nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-020-0644-4

Check for updates

Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory

Joao Barbosa[1,8], Heike Stein[1,8], Rebecca L. Martinez[1], Adrià Galan-Gadea[1], Sihai Li[2], Josep Dalmau[1,3,4,5,6], Kirsten C. S. Adam[7], Josep Valls-Solé[1], Christos Constantinidis[2] and Albert Compte[1] ✉

Persistent neuronal spiking has long been considered the mechanism underlying working memory, but recent proposals argue for alternative 'activity-silent' substrates. Using monkey and human electrophysiology data, we show here that attractor dynamics that control neural spiking during mnemonic periods interact with activity-silent mechanisms in the prefrontal cortex (PFC). This interaction allows memory reactivations, which enhance serial biases in spatial working memory. Stimulus information was not decodable between trials, but remained present in activity-silent traces inferred from spiking synchrony in the PFC. Just before the new stimulus, this latent trace was reignited into activity that recapitulated the previous stimulus representation. Importantly, the reactivation strength correlated with the strength of serial biases in both monkeys and humans, as predicted by a computational model that integrates activity-based and activity-silent mechanisms. Finally, single-pulse transcranial magnetic stimulation applied to the human PFC between successive trials enhanced serial biases, thus demonstrating the causal role of prefrontal reactivations in determining working-memory behavior.

The mechanisms by which information is maintained in working memory are still not fully understood. Ample evidence supports a role for sustained neural activity in prefrontal[1–3] and other cortices[4,5], possibly supported by attractor dynamics in recurrently connected circuits[6,7]. However, recent studies have argued that memories may be maintained without persistent firing-rate tuning during memory periods[8]. This 'activity-silent' memory can be mediated by slowly decaying intrinsic or synaptic mechanisms, such as short-term synaptic plasticity[9,10], or by activity-dependent intrinsic mechanisms with a long time constant[11–13] that could allow the reactivation of memories from latent storage. This computational proposal has received support from neuroimaging studies, whereby in some working memory tasks, despite good memory performance, stimulus information cannot be retrieved from neural delay activity, but later robustly reappears[14] during comparison or response periods (but see also ref. [15]).

The apparent incompatibility between activity-based and activity-silent memory maintenance has led to viewing them as exclusive alternatives[8]. However, modeling implementations of activity-silent conditions invariably require the network to be configured close to the same attractor regime[9] that enables persistent activity. This attractor nonlinearity is necessary to increase the signal-to-noise ratio of the fading subthreshold signal for successful memory reactivation[9]. At the same time, activity-silent memory mechanisms may stabilize persistent activity in attractor networks (for examples, see refs. [11,16–18]). Interestingly, modeling studies have argued that the interaction of these mechanisms during the delay period would be reflected behaviorally in serial biases[11,16], but this theoretically appealing hypothesis still lacks experimental support.

Serial biases in spatial working memory denote small but systematic shifts of memory reports toward nearby locations memorized in the previous trial[19–22], which reveal a lingering representation of previous memories. Uncleared memory remnants have long been viewed as limiting working memory performance (proactive interference[23]), but recent proposals suggest that they may be useful to inform working memory about the expected statistics in naturalistic conditions[24] (but see [25]), similar to other history biases with longer time scales and possibly different neural mechanisms (contraction bias[26–28]). The functional relevance of biases implicates specific roles of higher-order brain areas. On the one hand, these areas could suppress maladaptive biases to minimize performance degradation[29,30]. On the other hand, they might promote adaptive biases by maintaining a representation of stimulus history[26]. Whether association areas generate or suppress serial biases in primates is currently undefined, and a mechanistic understanding of the generation of any type of history biases is still lacking.

Both attractor dynamics[20] and activity-silent[11,16,31] mechanisms have been proposed to carry stimulus-selective information from one trial to the next to effect serial biases. However, dependencies of serial biases on inter-trial interval (ITI) durations[20–22] are largely consistent with activity-silent and not activity-based mechanisms[11,16,31]. Here, we sought to specify the interaction of activity-based and activity-silent PFC mechanisms in supporting serial biases while participants performed a spatial working memory task that engages attractor dynamics in the PFC[6]. Furthermore, this approach may offer indirect evidence that activity-silent and activity-based mechanisms co-occur during the delay period, as proposed by computational models (for examples, see refs. [11,16–18]).

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]Department of Neurobiology and Anatomy, Wake Forest School of Medicine, Winston-Salem, NC, USA. [3]Service of Neurology, Hospital Clínic, Barcelona, Spain. [4]University of Barcelona, Barcelona, Spain. [5]ICREA, Barcelona, Spain. [6]Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Psychology and Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA. [8]These authors contributed equally: Joao Barbosa and Heike Stein. ✉e-mail: acompte@clinic.cat

Telling these mechanisms apart in the delay period is problematic because of their coactivation. By extending the relevant task periods to the ITI, we propose a way to disentangle them and to study the effect of their interaction on upcoming memories.

