better described by the model with two distinct timescales (M_2) than with the one-timescale model (Fig. 2a,b). The presence of two distinct timescales (fast τ_1 and slow τ_2) was consistent across both spontaneous (i.e. in the absence of visual stimuli, $\tau_{1,MAP} = 8.87 \pm 0.78$ ms, $\tau_{2,MAP} = 85.82 \pm 15.9$ ms, mean $\pm$ s.e.m. across sessions, MAP: Maximum *a posteriori* estimate from the multivariate posterior distribution) and stimulus-driven activity ($\tau_{1,MAP} = 5.05 \pm 0.51$ ms, $\tau_{2,MAP} = 135.87 \pm 9.35$ ms, mean $\pm$ s.e.m.), and across all monkeys, while the precise values of timescales were heterogeneous reflecting subject- or session-specific characteristics (Fig. 2c). These results show that ongoing spiking activity generally exhibits at least two distinct timescales that arise from intrinsic neural dynamics.

Slow timescales are modulated during spatial attention. Next, we examined whether the intrinsic timescales of spiking activity were modulated during spatial attention. We compared the timescales estimated from the stimulus-driven activity on trials when the monkeys attended toward the RFs location of the recorded neurons (attend-in condition, covert or overt) versus the trials when they attended outside the RFs location (attend-away condition). In this analysis, we included recording sessions in which the autocorrelations were better fitted with two timescales in both attend-away and attend-in (covert or overt) conditions. We compared the MAP estimates of the fast τ_1 and slow τ_2

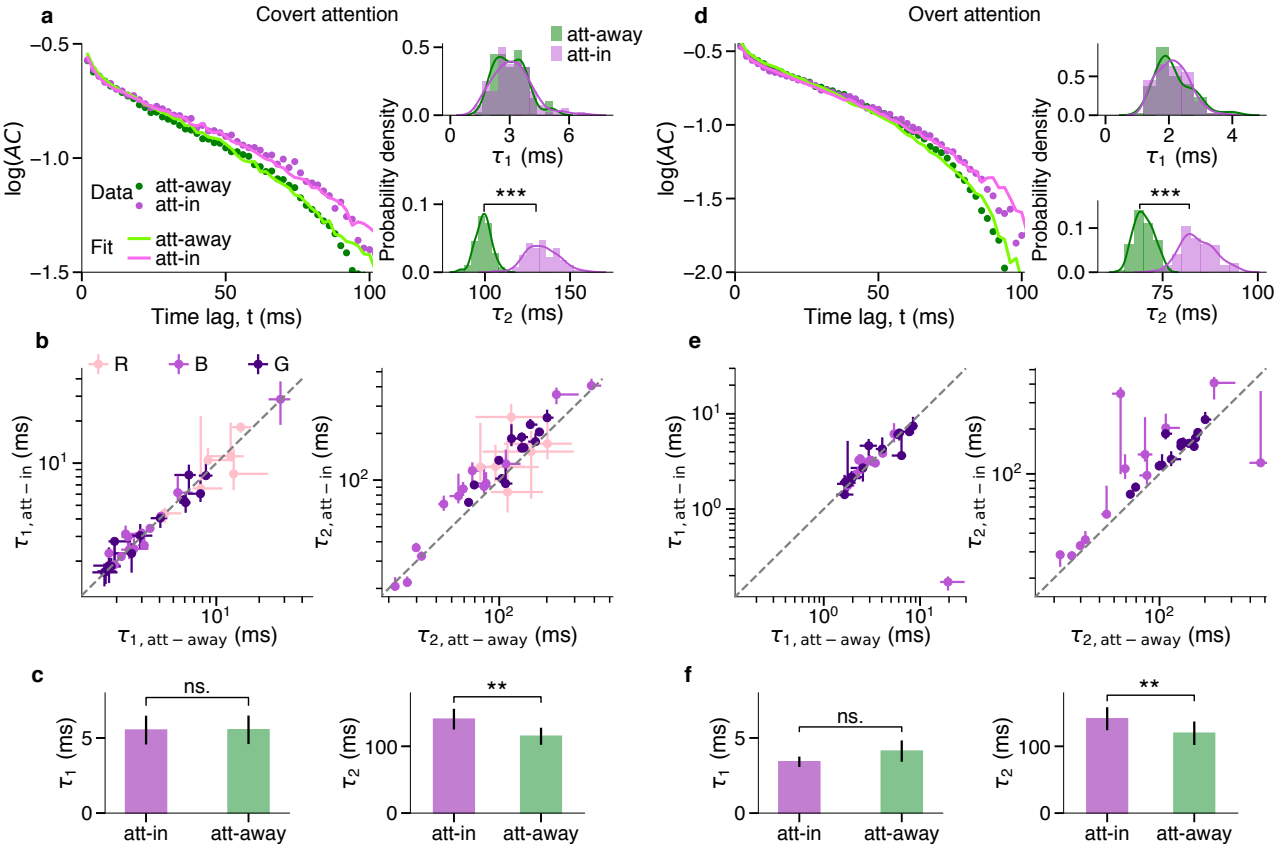

Fig. 3. Slow timescales increase during spatial attention. **(a)** Autocorrelations of neural data with two-timescale fits (left) and the corresponding posterior distributions (right) during covert attention and attend-away condition for an example recording session. The fitted lines are autocorrelations of synthetic data from the two-timescale model with MAP parameters. The posterior distribution of the slow timescale (τ_2) has significantly larger values in attend-in than in attend-away condition. Statistics: Wilcoxon rank-sum test. **(b)** The increase of the slow timescale (τ_2, right) during attention was visible on most sessions (points - MAP estimates for individual sessions, error bars - the first and third quartiles of the marginal posterior distribution, dashed line - the unity line). If the MAP estimate was smaller than the first or larger than the third quartile, the error bar was discarded. Larger error bars indicate wider posteriors, i.e. larger estimation uncertainty. Number of included sessions from the total fitted sessions for each monkey: $N_G = 13/19,\ N_B = 13/13,\ N_R = 6/12$. Color of the dots indicates different monkeys. **(c)** Across sessions, the fast timescale (τ_1, left) did not change, while the slow timescale (τ_2, right) significantly increased during covert attention relative to the attend-away condition. Bar plots show the mean $\pm$ s.e.m of MAP estimates across sessions. Statistics: Wilcoxon signed-rank test. ns., **, *** indicate $p > 0.05/4$, $p < 10^{-2}$, $p < 10^{-3}$, respectively (Bonferroni corrected for 4 comparisons). **(d-f)** Same as **(a-c)** but for the overt attention. Number of included sessions (pairs) from the total fitted sessions for each monkey: $N_G = 14/19,\ N_B = 12/12$.

timescales between attend-in and attend-away conditions across recording sessions.

We found that the slow timescale was significantly longer during both covert and overt attention relative to the attend-away condition (covert: mean $\tau_{2,\text{att-in}} = 140.69$ ms, mean $\tau_{2,\text{att-away}} = 115.07$ ms, $p = 3 \times 10^{-4}$, $N = 32$; overt: mean $\tau_{2,\text{att-in}} = 141.31$ ms, mean $\tau_{2,\text{att-away}} = 119.58$ ms, $p = 7 \times 10^{-4}$,

$N = 26$; Wilcoxon signed-rank test) (Fig. 3), yet there was no significant change in the fast timescale during attention (covert: mean $\tau_{1,\text{att-in}} = 5.53$ ms, mean $\tau_{1,\text{att-away}} = 5.54$ ms, $p = 0.75$, $N = 32$; overt: mean $\tau_{1,\text{att-in}} = 3.42$ ms, mean $\tau_{1,\text{att-away}} = 4.12$ ms, $p = 0.39$, $N = 26$; Wilcoxon signed-rank test). The increase in the slow timescale with attention was evident on individual recording sessions when comparing the marginal posterior distributions of τ_2 for attend-in versus attend-away conditions (Fig. 3a,d). The significant increase of τ_2 was observed in 24 out of 32 individual sessions during covert attention, and 22 out of 26 individual sessions during overt attention. The increase in τ_2 was not due to increase in the firing rate with attention, since the aABC method accounts for the differences in the firing rate across behavioral conditions (Methods), and τ_2 was not correlated with the mean firing rate of population activity (Supplementary Fig. 4). The modulation of the slow timescale was consistent across both attention tasks (AT1 and AT2) and each monkey, and appeared in response to trial-to-trial changes in the cognitive state of the animal directed by the attention cue. These results suggest that the fast and slow timescales of ongoing spiking activity are controlled by different mechanisms, and the mechanisms underlying the slow timescale can flexibly and rapidly adapt according to behavioral demands.

Multiple timescales in local dynamics of a network model with spatial connectivity. To understand the mechanisms underlying the generation and modulation of the neural activity timescales, we developed a recurrent network model mimicking the local population dynamics and connectivity of primate visual cortex. Our model consists of units interconnected on a two dimensional lattice corresponding to lateral dimensions in the cortex. Each unit on the lattice represents a cortical minicolumn[32,33]. In primate cortex, each minicolumn has a diameter of $\sim 50\ \mu m$ and consists of $\sim 80-100$ vertically connected neurons spanning all cortical layers[32,33]. Minicolumns form local spatial clusters through short-range horizontal connections tiling the lateral dimensions of the cortex[33]. To model the local horizontal connectivity between minicolumns, each model unit is connected to the 8 nearby units in its Moore neighborhood (Fig. 4a). The activity of each unit i at time-step t' is described with a binary state variable $S_i(t') \in \{0, 1\}$. The activity $S_i(t')$ stochastically transitions between the active (1) and inactive (0) states driven by the self-excitation (probability p_s), horizontal excitation from neighboring units (probability p_r), and the stochastic external excitation (probability $p_\text{ext} \ll 1$) delivered to each unit (Methods). The self-excitation probability accounts for spontaneous activation of a minicolumn driven by the biophysical properties of constituent neurons and their vertical interactions across layers. The recurrent excitation p_r accounts for horizontal interactions between minicolumns through the spatial lateral connectivity. In this model, the sum of all interaction probabilities is the local branching parameter: $\text{BP} = p_\text{s} + 8p_\text{r}$. The branching parameter describes the expected number of units that can be activated by a single active unit i in the network. At the critical point $\text{BP} = 1$, each unit on average would activate one other unit creating the self-sustained activity.