We compared the encoding properties of brain activity in the delay and ITI periods to identify the mechanistic basis of the memory trace that spans consecutive trials. We used behavioral and electrophysiological data collected in monkeys and humans: prefrontal multiple-unit recordings in monkeys and scalp electroencephalography (EEG) in humans. Between successive persistent activity mnemonic codes, we found an activity-silent code in the PFC that carried stimulus information through inter-trial periods. In addition, we found correlational and causal evidence, using transcranial magnetic stimulation (TMS), to indicate that fixation-period PFC reactivation from this activity-silent trace enhances attractive serial biases. These findings underscore the behavioral relevance of the dynamic interplay between attractor and subthreshold network dynamics in the PFC and reconcile these seemingly conflicting mechanisms. Our data suggests that this interplay could be the basis of closely associated memory storage processes operating at different time scales, thereby possibly serving different behavioral purposes.

Results

We trained four rhesus monkeys to perform an oculomotor delayed response task. The task consisted of remembering spatial locations at fixed eccentricity while maintaining fixation during a delay period of 3 s (Fig. 1a; Methods). The extinction of the fixation cue triggered the monkey to execute a saccade toward the remembered location and marked the beginning of a fixed ITI of 3.1 s, lasting until the appearance of the stimulus cue of the new trial (Fig. 1b). In addition, we tested 35 human participants in variations of the task performed by the monkeys (Methods). In all cases, we recorded the reported location and computed behavioral errors as angular distances to corresponding target locations. Following the methods described in previous studies[19], we analyzed the dependence of the current-trial error on relative previous-trial location. Both monkeys and humans showed biased reports relative to previously remembered locations. These biases were attractive for short distances between previous-trial and current-trial locations, and repulsive for large previous±current distances (Figs. 1a and 2a). Our primary goal was to test the hypothesis that activity-silent and persistent activity working memory mechanisms interact to produce serial dependence effects. To this end, we investigated electrophysiological measurements in the ITI, including periods from the response to the subsequent fixation period.

Reactivation of previous memory information in the monkey dorsolateral PFC before new stimulus presentation. We collected single-unit responses from the dorsolateral PFC (dlPFC) of two monkeys while they performed the task. A substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period[6] ($n = 206$ out of 822, Methods). These specific neurons are part of bump-attractor dynamics that characterize the memory periods of this task[6]. Based on this evidence, we assumed an attractor dynamics mechanism for persistent activity, and these terms are used interchangeably to refer to this network regime. Based on our hypothesis that an interplay of activity-silent and attractor mechanisms support serial biases, we focused our analyses on these neurons, and we grouped them in simultaneously recorded ensembles for decoding analyses ($n = 94$ ensembles, size range of 1±6 neurons; Extended Data Fig. 1a).

The firing rates of dlPFC neurons exhibited strong dynamics in the ITI compared to the characteristic stable dynamics during mnemonic delay periods (Fig. 1b). Phasic rate increases at response execution ($R_{n\pm1}$, Fig. 1b) and fixation onset (F_n, Fig. 1b) were hallmarks

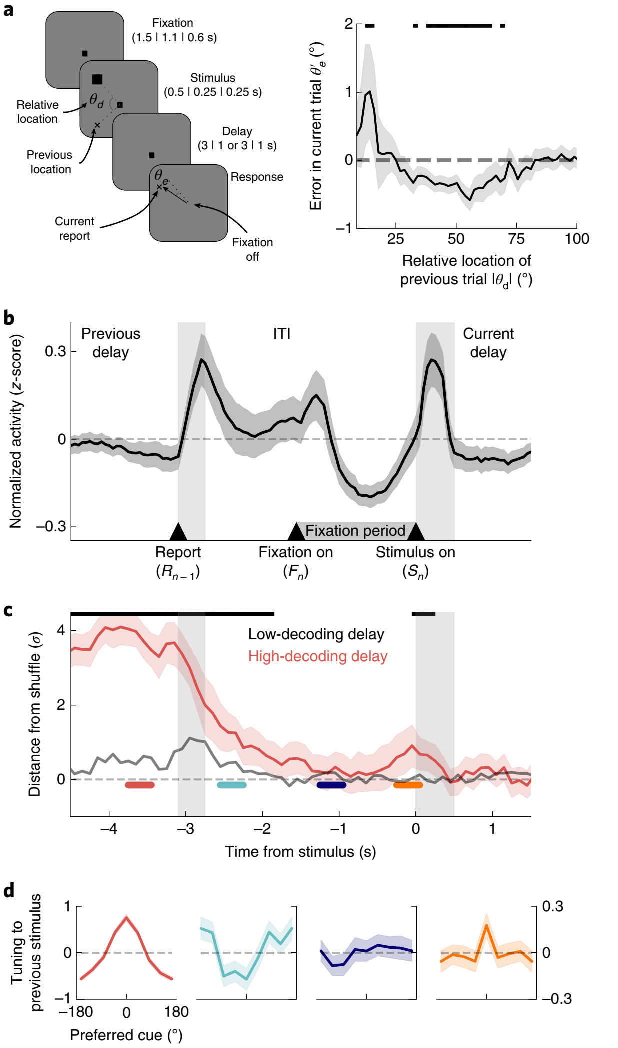


Fig. 1 | Previous-trial stimulus code reactivates before the forthcoming stimulus. a, General task design (left) and serial bias for four monkeys ($n = 11,670$ consecutive trial pairs; right). Trials with counter-clockwise previous reports relative to the current stimulus were collapsed into clockwise trials (folded errors, Methods). Positive (negative) values indicate response attraction (repulsion) toward previous locations presented at that relative distance from the current stimulus. Shading indicates bootstrapped ±s.e.m. Black horizontal solid bars represent $P < 0.05$ (one-sided permutation test). Durations in different experiments are separated by vertical bars (monkey | EEG | TMS). **b**, Averaged, normalized firing rate of $n = 206$ neurons during the ITI (spike counts of 300-ms causal square kernel, z-scored in the interval [−4.5 s, 1.5 s]). Gray vertical bars mark the response and stimulus cue periods. **c**, The decoding accuracy of previous-trial stimulus from $n = 94$ independent ensembles, computed as the distance from the mean of the decoding accuracy in shuffled surrogates, in units of their standard deviation σ (Methods), averaged over ensembles with strong (red) and weak (gray) decoding in the delay period (Methods). Aligned with anticipatory ramping in late fixation (**b**), the previous-trial stimulus code reappears specifically in ensembles with stronger delay code (Extended Data Fig. 1). Black bars mark time points for which a decoding accuracy of 99.5% CI is above zero. **d**, Tuning to previous-trial stimuli, aligning responses to the preferred cue as defined in the delay period, and computed in different trial epochs (color-coded in **c**; two-sided bootstrap test at preferred location: $P = 0.015$, CI = [−0.3, −0.03], Cohenís $d = -0.17$ (cyan); $P = 0.865$, CI = [−0.12, 0.14], Cohenís $d = 0.012$ (deep blue); $P = 0.025$, CI = [0.024, 0.33], Cohenís $d = 0.15$ (orange); $n = 206$ neurons, shading depicts ±s.e.m.). In all panels, unless stated otherwise, error shading marks bootstrapped 95% CI.