We measured the timescales of local dynamics in the model using the autocorrelations of individual units' activity. Similar to the V4 data, the activity of each model unit exhibited multiple distinct timescales arising from the recurrent network dynamics shaped by the spatial connectivity structure. The autocorrelation of the model units exhibited a steep decay at short time-lags and a shallower decay at longer time-lags (Fig. 4b).

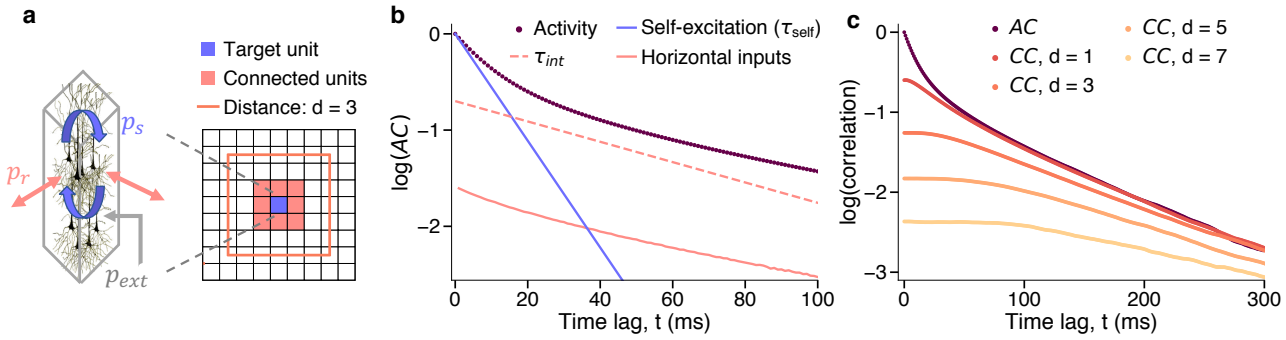

Fig. 4. Multiple timescales in dynamics of a network model with spatial connectivity structure. (a) Schematic of the local spatial connectivity in the network. Each unit (blue) on the lattice is connected to 8 other units (pink) in its Moore neighborhood. The units represent cortical minicolumns (left) that are driven by the self-excitation (probability p_s, blue arrows), horizontal excitation from neighboring units (p_r, pink arrows), and the stochastic external input (p_{ext}, gray arrow). Pink dashed line indicates the units at distance 3 from the blue unit. **(b)** Autocorrelation (AC) of individual units' activity (brown) exhibits multiple timescales. The fast timescale (τ_{self}) at short time-lags can be approximated by the autocorrelation of a 2-state Markov process driven only by p_s and p_{ext} (blue). The slower decay at longer time-lags is captured in simulations by the autocorrelation of horizontal inputs received by each unit (pink) and can be approximated analytically with a dominant interaction timescale (τ_{int}, dashed line). **(c)** Interaction timescales are visible in cross-correlations (CC) between activity of different units, but the self-excitation timescale vanishes. With increasing distance, the strength of cross-correlations decreases, faster interaction timescales (higher spatial frequency modes) vanish, and slower interaction timescales (lower spatial frequency modes) dominate. To compute cross-correlations, the same number of units were randomly selected for each distance ($p_s = 0.88, 8p_r = 0.11, p_{ext} = 10^{-4}$).

In the model, we were able to identify the mechanisms that give rise to the multiple timescales in network dynamics. The fast timescale (steep initial decay in the autocorrelation) corresponds to the self-excitation timescale (τ_{self}) induced by spontaneous self-activation of each unit in the absence of network interactions. We can estimate this timescale analytically as the autocorrelation timescale of a 2-state Markov process (2SMP) driven only by the self-excitation and the external input: $\tau_{self} = (-\log(p_s))^{-1}$. The autocorrelation decay rate at short time-lags agreed well with τ_{self} (Fig. 4b).

The slow timescales (shallower autocorrelation decay) are the interaction timescales induced by interactions among units in the network. In simulations, we can capture these timescales by computing the autocorrelation of the summed horizontal input received by a unit from all its connected neighbors (excluding the self-excitation input). The autocorrelation of the horizontal input closely corresponds with the slow decay of the autocorrelation at long time-lags (Fig. 4b). Using the master equation for binary units with Glauber dynamics[38], we derived dynamical equations for auto- and cross-correlations and computed the autocorrelation timescales analytically (Methods and Supplementary Note 1). We found that the slow decay of the autocorrelation consists of a mixture of interaction timescales $\tau_{int,\mathbf{k}}$. Each $\tau_{int,\mathbf{k}}$ arises from interactions between the units on a different spatial scale of the network. The spatial scales of interactions can be characterized by the correlated fluctuation modes with spatial

frequencies ($\mathbf{k}$) in the Fourier space (Methods). At each spatial frequency, the interaction timescale depends on both the probability of horizontal interactions (p_r) and the self-excitation probability (p_s): $\tau_{\mathrm{int},\mathbf{k}} \approx (1 - p_\mathrm{s} - p_\mathrm{r} f(\cos(\mathbf{k})))^{-1})$ (Methods Eq. 18). This analytical equation shows that shorter timescales arise from higher spatial frequency modes (larger $\mathbf{k}$) which correspond to persistent activity in local neighborhoods. Our calculations further show that shorter timescales created by local interactions between the neighbors (large $\mathbf{k}$, high spatial frequency modes) have a larger weight in the autocorrelation than longer timescales created by more global interactions (small $\mathbf{k}$, low spatial frequency modes). Thus, longer timescales appear only in the tail of autocorrelation where the correlations on shorter timescales decay to zero. We can approximate the slow decay of the autocorrelation with a single dominant interaction timescale (τ_int) defined based on the dominant spatial frequency mode (Methods). Therefore, the shape of individual units' autocorrelation is well captured with two timescales: the fast self-excitation timescale and the slow dominant interaction timescale.

The interaction timescales observed in autocorrelations are also evident in the cross-correlations between the units' activity, because the neighboring units share recurrent inputs (Fig. 4c). The self-excitation timescale, on the other hand, disappears in cross-correlations, since self-excitation dynamics are independent in different units. With increasing distance between the units, interaction timescales created by local interactions (high spatial frequency modes) vanish from cross-correlations and more globally generated interaction timescales (low spatial frequency modes) dominate. Moreover, the strength of cross-correlations decays with increasing distance, which is consistent with the reduction of the time-resolved pairwise correlations between distant neurons in primate visual area V4[39].

These results suggest that the experimentally observed fast and slow timescales in local neural dynamics correspond to the self-excitation and interaction timescales in the model. The fast timescale is generated from intrinsic dynamics of each minicolumn and the slow timescales are created by recurrent interactions between the minicolumns. Multiple timescales are present in local dynamics of spatially connected networks with different sizes and can be observed in smaller networks with a realistic number of minicolumns within a single cortical column (Supplementary Fig. 5). Moreover, our model makes testable predictions beyond our experimental observations. The model predicts that if we simultaneously record from multiple distant columns in the cortex (e.g., with multiple laminar probes), then with increasing lateral distance the cross-correlations become weaker and increasingly dominated by slower timescales.

Local timescales are shaped by the spatial network structure. Next, we investigated how the timescales of local dynamics depend on the spatial connectivity structure. We systematically modified the range of recurrent connections between units in the model, while keeping the strength and number of connections constant. We fixed a connectivity radius r and connected each unit to its 8 randomly chosen neighbors within this radius (Fig. 5a). For small r in the model, the connections between units are local (as in Fig. 4). With increasing r, the connections become more dispersed, and when r reaches the network size we get a randomly connected network.

We found that timescales of local dynamics reflect the underlying spatial network structure (Fig. 5b). Autocorrelations (AC) measured from the activity of individual units decay faster in networks with

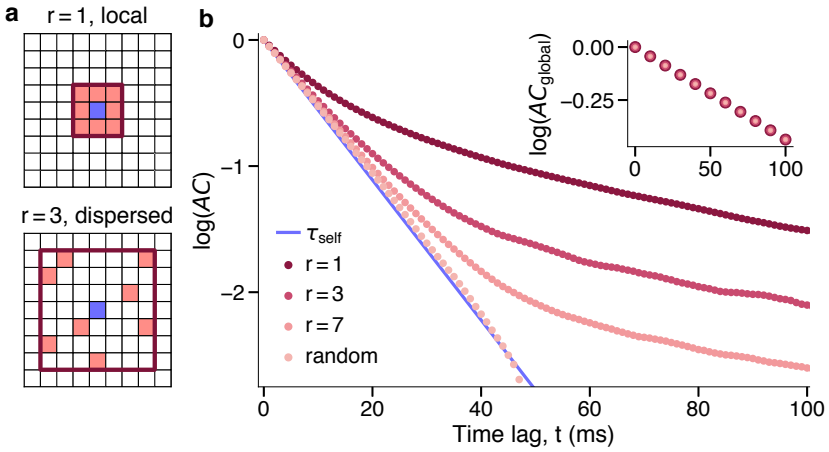

Fig. 5. Local timescales reveal the spatial network structure. **(a)** Schematic of local ($r = 1$) and dispersed ($r > 1$) spatial connectivity in the network model. Each unit (blue) is connected to 8 other units (pink) selected randomly within the connectivity radius r (brown line). **(b)** Shape of the autocorrelations of individual units (AC) reflect the underlying local connectivity structure. Interaction timescales disappear and the self-excitation timescale (τ_{self}) dominates local autocorrelations when the connectivity radius increases while the connection strengths are kept constant ($p_{\text{s}} = 0.88, 8p_{\text{r}} = 0.11, p_{\text{ext}} = 10^{-4}$). The autocorrelation of the the global network activity (AC_{global}, inset) does not depend on the connectivity structure.

more dispersed connectivity compared to locally connected networks. In addition, the weight of interaction timescales in the autocorrelation decreases with increasing connectivity radius. As a result, the transition point between the self-excitation and interaction timescales shifts towards longer time-lags so that the self-excitation timescale dominates the autocorrelation in networks with dispersed connectivity.