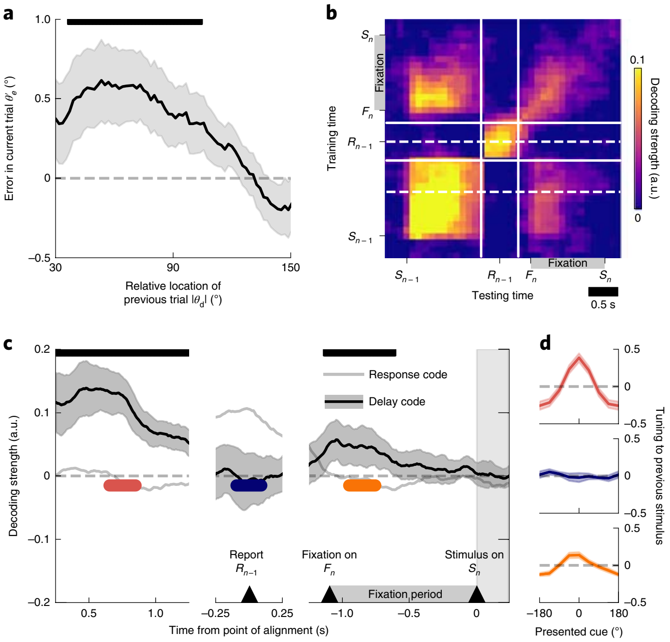


Fig. 2 | In human EEG, the delay code also reactivates in the fixation period. a, Serial bias for human participants. Shading represents ±s.e.m. **b**, Temporal generalization of previous⊠ timulus code for all combinations of training and testing times from previous⊠ rial stimulus onset ($S_{n\text{ñ}1}$) and response ($R_{n\text{ñ}1}$) to current⊠trial fixation (F_n) and stimulus onset (S_n). Solid white lines mark the discontinuity of EEG fragments aligned to $S_{n\text{ñ}1}$, $R_{n\text{ñ}1}$ and S_n. Dashed lines indicate the temporal center of transversal sections shown in **c**. a.u., arbitrary units. **c**, The decoding of previous stimulus during previous⊠ rial delay (left), response (middle) and current⊠ rial fixation period (right) for decoders trained during previous⊠ rial delay (black line, 0.5 s ñ1. s after $S_{n\text{ñ}1}$, lower dashed line in **b**) and during previous⊠ rial response (gray line, 0.5⊠ window centered on $R_{n\text{ñ}1}$, upper dashed line in **b**). The delay code is stable during the delay period, disappears during the response and reappears in current⊠ rial fixation; see also **d**. In contrast, previous⊠ rial response⊠ elated information is dynamic and not present in the fixation period. Error shading represents 95% CI. **d**, De⊠ eaned reconstruction of tuning to the previous stimulus at different epochs for the delay decoder, marked in **c** (two⊠ ided bootstrap⊠ est preferred versus anti⊠ referred location: $P<1\times10^{\text{ñ}6}$, CI = [0.55, 0.73], Cohenís $d=3.6$ (red); $P=0.69$, CI = [⊠ .22, 0.16], Cohenís $d=0.10$ (blue); $P=1\times10^{\text{ñ}6}$, CI = [0.17, 0.36], Cohenís $d=1.35$ (orange); shading represents ±s.e.m.). In **a** and **b**, the black horizontal bars indicate significant deviation from zero (bootstrap), $P<0.05$ in **a**, $P<0.005$ in **c** (both two⊠ ided). For all panels, $n=15$ independent participants.

in these dynamics, but we also noted an increase in the firing rate before stimulus presentation (S_n, Fig. 1b), which could reflect the anticipation of the upcoming stimulus due to fixed-length fixation periods. We wondered whether these rate changes were also related to dynamic changes in stimulus selectivity. Under the attractor-based hypothesis for serial biases[32], sustained stimulus selectivity would be expected to extend from the delay period of the previous trial into the fixation period of the next trial. We measured selectivity by training a linear decoder on the spike counts of our neuronal ensembles and referenced its accuracy to that obtained by chance using a resampling approach (Methods). During the delay period, neuronal ensembles carried stimulus information and single neurons showed stimulus tuning (Fig. 1c,d, red). After report, the memorized location was still decodable from ensemble activity, but the tuning curves of single neurons showed a selective suppression of responses in their mnemonic preferred locations (Fig. 1c, cyan). This could reflect neuronal adaptation mechanisms or saccade preparation toward the opposite direction to regain fixation. In the middle of the ITI, decoding accuracy was not different from chance and neurons were no longer tuned to the previous stimulus (Fig. 1c,d, deep blue), which suggests that the encoding of the previous stimulus had disappeared from neural activity. However, immediately before the presentation of the new stimulus and aligned with anticipatory ramping activity (Fig. 1b), the previous stimulus was again decoded and single-neuron tuning reappeared (Fig. 1c,d, orange). This reemergent stimulus information is consistent with previously-reported spiking selectivity during the ITI[32], but we show here that there is a period in the ITI in which stimulus information cannot be decoded before it reappears at the end of the fixation period (late fixation). Furthermore, this code in late fixation is a reactivation of the representation active in the previous trial delay. This is supported by two pieces of evidence. First, information reappearance occurred more strongly in those neuronal ensembles that maintained more stimulus information during the delay period (Fig. 1c; Extended Data Fig. 1). Second, the converging pattern of noise correlations at the end of the delay[6] and in late fixation suggested a similar attractor-like network activation in both periods. Indeed, when the preceding stimulus appeared between the preferred locations of two neurons, these PFC neuron pairs exhibited negative noise correlations in late fixation (Extended Data Fig. 2). These negative noise correlations are a signature of a fixed-shape bump that diffuses from the initial stimulus location: as it moves closer to the preferred location of one neuron and away from the other, the firing rate increases for one neuron and decreases for the other[6]. Negative noise correlations appeared exclusively during late fixation, which strongly suggests that a bump is reactivated at that specific time point (Extended Data Fig. 2). Taken together, these results support that there is a reactivation of memory-period representation in the fixation period (reactivation period) following a period of absent selective neuronal firing in the dlPFC. This reactivation points at a relationship between mechanisms of delay memory encoding and mechanisms bridging the ITI to facilitate reactivation before the new stimulus.

Previous trial memory information reactivation in the fixation period of human EEG traces. In line with the monkey electrophysiology data, we found similar previous-trial traces in human EEG data ($n=15$). We extracted alpha power from all electrodes and used a linear decoder to reconstruct the target location from EEG signals in each trial[33] (Methods). The target representation was significantly sustained during delay and response periods and in the fixation period of the next trial (Fig. 2b, diagonal axis). Importantly, at each time point, this dynamic EEG decoder uses signals originating from different cortical regions and could therefore combine temporally overlapping but spatially distinct representational components (for example, mnemonic versus response-related components). We therefore trained different linear decoders during the delay period (500±1,000 ms after stimulus onset, `delay code') and around the time of the response (250 ms before to 250 ms after response, `response code'), and used the respective weights to extract previous-stimulus information throughout different periods of the trial (Fig. 2c). The delay code was stable during stimulus presentation and delay, but disappeared during the ITI, around the time of the response. In contrast, the response code did not generalize beyond the time at which the decoder was trained (Fig. 2c). We found that the delay code of the previous trial reappeared during the fixation period (Fig. 2c,d, orange), similarly to what we found in the monkey neurophysiology data (Fig. 1c), but slightly earlier in the ITI. In our human data, reactivation was possibly triggered by the onset of the fixation dot, while reactivation in the monkey PFC could be triggered by a ramping anticipatory signal in the fixed-duration ITI (Fig. 1b). These results provide a confirmatory correspondence with the time course of mnemonic decoding in the monkey data, but they also show the temporal continuity between qualitatively distinct memory and response codes. The bidirectional transfer of information between memory and response representations in different brain areas could provide a bridge between the memory and reactivation periods observed in the PFC. Alternatively, response codes may just reflect the output motor commands, and mnemonic codes may subsist at a subthreshold level in the PFC to allow reactivations. We tested this hypothesis with a cross-correlation analysis of PFC units.

Increased cross-correlation suggests a latent trace during the ITI. We sought experimental validation for whether activity-silent mechanisms in the dlPFC still maintained stimulus information during the ITI between consecutive trials. We reasoned that if such latent activation (for example, a synaptic trace[9]) affected a group of interconnected neurons, these would be more likely to exceed their spiking threshold in synchrony[8,34]. Following a preferred cue, neurons would increase their activity in the delay period and maintain latent activity-silent traces in the subsequent ITI that would be reflected in enhanced synchrony[34], but not enhanced rates. Moreover, we deduced that this reasoning was pertinent only to effective excitatory interactions (exc); that is, neurons interacting through effective inhibition (inh) should instead show a reduced probability of coactivation following a possible inhibitory efficacy enhancement by preferred stimuli in the previous trial[34].

To test this hypothesis, we selected pairs of neurons with similar selectivity ($n=67$ pairs, Methods) so that they had consistent activation (high or low firing rate) in the delay period. As per previous studies[35,36], we divided the selected pairs on the basis of their whole-trial cross-correlation peak sign in exc and inh interactions (Methods). We considered the following two conditions (Fig. 3a; Methods): trials in which the previous stimulus was shown close to either preferred location (pref; Methods) or far from preferred locations (anti-pref). Then, we computed a cross-correlation selectivity index (CCSI) by subtracting the amplitude of the central peak of the jitter-corrected cross-correlation function (coincident spikes within 20 ms; Methods, similar to ref. [37]) for pref and anti-pref trials for each neuron pair (Fig. 3b). Our hypothesis predicts positive (negative) CCSI for exc (inh) pairs in the ITI; that is, higher (lower) spike synchrony following preferred stimuli.

The CCSI computed in a period of the ITI where the firing rate had ceased to represent the stimulus (activity-silent period, Fig. 1c,d, deep blue) was positive, which reflects selectivity in neuronal synchrony to the previous stimulus for all interactions (Fig. 3c). We then investigated changes in CCSI values for exc and inh interactions across our two periods of interest: the activity-silent and reactivation periods (Fig. 1c, deep blue and orange, respectively). We found that their reactivation-period CCSI values significantly differed, being negative for inh interactions and positive for exc interactions (Fig. 3c). Finally, we explored the CCSI dynamics throughout the trial (Fig. 3d) and found that with the exception of immediately after the previous response, in which neurons showed anti-tuning to previous-trial stimulus (Fig. 1c), the CCSI for exc pairs was always positive, indicating stronger central-peak cross-correlation when the previous stimulus was preferred (Fig. 3d, orange). Conversely, for inh interactions, the CCSI was negative (stronger inh interactions following a preferred stimulus) only during reactivation and the previous-trial delay period (Fig. 3d, cyan), the periods in which PFC firing rates showed stimulus selectivity (Fig. 1c). This pattern is consistent with the latent memory mechanism residing in excitatory neurons and only being reflected in inhibitory interactions through collective engagement in bump-attractor dynamics during the delay period and at the time of reactivation. Importantly, this analysis was done during a period without firing-rate selectivity (Fig. 3f), thus free of a potential confound from firing rates (see Extended Data Fig. 3 for the same analysis performed during the delay period, where that caveat cannot be ignored.)