While networks with local versus dispersed connectivity generate distinct local timescales, the timescale of their global dynamics is similar (Fig. 5b, inset). The timescale of global network activity is a direct measure for distance of system's dynamics from the critical point, with longer timescales observed in systems closer to criticality[40]. We measured the global timescale from the autocorrelation of the activity summed across all units in the network (AC_{global}). This autocorrelation exhibits only one timescale that is consistent across networks with different connectivity structure. The global timescale is equal to the slowest interaction timescale in the autocorrelations of local activity related to the zero spatial frequency mode: $\tau_{\text{global}} = (1 - p_{\text{s}} - 8p_{\text{r}})^{-1}$ (Methods). However, this timescale has a vanishingly small weight in local autocorrelations and is hard to observe empirically in local autocorrelations (it requires data with excessively long trial duration).

The relation between the spatial network structure and the timescales of local dynamics is a general phenomenon that is observed in the networks with linear (Fig. 5b) and non-linear (Supplementary Fig. 6, Methods) dynamics. Moreover, this relation is independent of the number of connections per unit as long as they are uniformly distributed within the connectivity radius and their strength is normalized to the same total strength. For instance, if instead of selecting 8 random neighbors, we connect each unit to all other units within the radius r, the qualitative results stay the same (Supple-

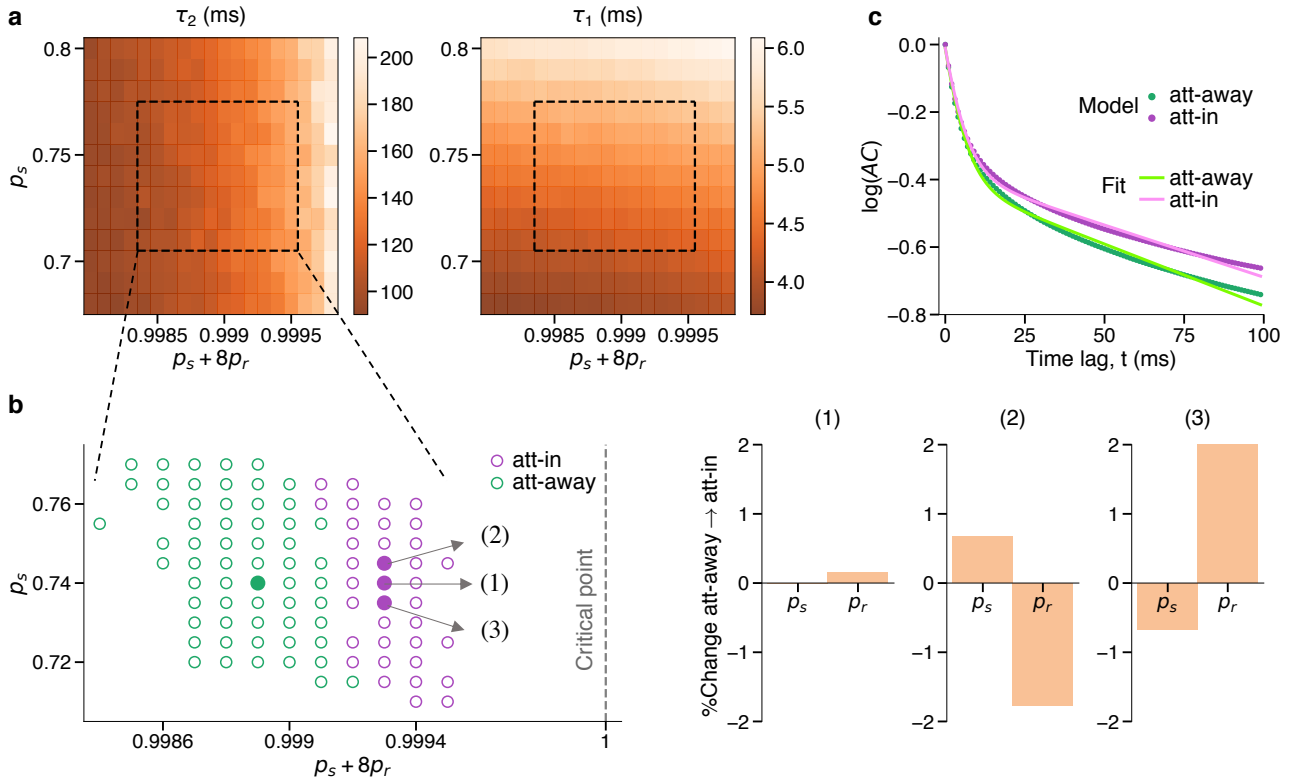

Fig. 6. Modulation of the slow timescale during attention is mediated by an increase in the efficacy of network interactions. **(a)** Effect of connectivity parameters on local timescales in the model. The fast timescale (τ_1, right) mainly depends on the self-excitation probability (p_s), whereas the slow timescale (τ_2, left) depends on both the self-excitation (p_s) and recurrent horizontal interactions (p_r). The dashed rectangles indicate the range of parameters reproducing V4 timescales (mean $\pm$ s.e.m. of MAP estimates, Methods). **(b)** The slow timescale increases with the network excitability ($p_s + 8p_r$, left panel). Green and magenta dots indicate the parameters reproducing attend-away and attend-in timescales, respectively. Filled dots show examples of experimentally observed $\sim$20% increase in τ_2 for three possible scenarios based on different changes in p_s or p_r (right panels). Larger changes of parameters in scenarios (2) and (3) are due to coarser grid of p_s used to fit the timescales. A similar change of τ_2 can be achieved also with smaller changes in p_s and p_r (e.g., for all $0.74 < p_s < 0.745$ in scenario 2). **(c)** Example autocorrelations (ACs) from the model simulations with the attend-in and attend-away parameters for the scenario (2) in b. We fitted unbiased autocorrelations from the model simulations with double exponential functions (green and pink lines) to estimate the two timescales (Methods).

mentary Fig. 7). Our analytical and numerical results for the network model suggest that the local connectivity structure between minicolumns in primate cortex[32,33] can give rise to multiple timescales in the dynamics of local neural population activity, as we detected in our V4 data.

Changes in the efficacy of network interactions modulate local timescales. Finally, we investigated which mechanisms can underlie the modulation of the slow timescales during attention using our network model. We matched the timescales between the model and experimental data to determine what changes of the model parameters can explain the attentional modulation of timescales in

V4. In the network model with local connectivity ($r = 1$), we adjusted the parameters to match the self-excitation and dominant interaction timescales of a model unit to, respectively, the fast and slow timescales of V4 activity (mean timescale $\pm$ s.e.m., Methods) for both the attend-away and attend-in (averaged over covert and overt) conditions (Fig. 6). We used a combination of analytical approximations and model simulations to find parameters that produce timescales similar to the V4 data (Methods).

We found that to reproduce the increase of the slow timescale in V4 during attention, the total excitability of the network interactions should increase, shifting the network dynamics closer to the critical point (Fig. 6b). The model parameters indicate that the overall increase in the interaction strength can be achieved by increasing the strength of either the self-excitation (p_s) or the horizontal interactions (p_r). Small changes of horizontal recurrent interactions during attention[41] can be mediated by weak top-down inputs from higher cortical areas[42]. Increasing p_r while keeping p_s constant allows for substantial changes in the slow timescale and nearly unchanged fast timescale consistent with the V4 data. The constant p_s may reflect the fixed biophysical properties of neurons within a minicolumn.

The increase in p_s, on the other hand, is consistent with the observation that interactions between cortical layers in V4 increase during attention[43]. The increase in p_s can be counterbalanced by a reduction in p_r mediated by neuromodulatory effects that reduce the efficacy of lateral connections in the cortex during attention[44]. The increase of p_s in the model produces a slight increase in the fast timescale (τ_1) (about ~ 0.4 ms on average), but such small changes in τ_1 would not be detectable with our available data amount (the uncertainty of τ_1 MAP estimate is ± 0.9 ms on average, Fig. 3b,e). Altogether, our model suggests that attentional modulation of timescales can arise from changes in the efficacy of vertical or horizontal recurrent interactions in visual cortex.

Discussion

We found that ongoing spiking activity of local neural populations within columns of the area V4 unfolded across fast and slow timescales, both in the presence and absence of visual stimuli. The slow timescale increased when monkeys attended to the receptive fields location, showing that local intrinsic timescales can change flexibly from trial to trial according to selective attention. To understand the mechanisms underlying the multiplicity and flexible modulation of timescales observed in our V4 data, we developed a network model with spatially structured connections. Our model suggested that multiple timescales in local neural activity can arise from the local spatial network structure of primate visual cortex. A fast timescale is induced by the recurrent excitation within a minicolumn, and a set of slow timescales are induced by recurrent interactions via short-range horizontal connections between minicolumns. The model also indicated that modulation of the slow timescale during attention can arise from an increase in the efficacy of recurrent interactions between the neurons.

Multiple intrinsic timescales in neural activity. Previous studies characterized the autocorrelation of ongoing neural activity with a single intrinsic timescale[11–13,18]. The intrinsic timescale was usually

measured for neural populations either by averaging autocorrelations of single neurons in one area[11] or using coarse-grained measurements such as ECoG[12] or fMRI[13,18]. Thus, ongoing dynamics in each area were described with a single intrinsic timescale that varied across areas. We extended this view by showing that, within one area, local population activity exhibits multiple intrinsic timescales. These timescales reflect ongoing dynamics on single trials and are not driven by task events. Our results suggest that the multiplicity of timescales is an intrinsic property of neural activity arising from inherent cellular and network properties of the cortex.