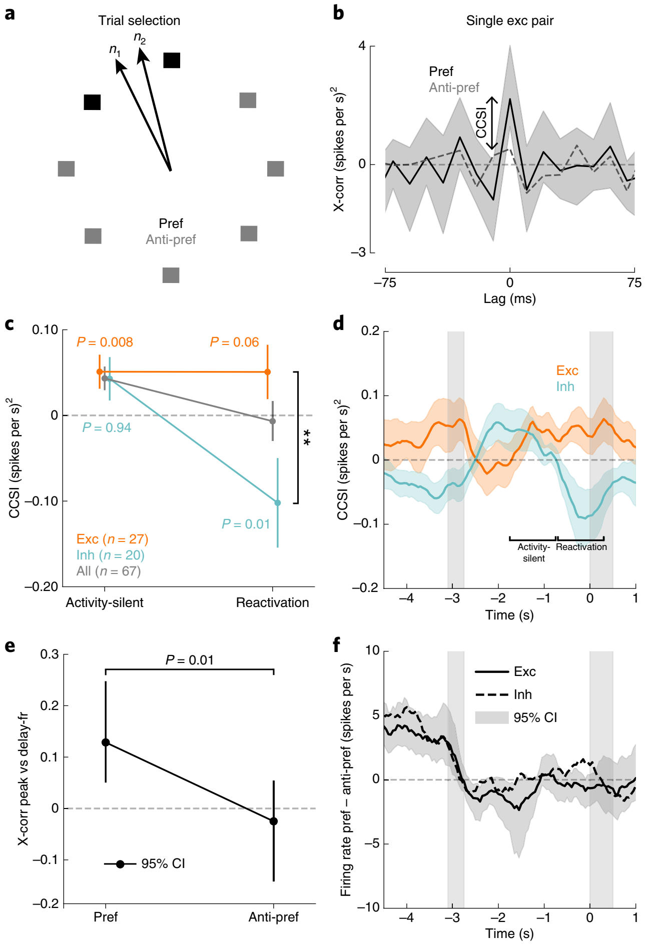


Fig. 3 | Cross-correlation selectivity to previous-trial stimulus suggests an activity-silent trace in the PFC. a, Schematic of trial selection. For neuron pairs with a similar preferred location (<60- w e separated trials with stimulus near the preferred locations (pref) of the pair from trials with far locations (anti- ref). **b**, Cross- orrelation (X- orr) of a sample PFC pair shows zero-ag p eak selectivity to a previous- rial stimulus in the activity- ilent period (one- ided permutation test, $P=0.025$, Cohenís $d=0.10$, $n=44$ independent trials). **c**, The CCSI was consistently positive in the activity- ilent period, but became negative for inh interaction pairs during reactivation (two-sided permutation test, interaction period × exc/inh, $P=0.03$, Cohenís $d=-0.6$). At reactivation, the CCSI for exc ($n=27$) and inh pairs ($n=20$) significantly differed (two- ided permutation test, $^{**}P=0.006$, $d=0.75$). P values report results of one- ailed permutation tests according to our hypotheses (CCSI > 0 for exc, CCSI < 0 for inh). **d**, The CCSI in the ITI (1- windows, 50- s steps) for exc ($n=27$) and inh pairs ($n=20$). Except immediately after the report, where neurons show anti- uning (Fig. 1d), the CCSI was positive for exc interactions. The CCSI was negative for inh interactions during previous delay and reactivation. Data were smoothed with a five- ample square filter. **e**, Trial- y- rial correlation between previous- elay spike counts for exc pairs and the ITI cross- orrelation central peak (activity- ilent period in **d**, Methods) is positive only for the pref condition (one- ided permutation test $P=0.017$, interaction $P=0.01$; $n=320$ and 769 trials for pref and anti- ref, respectively). **f**, The absence of a mean firing rate difference between the pref and anti- ref conditions (same pairs as in **d**) discards a confound between the rate selectivity and the CCSI. Error bars represent bootstrapped 95% CI (**b** and **e**) or s.e.m (**c** and **d**).

This proves the existence of a latent trace of the stimulus in the PFC during the ITI, but it could still be reflecting selective subthreshold inputs from a different area that maintains tuned persistent activity instead of selective local modulations in the PFC. To rule out this possibility and to strengthen the idea that stimulus information is directly transferred from an activity-based to an activity-silent code in the PFC, we tested whether the selectivity of exc interactions during the activity-silent period depended on the spiking activity of corresponding neurons in the previous delay period. Assuming a neuron-specific activity-dependent mechanism supporting the activity-silent code in the ITI, we predicted that the magnitude of the cross-correlation central peak in the activity-silent period would correlate on a trial-by-trial basis with the mean spike count recorded in the preceding delay period and specifically for pref (and not for anti-pref) trials (Methods). This prediction was confirmed in the experimental data (Fig. 3e). Thus, this cross-correlation analysis supports the hypothesis that previous, currently irrelevant, stimulus information remains in prefrontal circuits in latent states, undetected by linear decoders that do not take spike timing into consideration (Figs. 1c and 3f).

Bump reactivation as a mechanism for stimulus information reappearance. Based on our electrophysiology results and on prior modeling studies[9], we formulated the bump-reactivation hypothesis to explain our data. We hypothesized that information held in memory as an activity bump during the delay period of the previous trial[6] would be imprinted in neuronal synapses as a latent activity-silent trace during the ITI. This latent bump could be reactivated by the nonspecific anticipatory signal seen in the mean firing activity in the PFC (Fig. 1b) or by anticipatory mechanisms following an external cue that predicts stimulus presentation, such as the onset of a fixation dot (Fig. 2c). In fact, in a separate EEG experiment in which fixation lengths were jittered so as to make stimulus onsets unpredictable, we could not find any delay code reactivation (Extended Data Fig. 4).