We show that multiple timescales in local dynamics can emerge from the spatial connectivity structure between minicolumns in a recurrent network model. The presence of two dominant timescales (τ_{self}, τ_{int}) in local dynamics depends on the combination of the structured connectivity and strong, mean-driven interactions between units. Networks with random connectivity (Fig. 5,b) or weak, diffusion-type interactions[45] exhibit only one dominant timescale in local activity (Supplementary Note 2). Biophysical mechanisms such as adaptation currents can induce slow timescales in firing of single neurons. However, the slow adaptation in single neurons cannot generate slow timescales on the population level[46], highlighting the importance of the network structure for emergence of slow timescales in population activity.

In our network model with local spatial connectivity, recurrent interactions across different spatial scales induce multiple slow timescales. To generate multiple slow timescales, our network operates close to a critical point. Spiking networks with spatial connectivity can generate fast correlated fluctuations that emerge from instability at particular spatial frequency modes[47]. Slow fluctuations of firing rates can also arise in networks with clustered random connectivity, but interactions between clusters induce only a single slow timescale[30]. We show that more local spatial connectivity (smaller r) leads to slower dynamics and modifies the weights and composition of timescales in the local activity. The timescale of the global activity, on the other hand, is the same across networks with distinct local timescales and different connectivity structures. Therefore, comparing the timescales of local and global dynamics can reveal the underlying network structure to distinguish between random and spatially organized connectivity.

In our model, integrating activity over larger spatial scales leads to disappearance of faster interaction timescales (higher spatial frequencies) leaving only slower interaction timescales (lower spatial frequencies) in the coarse-grained activity. At the extreme, the global network activity exhibits only the slowest interaction timescale (the global timescale). This mechanism may explain the prominence of slow dynamics in meso- and macroscale measures of neural activity such as LFP or fMRI[48], while faster dynamics dominate in local measures such as spiking activity. The model predicts that the slowest interaction timescales have very small weights in the autocorrelation of local neural activity and thus can be detected in local activity only with excessively long recordings. Indeed, infraslow timescales (on the order of tens of seconds and minutes) are evident in the cortical spiking activity recorded over hours[49].

Functional relevance of neural activity timescales. Intrinsic timescales are thought to define the predominant role of neurons in the cognitive processes[19]. For example, in the orbitofrontal cortex, neurons with long intrinsic timescales are more involved in decision-making and the maintenance of

value information[50]. In the prefrontal cortex (PFC), neurons with short intrinsic timescales are primarily involved in the early phases of working memory encoding[27], while neurons with long timescales play a significant role in coding and maintaining information during the delay period[27,51]. In addition to intrinsic timescales, neurons exhibit changes of firing rates over multiple trials encoding various task events, which are characterized by task-induced timescales[16,17]. The task-induced timescales of single neurons do not correlate with intrinsic timescales measured over the entire task duration[17]. Our finding that intrinsic timescales can flexibly change from trial to trial (and across epochs within a trail[12]) suggests a possibility that task-induced timescales may correspond with intrinsic timescales only during specific task phases.

We found that timescales of local neural activity changed from trial to trial depending on the attended location. A previous ECoG study found that the intrinsic timescale of neural activity in cortical association areas increased after engagement in a working memory task[12]. Our findings go beyond this earlier work by showing that the modulation of timescales can be functionally specific as it selectively affects only neurons representing the attended location within the retinotopic map. Moreover, the timescales change rapidly from one trial to the next. While changes in timescale due to task engagement could be mediated by slow global processes such as arousal, the retinotopically precise modulation of timescales requires local changes targeted to task-relevant neurons. Our results further show that the modulation of timescales also occurs in sensory cortical areas and cognitive processes other than memory maintenance. The increase of neural timescales with selective attention may be functionally relevant, potentially allowing neurons to integrate information over longer durations.

Longer timescales during attention in the model are associated with shifting the network dynamics closer to a critical point. Shifting closer to criticality was also suggested as a mechanism for the increase in gamma-band synchrony and stimulus discriminability during attention[52]. Operating closer to the critical point during attention might help to optimize neural responses to environmental cues and improves information processing[53].

Mechanisms for attentional modulation of timescales. Changes in the slow timescale of neural activity due to attention occurred rapidly from one trial to another. Such swift changes cannot be due to significant changes in the underlying network structure and require a fast mechanism. Our model suggests that the modulation of slow timescales during attention can be explained with a slight increase in the network excitability mediated by an increase in the efficacy of vertical or horizontal recurrent interactions. In particular, an increase in the efficacy of vertical interactions can be accompanied by a decrease in the strength of horizontal interactions.

Several physiological processes may underlie these network mechanisms in the neocortex. Top-down inputs during attention can enhance the local excitability in cortical networks[42] and increase the effective horizontal interactions between neurons[41]. Furthermore, feedback connections from higher visual areas like PFC or the temporal-occipital area (TEO) to lower visual areas have broader terminal arborizations than the size of the receptive fields in lower areas [54,55]. These feedback inputs can coordinate activity across minicolumns in V4. Moreover, vertical interactions in V4 measured with local field potentials (LFPs) increase during attention[43], while neuromodulatory mechanisms can reduce horizontal interactions. The level of Acetylcholine (ACh) can modify the efficacy of synaptic

interactions during attention in a selective manner[44]. Increase in ACh strengthens the thalamocortical synaptic efficacy by affecting nicotinic receptors and reduces the efficacy of horizontal recurrent interactions by affecting muscarinic receptors. Decrease in horizontal interactions is also consistent with the proposed reduction of spatial correlations length during attention[45]. These observations suggest that an increase in vertical interactions and a decrease in horizontal interactions is a likely mechanism for modulation of the slow timescale during attention.

To identify biophysical mechanisms of timescales modulation, experiments with larger number of longer trials are required to provide tighter bounds for estimated timescales. Additionally, detailed biophysical models can help distinguish different mechanisms, since biophysical and cell-type specific properties of neurons might also be involved in defining neural timescales[56]. Finally, perturbation experiments that modulate selectively top-down inputs or neuromodulatory levels can provide the most direct test of the underlying mechanisms.

Our findings reveal that targeted neural populations can integrate information over variable timescales following demands of a cognitive task. Our model suggests that local interactions between neurons via the spatial connectivity of primate visual cortex can underlie the multiplicity and flexible modulation of intrinsic timescales. Our experimental observations combined with the computational model provide a basis for studying the link between the network structure, functional brain dynamics, and flexible behavior.

Methods

Behavioral tasks and electrophysiology recordings. Experimental procedures were described previously[34,35]. Experimental procedures for the fixation task and attention task 1 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. Experimental procedures for the attention task 2 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.

In brief, on each trial of the fixation task (FT, monkey G), the monkey was rewarded for fixating a central dot on a blank screen for 3 s. In attention task 1 (AT1, monkeys G, B), the monkey detected orientation changes in one of the four peripheral grating stimuli while maintaining central fixation. Each trial started by fixating a central fixation dot on the screen and after several hundred milliseconds (170 ms for monkey B and 333 ms for monkey G), four peripheral stimuli appeared. Following a $200-500$ ms period, a central attention cue indicated the stimulus that was likely to change with $\sim 90\%$ validity. Cue was a short line from fixation dot pointing toward one of the four stimuli, randomly chosen on each trial with equal probability. After a variable interval ($600-2200$ ms), all four stimuli disappeared for a brief moment and reappeared. Monkeys were rewarded for correctly reporting the change in orientation of one of the stimuli (50% of trails) with an antisaccade to the location opposite to the change, or maintaining fixation if none of the orientations changed. Due to the anticipation

of antisaccade response, the cued stimulus was the target of covert attention, while the stimulus in location opposite to the cue was the target of overt attention. In attend-in conditions, the cue pointed either to the stimulus in the RFs of the recorded neurons (covert attention) or to the stimulus opposite to the RFs (overt attention). The remaining two cue directions were attend-way conditions.

In attention task 2 (AT2, monkey R), the monkey detected a small luminance change within the white phase of a square wave static grating. The monkey initiated a trial by holding a bar and visually fixating a fixation point. The color of the fixation point indicated the level of spatial certainty (red: narrow focus, blue: wide focus). After 500 ms a cue appeared indicating the location and focus of the visual field to attend to. The cue was switched off after 250 ms. After another second two gratings appeared, one in the center of the RFs and one diametrically opposite with respect to the fixation point. The grating at the position indicated by the cue was the test stimulus. The other grating served as the distractor. After at least 500 ms a small luminance change (dimming) occurred either in the center of the grating (narrow focus) or in one of 12 peripheral positions (wide focus). If the dimming occurred in the distractor grating first the monkey had to ignore it. The monkey was rewarded for a bar release within 750 ms of the dimming in the test grating. The faster the monkey reacted, the larger reward it received. Two grating sizes (small and large) were used in this experiment. We analyzed trials with the small grating to avoid surround-suppression effects created by the large grating sizes extending beyond the neurons' summation area[57].

Recordings were performed in the visual area V4 with linear array microelectrodes inserted perpendicularly to the cortical layers. Arrays were placed such that receptive fields of recorded neurons largely overlapped. Each array had 16 channels with 150 μm center-to-center spacing. In AT1 and FT, all 16 channels were visually responsive. In AT2, the number of visually-responsive channels per recording ranged between 8 and 12 with the median at 9.

Computing autocorrelations of neural activity. We computed autocorrelations from multi-unit (MUA) spiking activity recorded in the presence (stimulus-driven) and absence (spontaneous) of visual stimuli (brown and yellow frames in Supplementary Fig. 1). For spontaneous activity, we analyzed spikes during the 3s fixation epoch in FT, and during the 800 ms epoch from 200 ms after the cue offset until the stimulus onset in AT2. For stimulus-driven activity, we analyzed spikes in the epoch from 400 ms after the cue onset until the stimulus offset in AT1, and from 200 ms after the stimulus onset until the dimming in AT2. For the stimulus-driven activity, trials in both attention tasks had variable durations ($500-2200$ ms). Thus, we computed autocorrelations in non-overlapping windows of 700 ms for AT1 and 500 ms for AT2. On long trials, we used as many windows as would fit within the trial duration, and we discarded trials that were shorter than the window size. The duration of windows were selected such that we had at least 50 windows for each condition in each session. 3 out of 25 recording sessions in monkey G (AT1) were excluded due to short trial durations. For spontaneous activity, the windows were 3 s in FT and 800 ms in AT2.

We computed the average spike-count autocorrelation for each recording session. On each trial we pooled the spikes from all visually-responsive channels and counted the pooled spikes in 2 ms bins. For each behavioral condition (stimulus orientation, attention condition), we averaged spike-counts at each time-bin across trials, and subtracted the trial-average from the spike-counts at each bin[11] to

remove correlations due to changes in firing rate locked to the task events. We segmented the mean-subtracted spike-counts $A(t'_i)$ into windows of the same length N, where t'_i $(i = 1 \dots N)$ indexes bins within a window. We then computed the autocorrelation in each window as a function of time-lag t_j[37]:

$$AC(t_j) = \frac{1}{\hat{\sigma}^2(N-j)} \sum_{i=1}^{N-j} \left(A(t'_i) - \hat{\mu}_1(j)\right)\left(A(t'_{i+j}) - \hat{\mu}_2(j)\right). \quad (1)$$

Here $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(A(t'_i)^2 - \frac{1}{N^2}(\sum_{i=1}^{N} A(t'_i))^2)$ is the sample variance, and $\hat{\mu}_1(j) = \frac{1}{N-j}\sum_{i=1}^{N-j} A(t'_i)$ and $\hat{\mu}_2(j) = \frac{1}{N-j}\sum_{i=j+1}^{N} A(t'_i)$ are two different sample means. In Eq.(1) for autocorrelation, we subtracted window-specific mean to remove correlations due to slow changes in firing rate across trials, such as slow fluctuations related to changes in the arousal state. Finally, we averaged the autocorrelations over windows of the same behavioral condition separately for each recording session. The exact method of computing autocorrelations does not affect the estimated timescales, since we use the same method for computing autocorrelations of synthetic data when fitting generative models with the aABC method[37].

In AT1, we averaged autocorrelations over trials with different stimulus orientation for each attention condition, since all attention conditions contained about the same number of trials with each orientation. For stimulus-driven activity in AT2, we first estimated timescales separately for focus wide and narrow conditions and found no significant differences (Wilcoxon signed rank test between MAP estimates, $p > 0.05$). Thus, we averaged autocorrelations of the focus narrow and wide conditions and refitted the average autocorrelations. The same procedure was applied to the spontaneous activity in AT2, and since there was no significant differences in timescales between different focus or attention conditions (Wilcoxon signed rank test between MAP estimates for the two-by-two conditions, $p > 0.05$), we averaged the autocorrelations over all conditions and refitted the average autocorrelation.

For estimating the timescales, we excluded sessions with autocorrelations dominated by noise or strong oscillations that could not be well described with a mixture of exponential decay functions. We excluded a session if the autocorrelation fell below 0.01 ($\log(AC)$ fell below -2) in lags smaller or equal to 20 ms (Supplementary Fig. 8). Based on this criterion, we excluded 3 out of 22 sessions for monkey G in AT1, 8 out of 21 sessions during covert attention and 9 out of 21 during overt attention for monkey B in AT1, 2 out 20 sessions for spontaneous activity and 8 out 20 sessions for stimulus-driven activity for monkey R in AT2. The difference in the number of excluded sessions for monkey R during spontaneous and stimulus-driven activity is explained by the larger amount of data availabile for computing autocorrelations during spontaneous activity due to averaging over attention conditions and longer window durations (800 ms vs. 500 ms).

For visualization of autocorrelations, we omitted the zero time-lag ($t = 0$ ms) (examples with the zero time-lag are shown in Supplementary Fig. 8). The autocorrelation drop between the zero and first time-lag ($t = 2$ ms) reflects the difference between the total variance of spike counts and the variance of instantaneous rate according to the law of total variance for a doubly stochastic process[37]. This drop is fitted by the aABC algorithm when estimating the timescales.

Estimating timescales with adaptive Approximate Bayesian Computations. We estimated the autocorrelation timescales using the aABC method that overcomes the statistical bias in empirical autocorrelations and provides the posterior distributions of unbiased estimated timescales[37]. The width of inferred posteriors indicates the uncertainty of estimates. For more reliable estimates of timescales (i.e. narrower posteriors), we selected epochs of experiments with longer trial durations (brown and yellow frames in Supplementary Fig. 1).

The aABC method estimates timescales by fitting the spike-count autocorrelation with a generative model. We used a generative model based on a doubly stochastic process with one or two timescales. Spike-counts were generated from a rate governed by a linear mixture of OrnsteinUhlenbeck (OU) processes (one OU process A_{τ_k} for each timescale τ_k)

$$A_{OU}(t') = \sum_{k=1}^{n} \sqrt{c_k} A_{\tau_k}(t'), \quad \sum_{k=1}^{n} c_k = 1, \quad c_k \in [0,1], \tag{2}$$

where n is the number of timescales and c_k are their weights. The aABC algorithm optimizes the model parameters to match the spike-count autocorrelations between V4 data and synthetic data generated from the model. We generated synthetic data with the same number of trials, trial duration, mean and variance of spike counts as in the experimental data. By matching these statistics, the empirical autocorrelations of the synthetic and experimental data are affected by the same statistical bias when their shapes match. Therefore, the timescales of the fitted generative model represent the unbiased estimate of timescales in the neural data.

The spike-counts s are sampled for each time-bin $[t'_i, t'_{i+1}]$ from a distribution $p_{\text{count}}(s|\lambda(t'_i))$, where $\lambda(t'_i) = A_{OU}(t'_i)\Delta t'$ is the mean spike-count and $\Delta t' = t'_{i+1} - t'_i$ is the bin size. To capture the possible non-Poisson statistics of the recorded neurons, we introduce a dispersion parameter α defined as the variance over mean ratio of the spike-counts distribution $\alpha = \sigma^2_{s|\lambda(t'_i)}/\lambda(t'_i)$. For a Poisson distribution α is equal to 1. We allow for non-Poisson statistics by sampling the spike counts from a gamma distribution and optimize the value of α together with the timescales and the weights.

On each iteration of the aABC algorithm, we draw sample parameters from a prior distribution (first iteration) or a proposal distribution (subsequent iterations) defined based on the prior distribution and parameters accepted on the previous iteration. Then, we generate synthetic data from the sampled parameters and compute the distance d between the autocorrelations of synthetic and experimental data:

$$d(t_m) = \frac{1}{m} \sum_{j=0}^{m} \left(AC_{\text{experimental}}(t_j) - AC_{\text{synthetic}}(t_j) \right)^2, \tag{3}$$

where t_m is the maximum time-lag considered in computing the distance. We set t_m to 100 ms to avoid over-fitting the noise in the tail of the autocorrelations. If the distance is smaller than a predefined error threshold ε, the sample parameters are accepted and added to the posterior distribution. Each iteration continued until 100 sample-parameters were accepted. The initial error threshold was set to $\varepsilon_0 = 0.1$, and in subsequent iterations, the error threshold was updated to the first quartile of the distances for the accepted samples. The fraction of accepted samples out of all drawn parameter samples is recorded as the acceptance rate $accR$. The algorithm stops when the acceptance rate reaches $accR < 0.0007$. The final accepted samples are considered as an approximation for the posterior distribution. We

computed the MAP estimates by smoothing the final joint posterior distribution with a multivariate Gaussian kernel and finding its maximum with a grid search.

We used a multivariate uniform prior distribution over all parameters. For the two-timescale generative model (M_2), the priors' ranges were set to

$$\tau_1 : U[0, 60], \quad \tau_2 : U[0, 400], \quad c_1 : U[0, 1], \quad \alpha : U[0.7, 1.3], \tag{4}$$

and for the one-timescale generative model (M_1) they were set to

$$\tau : U[0, 400], \quad \alpha : U[0.7, 1.3]. \tag{5}$$

Model comparison with adaptive Approximate Bayesian Computations. We used the inferred posteriors from the aABC fit to determine whether the V4 data autocorrelations were better described with the one-timescale (M_1) or the two-timescale (M_2) generative models[37]. First, we measured the goodness of fit for each model based on the distribution of distances between the autocorrelation of synthetic data from the generative model and the autocorrelation of V4 data. We approximated the distributions of distances by generating 1000 realizations of synthetic data from each model with parameters drawn from the posterior distributions and computing the distance for each realization. If the distributions of distances were significantly different (Wilcoxon ranksum test), we approximated the Bayes factor, otherwise the summary statistics were not sufficient to distinguish these two models[58].

Bayes factor is the ratio of marginal likelihoods of the two models and takes into account the number of parameters in each model[59]. In the aABC method, the ratio between the acceptance rates of two models for a given error threshold ε approximates the Bayes factor (BF) for that error threshold[37]:

$$\text{BF}(\varepsilon) = \frac{accR_{M_2}(\varepsilon)}{accR_{M_1}(\varepsilon)}. \tag{6}$$

Acceptance rates can be computed using the cumulative distribution function (CDF) of the distances for a given error threshold ε,

$$\text{CDF}_{M_i}(\varepsilon) = p_{M_i}(d < \varepsilon) = accR_{M_i}(\varepsilon), \quad i = 1, 2, \tag{7}$$

where $p_{M_i}(d)$ is the probability distribution of distances for the model M_i. Thus, the ratio between the CDF of distances approximates the Bayes factor for every chosen error threshold. To eliminate the dependence on a specific error threshold, we computed the acceptance rates and the Bayes factor for varying error thresholds. Since only small errors indicate a well-fitted model, we computed the Bayes factor for all error thresholds that were smaller than the largest median of distance distributions of two models.

The M_2 model was selected if its distances were significantly smaller than for the M_1 model (Wilcoxon ranksum test) and $\text{CDF}_{M_2}(\varepsilon) > \text{CDF}_{M_1}(\varepsilon)$, i.e. $\text{BF} > 1$, for all $\varepsilon < \max_{M_1,M_2}[\text{median}(\varepsilon)]$ (Supplementary Fig. 3). The same procedure was applied for selecting the M_1 model. Although the Bayes factor threshold was set at 1, in most cases we obtained $\text{BF} \gg 1$, indicating strong evidence for the two-timescale model. If the distribution of distances for the two models were not significantly different or the condition for the ratio between CDFs did not hold for all selected ε (CDFs were crossing), we classified the outcome as inconclusive, meaning that data statistics were not sufficient to make the comparison.