To test the bump-reactivation hypothesis, we built a bump-attractor network model of spiking excitatory and inhibitory neurons. Based on our electrophysiology findings, short-term plasticity (STP) dynamics were included only in excitatory synapses (Methods). In each trial, stimulus information was maintained in activity bumps during the delay period by virtue of recurrent connectivity between neurons selective to the corresponding stimulus. During the ITI period, model neurons did not exhibit detectable tuning to the previous-trial stimulus (Fig. 4a, black, and Fig. 4b, deep blue)[16,31]. However, the synapses of neurons that had participated in memory maintenance in the previous delay period were facilitated due to STP (Fig. 4a, deep blue). Parallel to our analysis presented in Fig. 3, this was reflected in the central peak of the ITI cross-correlation for pairs of excitatory model neurons, which maintained selectivity to the previous stimulus (Fig. 4a) even in the absence of single-neuron firing-rate selectivity (Fig. 4a, deep blue). We found that single-neuron tuning could be recovered from the hidden synaptic trace using a nonspecific input (drive*)* to the entire population (Fig. 4a,c; Methods, see also refs. [9,38]). Our biologically constrained computational model was therefore an explicit implementation of the bump-reactivation hypothesis that we had formulated.

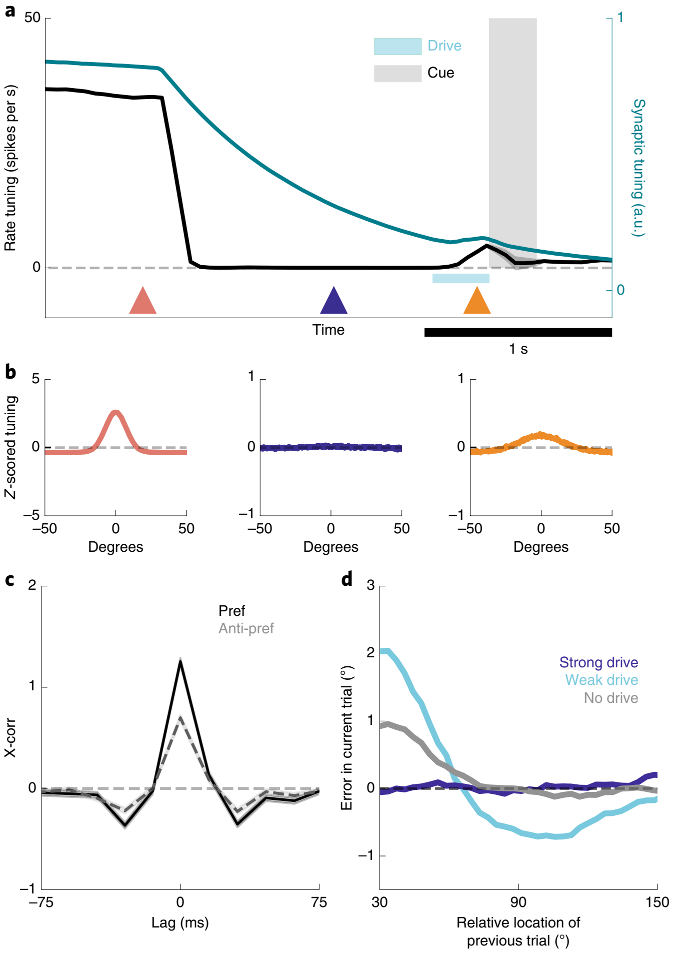


Fig. 4 | Bump-attractor model with STP accounts for serial dependence and neurophysiology. a, The average firing-rate tuning (black) and synaptic tuning (green) for 5,000 network simulations of two successive trials during the delay period (Methods). In the mnemonic period (red triangle), both rate and synaptic tuning are at their maximum, both driven by persistent bump-attractor activity (red plot in **b**). Following the memory period, a brief nonspecific hyperpolarizing input resets the baseline network state for the duration of the ITI (deep blue triangle and plot in **b**). This is reflected in a vanishing rate tuning, but long-lasting synaptic tuning that can regenerate firing-rate tuning (orange triangle and plot in **b**) through reactivation by a nonspecific input drive (cyan bar). **b**, Averaged single-neuron tuning to the previous-trial stimulus at different epochs, marked as colored triangles in **a**. **c**, Cross-correlation of model neurons in the ITI differed for the previous-trial stimulus in the preferred location (pref, black) and for anti-pref trials (gray) despite no firing-rate selectivity (**a** and **b**, deep blue). **d**, Serial bias plots computed from ‘behavioral responses’ (Methods) in three different conditions of nonspecific depolarizing drive. A weak anticipatory drive increases attractive serial biases and produces repulsion from more distinct previous memories, while a strong drive removes serial biases.