Recurrent network model with spatially structured connections. The network model operates on a two-dimensional square lattice of size 100×100 with periodic boundary conditions. Each unit in the model is connected to 8 other units taken either from its direct Moore neighborhood (local connectivity, Fig. 5a, top) or randomly selected within the connectivity radius r (dispersed connectivity, Fig. 5a, bottom). Activity of each unit is represented by a binary state variable $S_i \in \{0, 1\}$ ($i = 1 \ldots N$, where $N = 10^4$ is the number of units). The units act as probabilistic integrate-and-fire units[60] following linear or non-linear integration rules. States of the units are updated in discrete time-steps t' based on a self-excitation probability (p_s), probability of excitation by the connected units (p_r), and the probability of external excitation ($p_\mathrm{ext} \ll 1$). The transition probabilities for each unit S_i at time-step t' are either governed by additive interaction rules (linear model):

$$\begin{aligned} p(S_i = 0 \to 1) &= p_\mathrm{ext} + p_\mathrm{r} \sum_j S_j, \\ p(S_i = 1 \to 0) &= 1 - \left(p_\mathrm{ext} + p_\mathrm{s} + p_\mathrm{r} \sum_j S_j \right), \end{aligned} \tag{8}$$

or multiplicative interaction rules (non-linear model):

$$\begin{aligned} p(S_i = 0 \to 1) &= 1 - (1 - p_\mathrm{ext})(1 - p_\mathrm{r})^{\sum_j S_j}, \\ p(S_i = 1 \to 0) &= (1 - p_\mathrm{ext})(1 - p_\mathrm{s})(1 - p_\mathrm{r})^{\sum_j S_j}. \end{aligned} \tag{9}$$

Here, $\sum_j S_j$ indicates the number of active neighbors of unit S_i at time-step t'. For the analysis in the main text, we used the linear model. The non-linear model generates similar local temporal dynamics (Supplementary Fig. 6). In the linear model, the sum of connection probabilities $\mathrm{BP} = p_\mathrm{s} + 8p_\mathrm{r}$ is the branching parameter that defines the state of the dynamics relative to a critical point at $\mathrm{BP} = 1$[60].

To compute the average local autocorrelation in the network, we simulated the model for 10^5 time-steps and averaged the autocorrelations of individual units. The global autocorrelations were computed from the pooled activity of all units in the network. To compute the autocorrelation of horizontal inputs for a unit i, we simulated the network with an additional "shadow" unit, which was activated by the same horizontal inputs (p_r) as the unit i but without the inputs p_s and p_ext. The shadow unit did not activate other units in the network. The autocorrelation of horizontal inputs was computed from the shadow unit activity. Each simulation started with a random configuration of active units based on the analytically computed steady-state mean activity (Eq. 15). Running simulations for long periods allowed us to avoid the statistical bias in the model autocorrelations. We set $p_\mathrm{ext} = 10^{-4}$, but the strength of external input in the linear model does not affect the autocorrelation timescales.

Analytical derivation of local timescales in the network model. For analytical derivations, we approximated the linear probabilistic network model (Eq. 8) by a continuous-time rate model with the transition rates defined as

$$\begin{aligned} w(S_i = 0 \to 1) &= \alpha_1 + \beta_1 \sum_j S_j, \\ w(S_i = 1 \to 0) &= \alpha_2 - \beta_2 \sum_j S_j. \end{aligned} \tag{10}$$

These equations contain two non-interaction terms $\alpha_1 = p_\mathrm{ext}/\Delta t'$ and $\alpha_2 = (1 - p_\mathrm{ext} - p_\mathrm{s})/\Delta t'$, and two interaction terms $\beta_1 = \beta_2 = p_\mathrm{r}/\Delta t'$, where $\Delta t' = 1$ ms is the duration of each time step. For this

model, the probability of units to stay in a certain configuration $\{S\} = \{S_1, S_2, ..., S_N\}$ at time t' is denoted as $P(\{S\}, t')$. The master equation describing the time evolution of $P(\{S\}, t')$ is given by[38]:

$$\frac{d}{dt'}P(\{S\}, t') = -P(\{S\}, t') \sum_i w(S_i) + \sum_i P(\{S\}^{i*}, t')w(1 - S_i) \, , \tag{11}$$

where $\{S\}^{i*} = \{S_1, S_2, ..., 1-S_i, ..., S_N\}$. Using the master equation, we can write the time evolution for the first and second moments as

$$\frac{d}{dt'}\langle S_i \rangle(t) = \sum_{\{S\}} P(\{S\}, t')[w(S_i)(1 - 2S_i)] \, , \tag{12}$$

$$\frac{d}{dt'}\langle S_i S_j \rangle(t') = \sum_{\{S\}} P(\{S\}, t')[w(S_i)(1 - 2S_i)S_j + w(S_j)(1 - 2S_j)S_i], \tag{13}$$

and for the time-delayed quadratic moment at time-lag t as

$$\frac{d}{dt}\langle S_i(t')S_j(t' + t)\rangle = \langle S_i(t')(1 - 2S_j(t' + t))w(S_j(t' + t))\rangle. \tag{14}$$

By setting the right side of Eq. 12 to zero, we can compute the steady-state mean activity

$$\langle S \rangle = \frac{\alpha_1}{\alpha_1 + \alpha_2 - n\beta_1} = \frac{p_{\text{ext}}}{1 - (p_{\text{s}} + 8p_{\text{r}})}, \tag{15}$$

where $n = 8$ is the number of incoming connections to each unit.

We compute the timescales analytically for the network with local connections ($r = 1$). From Eq. 14, we can derive the equation for the average autocorrelation of each unit $AC(t)$ as

$$\frac{1}{\alpha_1 + \alpha_2}\frac{d}{dt}AC(t) = -AC(t) + \frac{\beta_1}{\alpha_1 + \alpha_2}\sum_{\mathbf{x}} CC(\mathbf{x}, t) \, . \tag{16}$$

Here $CC(\mathbf{x}, t)$ is the cross-correlation between each unit at location (i, j) and its 8 nearest neighbors $\mathbf{x} = (i \pm 1, j \pm 1)$. The cross-correlation term in this equation gives rise to the interaction timescales in the autocorrelation. By neglecting the cross-correlation term, we can solve the Eq. 16 to get the self-excitation timescale

$$\tau_{\text{self}} = \frac{1}{\alpha_1 + \alpha_2} = \frac{\Delta t'}{1 - p_{\text{s}}} \, . \tag{17}$$

Solving the dynamical equation for the time-delayed cross-correlation (Eq. 14) in the Fourier domain gives the interaction timescales (see Supplementary Note 1 for details):

$$\begin{aligned} \tau_{\text{int},\mathbf{k}}(\mathbf{k} = (k_1, k_2)) &= \frac{\tau_{\text{self}}}{1 - \frac{n}{4}\frac{\beta_1}{\alpha_1+\alpha_2}[\cos(k_1) + \cos(k_2) + 2\cos(k_1)\cos(k_2)]} \\ &= \frac{1}{1 - p_{\text{s}} - 2p_{\text{r}}[\cos(k_1) + \cos(k_2) + 2\cos(k_1)\cos(k_2)]} \, , \end{aligned} \tag{18}$$

where $\mathbf{k} = (k_1, k_2)$ are the spatial frequencies in the Fourier space. For each $\mathbf{k}$ we get a different interaction timescale. Smaller $\mathbf{k}$ (low spatial frequencies) correspond to interactions on larger spatial scales, whereas larger $\mathbf{k}$ (high spatial frequencies) correspond to interactions on more local spatial

scales. The largest interaction timescale (the global timescale) is defined based on the zero spatial frequency mode:

$$\tau_{\text{global}} = \tau_{\text{int},\mathbf{k}}(\mathbf{k} = (0,0)) = \frac{1}{\alpha_1 + \alpha_2 - n\beta_1} = \frac{\Delta t'}{1 - p_{\text{s}} - 8p_{\text{r}}} \,. \tag{19}$$

In these derivations, we defined distances between units as euclidean distances and discarded the contributions from third and higher moments.

In the spatial frequency domain, we found that the weight of different spatial frequencies is a continuous function of spatial frequency magnitude $|\mathbf{k}|$. The shallow decay of autocorrelation results from summation of contributions from different spatial frequency modes. Since the spatial correlations in the network can be approximately captured with a single exponential decay function characterized by a correlation length-scale ξ, the spatial frequency mode with $k_1 = k_2 = 1/\xi$ approximately makes the largest contribution (in the continuum limit, the weight of this mode reaches the peak value). Therefore, we can approximate the shallow decay of the autocorrelation with a single dominant interaction timescale, that corresponds to the dominant spatial frequency $k_1 = k_2 = 1/\xi$. The contribution of interaction timescale created by zero spatial frequency is negligible in local autocorrelations, so we neglect its contribution (see Supplementary Note 1 for details).

Substituting the dominant spatial frequency in Eq. 18, we can write the dominant interaction timescale as

$$\tau_{\text{int}} = \frac{\tau_{\text{self}}}{1 - \frac{n}{2}\frac{\beta_1}{\alpha_1+\alpha_2}[\cos(1/\xi) + \cos(1/\xi)^2]} = \frac{\Delta t'}{1 - p_{\text{s}} - 4p_{\text{r}}[\cos(1/\xi) + \cos(1/\xi)^2]} \,, \tag{20}$$

where $\xi = \sqrt{\frac{4p_{\text{r}}}{1-p_{\text{s}}-8p_{\text{r}}}}$ is the approximation for the length-scale of spatial correlations. The dominant interaction timescale corresponds to the slow timescale in the V4 data autocorrelation. The analytical approximation of the dominant interaction timescale is more accurate when the dynamics are away from the critical point. Close to the critical point (BP $\rightarrow$ 1), the dominant spatial frequency cannot be captured with the approximated correlation length-scale due to the presence of higher-order correlations. In this case, we estimate the dominant interaction timescale from simulations by fitting the shape of autocorrelations with a double exponential function (Eq. 24).

A more accurate estimation of the self-excitation timescale for the discrete time network model can be obtained using the autocorrelation of a 2-state Markov process (2SMP) driven by the self-excitation and external input. Using the transition matrix (considering the linear model)

$$\mathbb{P} = \begin{bmatrix} 1 - p_{\text{ext}} & p_{\text{ext}} \\ 1 - (p_{\text{s}} + p_{\text{ext}}) & p_{\text{s}} + p_{\text{ext}} \end{bmatrix} , \tag{21}$$

we can compute the autocorrelation of the Markov process at time-lag t (Supplementary Note 3):

$$AC_{\text{2SMP}}(t) = p_{\text{s}}^t. \tag{22}$$

The decay timescale of this autocorrelation is equivalent to the self-excitation timescale in the network model

$$\tau_{\text{self}} = -(\log(p_{\text{s}}))^{-1}, \tag{23}$$

which for large p_{s} and $\Delta t' = 1$ is equivalent to Eq. 17.

Matching the timescales of the network model to neural data. To match the timescales between the model and V4 data, we used the activity autocorrelation of one unit in the network model with local connections ($r = 1$). We searched for model parameters such that the model timescales fell within the range of timescales observed in the V4 activity, which was the mean $\pm$ s.e.m of the MAP timescale-estimates across recording sessions. We computed the range for the fast timescales from the pooled attend-in and attend-away conditions, since they were not significantly different: $\tau_{1,\text{att-away}} = \tau_{1,\text{att-in}} = 4.74 \pm 0.42$ ms. We used this range for the fast timescale in both the attend-in and attend-away conditions. For the slow timescales, we computed the ranges separately for the attend-in (averaged over covert and overt) and attend-away conditions: $\tau_{2,\text{att-away}} = 117.09 \pm 10.58$ ms, $\tau_{1,\text{att-in}} = 140.97 \pm 11.51$ ms.

We fitted the self-excitation and dominant interaction timescales obtained from the autocorrelation of an individual unit's activity in the model to the fast and slow timescales of V4 data estimated from the aABC method. Using Eq. 23 and Eq. 20, we found an approximate range of parameters p_s and p_r that reproduce V4 timescales. Then, we performed a grid search within this parameter range to identify the model timescales falling within the range of V4 timescales during attend-away and attend-in conditions. We used model simulations for grid search since the analytical results for dominant timescale are approximate. We used very long model simulations (10^5 time-steps) to obtain unbiased autocorrelations and then estimated the model timescales by fitting a double exponential function

$$AC(t) = c_1 e^{-t/\tau_1} + (1 - c_1) e^{-t/\tau_2}, \tag{24}$$

directly to the empirical autocorrelations. We fitted the exponential function up to the time-lag $t_m = 100$ ms, the same as used for fitting the neural data autocorrelations with the aABC method.

Data availability

The data are available from the corresponding authors upon request.

Code availability

Codes for the timescale estimation and Bayesian model comparison with the aABC method are available as a Python package at: https://github.com/roxana-zeraati/abcTau. Codes for simulating the network model are available at: https://github.com/roxana-zeraati/spatial-network.

References

1. Kiebel, S. J., Daunizeau, J. & Friston, K. J. A Hierarchy of Time-Scales and the Brain. *PLOS Computational Biology* **4**, e1000209 (2008). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000209. Publisher: Public Library of Science.

2. Wiltschko, A. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–1135 (2015). URL http://www.sciencedirect.com/science/article/pii/S0896627315010375.

3. Berman, G. J., Bialek, W. & Shaevitz, J. W. Predictability and hierarchy in Drosophila behavior. *Proceedings of the National Academy of Sciences* **113**, 11943–11948 (2016). URL https://www.pnas.org/content/113/42/11943. Publisher: National Academy of Sciences Section: Biological Sciences.

4. Uchida, N. & Mainen, Z. F. Speed and accuracy of olfactory discrimination in the rat. *Nature Neuroscience* **6**, 1224–1229 (2003). URL https://www.nature.com/articles/nn1142. Number: 11 Publisher: Nature Publishing Group.

5. Buracas, G. T., Zador, A. M., DeWeese, M. R. & Albright, T. D. Efficient Discrimination of Temporal Patterns by Motion-Sensitive Neurons in Primate Visual Cortex. *Neuron* **20**, 959–969 (1998). URL http://www.sciencedirect.com/science/article/pii/S0896627300804778.

6. Yang, Y., DeWeese, M., Otazu, G. & Zador, A. Millisecond-scale differences in neural activity in auditory cortex can drive decisions. *Nature Precedings* 1–1 (2008). URL https://www.nature.com/articles/npre.2008.2280.1. Publisher: Nature Publishing Group.

7. Bathellier, B., Buhl, D. L., Accolla, R. & Carleton, A. Dynamic Ensemble Odor Coding in the Mammalian Olfactory Bulb: Sensory Information at Different Timescales. *Neuron* **57**, 586–598 (2008). URL http://www.sciencedirect.com/science/article/pii/S0896627308001347.

8. Jonides, J. *et al.* The Mind and Brain of Short-Term Memory. *Annual Review of Psychology* **59**, 193–224 (2008). URL https://doi.org/10.1146/annurev.psych.59.103006.

093615. _eprint: https://doi.org/10.1146/annurev.psych.59.103006.093615.

9. Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364** (2019). URL https://science.sciencemag.org/content/364/6441/eaav8911. Publisher: American Association for the Advancement of Science Section: Research Article.

10. Shadlen, M. N. & Newsome, W. T. Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology* **86**, 1916–1936 (2001). URL https://journals.physiology.org/doi/full/10.1152/jn.2001.86.4.1916. Publisher: American Physiological Society.

11. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience* **17**, 1661–1663 (2014). URL https://www.nature.com/articles/nn.3862/. Number: 12 Publisher: Nature Publishing Group.

12. Gao, R., van den Brink, R. L., Pfeffer, T. & Voytek, B. Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife* **9**, e61277 (2020). URL https://doi.org/10.7554/eLife.61277. Publisher: eLife Sciences Publications, Ltd.

13. Raut, R. V., Snyder, A. Z. & Raichle, M. E. Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proceedings of the National Academy of Sciences* **117**, 20890–20897 (2020). URL https://www.pnas.org/content/117/34/20890. Publisher: National Academy of Sciences Section: Biological Sciences.

14. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017). URL https://www.nature.com/articles/nature23020. Number: 7665 Publisher: Nature Publishing Group.

15. Honey, C. *et al.* Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron* **76**, 423–434 (2012). URL http://www.sciencedirect.com/science/article/pii/S0896627312007179.

16. Bernacchia, A., Seo, H., Lee, D. & Wang, X.-J. A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience* **14**, 366–372 (2011). URL https://www.nature.com/articles/nn.2752. Number: 3 Publisher: Nature Publishing Group.

17. Spitmaan, M., Seo, H., Lee, D. & Soltani, A. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proceedings of the National Academy of Sciences* **117**, 22522–22531 (2020). URL https://www.pnas.org/content/117/36/22522. Publisher: National Academy of Sciences Section: Biological Sciences.

18. Fallon, J. *et al.* Timescales of spontaneous fMRI fluctuations relate to structural connectivity in the brain. *Network Neuroscience* **4**, 788–806 (2020). URL https://doi.org/10.1162/netn_a_00151. Publisher: MIT Press.

19. Cavanagh, S. E., Hunt, L. T. & Kennerley, S. W. A Diversity of Intrinsic Timescales

Underlie Neural Computations. *Frontiers in Neural Circuits* **14** (2020). URL https://www.frontiersin.org/articles/10.3389/fncir.2020.615626/full?field=&id=615626&journalName=Frontiers_in_Neural_Circuits. Publisher: Frontiers.

20. Wang, X.-J. Macroscopic gradients of synaptic excitation and inhibition in the neocortex. *Nature Reviews Neuroscience* **21**, 169–178 (2020). URL https://www.nature.com/articles/s41583-020-0262-x. Number: 3 Publisher: Nature Publishing Group.

21. Huntenburg, J. M., Bazin, P.-L. & Margulies, D. S. Large-Scale Gradients in Human Cortical Organization. *Trends in Cognitive Sciences* **22**, 21–31 (2018). URL http://www.sciencedirect.com/science/article/pii/S1364661317302401.

22. Elston, G. N. 4.13 - Specialization of the Neocortical Pyramidal Cell during Primate Evolution. In Kaas, J. H. (ed.) *Evolution of Nervous Systems*, 191–242 (Academic Press, Oxford, 2007). URL http://www.sciencedirect.com/science/article/pii/B0123708788001646.

23. Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H. & Wang, X.-J. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron* **88**, 419–431 (2015). URL http://www.sciencedirect.com/science/article/pii/S0896627315007655.

24. Glasser, M. F. & Essen, D. C. V. Mapping Human Cortical Areas In Vivo Based on Myelin Content as Revealed by T1- and T2-Weighted MRI. *Journal of Neuroscience* **31**, 11597–11616 (2011). URL https://www.jneurosci.org/content/31/32/11597. Publisher: Society for Neuroscience Section: Articles.

25. Burt, J. B. *et al.* Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nature Neuroscience* **21**, 1251–1259 (2018). URL https://www.nature.com/articles/s41593-018-0195-0. Number: 9 Publisher: Nature Publishing Group.

26. Hart, E. & Huk, A. C. Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network. *eLife* **9**, e52460 (2020). URL https://doi.org/10.7554/eLife.52460. Publisher: eLife Sciences Publications, Ltd.

27. Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K. & Stokes, M. G. Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nature Communications* **9**, 3499 (2018). URL https://www.nature.com/articles/s41467-018-05961-4. Number: 1 Publisher: Nature Publishing Group.

28. Safavi, S. *et al.* Nonmonotonic spatial structure of interneuronal correlations in prefrontal microcircuits. *Proceedings of the National Academy of Sciences* **115**, E3539–E3548 (2018). URL https://www.pnas.org/content/115/15/E3539. Publisher: National Academy of Sciences Section: PNAS Plus.

29. Demirta, M. *et al.* Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics. *Neuron* **101**, 1181–1194.e13 (2019). URL http://www.sciencedirect.com/science/article/pii/S0896627319300443.

30. Litwin-Kumar, A. & Doiron, B. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience* **15**, 1498–1505 (2012). URL https://www.nature.com/articles/nn.3220. Number: 11 Publisher: Nature Publishing Group.

31. Chaudhuri, R., Bernacchia, A. & Wang, X.-J. A diversity of localized timescales in network activity. *eLife* **3**, e01239 (2014). URL https://doi.org/10.7554/eLife.01239. Publisher: eLife Sciences Publications, Ltd.

32. Buxhoeveden, D. P. & Casanova, M. F. The minicolumn hypothesis in neuroscience. *Brain* **125**, 935–951 (2002). URL https://academic.oup.com/brain/article/125/5/935/328135. Publisher: Oxford Academic.

33. Mountcastle, V. B. The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997). URL https://academic.oup.com/brain/article/120/4/701/372118.

34. Engel, T. A. *et al.* Selective modulation of cortical state during spatial attention. *Science* **354**, 1140–1144 (2016). URL https://science.sciencemag.org/content/354/6316/1140. Publisher: American Association for the Advancement of Science Section: Report.

35. Steinmetz, N. & Moore, T. Eye Movement Preparation Modulates Neuronal Responses in Area V4 When Dissociated from Attentional Demands. *Neuron* **83**, 496–506 (2014). URL http://www.sciencedirect.com/science/article/pii/S0896627314005364.

36. van Kempen, J. *et al.* Top-down coordination of local cortical state during selective attention. *Neuron* (2021). URL http://www.sciencedirect.com/science/article/pii/S0896627320309958.

37. Zeraati, R., Engel, T. A. & Levina, A. Estimation of autocorrelation timescales with Approximate Bayesian Computations. *bioRxiv* 2020.08.11.245944 (2020). URL https://www.biorxiv.org/content/10.1101/2020.08.11.245944v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

38. Ginzburg, I. & Sompolinsky, H. Theory of correlations in stochastic neural networks. *Physical Review E* **50**, 3171–3191 (1994). URL https://link.aps.org/doi/10.1103/PhysRevE.50.3171. Publisher: American Physical Society.

39. Smith, M. A. & Sommer, M. A. Spatial and Temporal Scales of Neuronal Correlation in Visual Area V4. *Journal of Neuroscience* **33**, 5422–5432 (2013). URL https://www.jneurosci.org/content/33/12/5422. Publisher: Society for Neuroscience Section: Articles.

40. Wilting, J. & Priesemann, V. Inferring collective dynamical states from widely unobserved systems. *Nature Communications* **9**, 2325 (2018). URL https://www.nature.com/

articles/s41467-018-04725-4. Number: 1 Publisher: Nature Publishing Group.

41. Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M. & Miller, K. D. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron* **98**, 846–860.e5 (2018). URL http://www.sciencedirect.com/science/article/pii/S0896627318303258.

42. Anderson, J. C., Kennedy, H. & Martin, K. A. C. Pathways of Attention: Synaptic Relationships of Frontal Eye Field to V4, Lateral Intraparietal Cortex, and Area 46 in Macaque Monkey. *Journal of Neuroscience* **31**, 10872–10881 (2011). URL https://www.jneurosci.org/content/31/30/10872. Publisher: Society for Neuroscience Section: Articles.

43. Ferro, D., van Kempen, J., Boyd, M., Panzeri, S. & Thiele, A. Directed information exchange between cortical layers in macaque V1 and V4 and its modulation by selective attention. *Proceedings of the National Academy of Sciences* **118**, e2022097118 (2021). URL http://www.pnas.org/lookup/doi/10.1073/pnas.2022097118.

44. Thiele, A. & Bellgrove, M. A. Neuromodulation of Attention. *Neuron* **97**, 769–785 (2018). URL https://www.sciencedirect.com/science/article/pii/S0896627318300114.

45. Shi, Y.-L., Steinmetz, N. A., Moore, T., Boahen, K. & Engel, T. A. Influence of On-Off dynamics and selective attention on the spatial pattern of correlated variability in neocortex. preprint, Neuroscience (2020). URL http://biorxiv.org/lookup/doi/10.1101/2020.09.02.279893.

46. Beiran, M. & Ostojic, S. Contrasting the effects of adaptation and synaptic filtering on the timescales of dynamics in recurrent networks. *PLOS Computational Biology* **15**, e1006893 (2019). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006893. Publisher: Public Library of Science.

47. Huang, C. *et al.* Circuit Models of Low-Dimensional Shared Variability in Cortical Networks. *Neuron* **101**, 337–348.e4 (2019). URL http://www.sciencedirect.com/science/article/pii/S0896627318310432.

48. He, B. J., Snyder, A. Z., Zempel, J. M., Smyth, M. D. & Raichle, M. E. Electrophysiological correlates of the brain's intrinsic large-scale functional architecture. *Proceedings of the National Academy of Sciences* **105**, 16039–16044 (2008). URL https://www.pnas.org/content/105/41/16039. Publisher: National Academy of Sciences Section: Biological Sciences.

49. Okun, M., Steinmetz, N. A., Lak, A., Dervinis, M. & Harris, K. D. Distinct Structure of Cortical Population Activity on Fast and Infraslow Timescales. *Cerebral Cortex* **29**, 2196–2210 (2019). URL https://doi.org/10.1093/cercor/bhz023.

50. Cavanagh, S. E., Wallis, J. D., Kennerley, S. W. & Hunt, L. T. Autocorrelation structure at

rest predicts value correlates of single neurons during reward-guided choice. *eLife* **5**, e18937 (2016). URL https://doi.org/10.7554/eLife.18937. Publisher: eLife Sciences Publications, Ltd.

51. Kim, R. & Sejnowski, T. J. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nature Neuroscience* **24**, 129–139 (2021). URL https://www.nature.com/articles/s41593-020-00753-w. Number: 1 Publisher: Nature Publishing Group.

52. Tomen, N., Rotermund, D. & Ernst, U. Marginally subcritical dynamics explain enhanced stimulus discriminability under attention. *Frontiers in Systems Neuroscience* **8** (2014). URL https://www.frontiersin.org/articles/10.3389/fnsys.2014.00151/full. Publisher: Frontiers.

53. Muoz, M. A. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics* **90**, 031001 (2018). URL https://link.aps.org/doi/10.1103/RevModPhys.90.031001. Publisher: American Physical Society.

54. Rockland, K. S., Saleem, K. S. & Tanaka, K. Divergent feedback connections from areas V4 and TEO in the macaque. *Visual Neuroscience* **11**, 579–600 (1994). URL https://www.cambridge.org/core/journals/visual-neuroscience/article/abs/divergent-feedback-connections-from-areas-v4-and-teo-in-the-macaque/9F954B564C8C1406793B1246F3B251E4. Publisher: Cambridge University Press.

55. Shou, T.-D. The functional roles of feedback projections in the visual system. *Neuroscience Bulletin* **26**, 401–410 (2010). URL https://doi.org/10.1007/s12264-010-0521-3.

56. Gjorgjieva, J., Drion, G. & Marder, E. Computational implications of biophysical diversity and multiple timescales in neurons and synapses for circuit performance. *Current Opinion in Neurobiology* **37**, 44–52 (2016). URL http://www.sciencedirect.com/science/article/pii/S0959438815001865.

57. Gieselmann, M. A. & Thiele, A. Comparison of spatial integration and surround suppression characteristics in spiking activity and the local field potential in macaque V1. *European Journal of Neuroscience* **28**, 447–459 (2008). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2008.06358.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2008.06358.x.

58. Marin, J.-M., Pillai, N. S., Robert, C. P. & Rousseau, J. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 833–859 (2014). URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12056. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12056.

59. Bishop, C. M. *Pattern Recognition and Machine Learning: All "just the Facts 101" Material* (Springer (India) Private Limited, 2013). Google-Books-ID: HL4HrgEACAAJ.

60. Larremore, D. B., Shew, W. L., Ott, E., Sorrentino, F. & Restrepo, J. G. Inhibition Causes Ceaseless Dynamics in Networks of Excitable Nodes. *Physical Review Letters* **112**, 138103 (2014). URL https://link.aps.org/doi/10.1103/PhysRevLett.112.138103. Publisher: American Physical Society.

Acknowledgements

This work was supported by a Sofja Kovalevskaja Award from the Alexander von Humboldt Foundation, endowed by the Federal Ministry of Education and Research (R.Z., A.L.), SMARTSTART2 program provided by Bernstein Center for Computational Neuroscience and Volkswagen Foundation (R.Z.), the NIH grant R01 EB026949 (T.A.E.), the Swartz Foundation (Y.S.), the Pershing Square Foundation (T.A.E.), the Sloan Research Fellowship (Y.S.,T.A.E.), the NIH grant EY014924 (T.M.), the MRC grant MR/P013031/1 (M.A.G., A.T.). The authors acknowledge the support from the BMBF through the Tübingen AI Center (FKZ: 01IS18039B) and the International Max Planck Research School for the Mechanisms of Mental Function and Dysfunction (IMPRS-MMFD).

Author Contributions

R.Z., A.L., and T.A.E. designed the study. N.A.S., M.A.G, A.T., and T.M. designed the experiments. N.A.S. and M.A.G performed the experiments and spike sorting. R.Z., Y.S., A.L., and T.A.E. developed the analysis methods and mathematical models. R.Z. analyzed the data and performed model simulations. Y.L. performed the analytical calculations for the network model. R.Z., Y.S., N.A.S., M.A.G, A.T., T.M., A.L., and T.A.E. discussed the findings and wrote the paper.

Supplementary Information

Supplementary Figures 1–8
Supplementary Notes 1–3

Competing interests

The authors declare no competing interests.