The impact of bump reactivation on serial biases. We next used our computational model to derive behavioral and physiological predictions to test in our data, in particular in relation to serial biases. To simulate serial biases with our computational model, we ran pairs of consecutive trials with varying distance between the two stimuli presented in each simulation. We used the final location of the bump in the second trial (current-trial memory) as the 'behavioral' output of the model in that trial. We were able to model the profile of serial biases that were experimentally observed (Fig. 4d; Extended Data Fig. 5), similar to previous models[16,31]. To test the impact of bump reactivation on serial biases, we compared the behavioral output of simulations with and without drive before the second trial stimulus (Methods). Bump reactivation resulted in stronger attractive biases for similar successive stimuli, and in repulsive biases for more dissimilar successive stimuli (Fig. 4d, cyan). We found that tuned intracortical inhibition[39,40] was necessary for this emergence of repulsive biases after bump reactivation (Extended Data. Fig. 5; see refs. [31,41] for an alternative mechanism). Finally, we tested the dependence of this behavioral effect on the strength of the nonspecific drive. A very short but strong impulse to the entire network during the ITI quickly saturated all the synaptic facilitation variables, effectively removing all serial biases in the output of the network (Fig. 4d, deep blue). Thus, in this model, bump reactivation nonlinearly affects serial biases as the reactivation strength is varied. In summary, our model reproduced the behavioral and neurophysiological findings described in Figs. 1±3 and derived predictions concerning memory reactivations from silent traces that we then tested in the data.

Previous stimulus reactivation increases serial biases. The model predicts that higher reactivation of previous memories in the fixation period should be associated with stronger serial biases (Fig. 4d). We tested this prediction in our neural recordings from monkey PFC and in EEG recordings from the human scalp.

Monkey PFC. We first classified each trial on the basis of leave-one-out decoding of the previous stimulus trained and tested on activity from two different time windows during fixation: during a period with no stimulus information (activity-silent period; Fig. 1, deep blue) and at the time of reactivation (Fig. 1, orange). For each of these two windows, we separated high-decoding trials (first quartile) from low-decoding trials (all other trials) and computed bias curves separately. We found that serial biases were indistinguishable in the activity-silent period (Fig. 5a), but they were stronger for high-decoding than for low-decoding trials at the time of bump reactivation (Fig. 5b). This follows the prediction of our computational model, and it confirms the behavioral relevance of the bump reactivation before stimulus onset. This result was not dependent on a singular selection of trial separations, because for different proportions of high-decoding and low-decoding trials, the serial bias strengths (Methods) changed smoothly and remained consistent with the reported result (Extended Data Fig. 6). We then repeated the same analysis at different time points of the ITI. A significant difference in serial bias strength (Methods) emerged only when trials were classified as low-decoding versus high-decoding in the reactivation period (Figs. 1c and 5c, orange), and serial biases remained virtually indistinguishable at all other time points (Fig. 5c).

Human EEG. Analogous to the analysis of the monkey data, we grouped trials on the basis of their leave-one-out decoding accuracy of the previous stimulus (Methods). We separated high-decoding and low-decoding trials at two different time points: at the time of reactivation (Figs. 2 and 5f, orange) and at a fixation-period time point without stimulus information (activity-silent; Fig. 5c, black). Consistent with the monkey data and the prediction from our model, we found a stronger serial bias for high-decoding than for low-decoding trials for the reactivation period (Fig. 5e), but not for the activity-silent period (Fig. 5d), during which previous memory content was not decodable (Fig. 2c). The analysis was repeated for all other time points during the fixation period (Fig. 5f). Indeed, behavior exclusively depended on decoding accuracy at the time of delay code reactivation (Fig. 2, orange). Taken together, these results support the hypothesis that previous-trial memory reactivation before stimulus onset controls serial biases.

TMS-induced reactivations modulate serial biases. As a causal validation of the influence of pre-stimulus PFC reactivation on serial biases, we designed a TMS study. This is a relevant experiment because memory-dependent changes in human EEG alpha power cannot be unequivocally ascribed to a specific brain region, which limits the correspondence of our EEG and monkey dlPFC data. In particular, representations in larger and more organized occipital cortices might strongly contribute to visual EEG signals (for example, see ref. [33]), but could yet be driven by top-down projections from association cortices[42]. Inspired by a previous study[14] that reported reactivation of latent memories using TMS, we causally tested the role of the dlPFC in serial biases by applying single-pulse TMS during the fixation period. We had two control conditions to test our hypotheses: (1) we targeted the TMS coil at the dlPFC and the vertex in interleaved blocks, and (2) we randomly chose the TMS intensity in each trial (sham: 0%, weak-TMS: 70%, and strong-TMS: 130% of the resting motor threshold (RMT) of each participant; Methods). We found that TMS modulated serial biases when targeted at the dlPFC but not at the vertex (Fig. 6). Moreover, our computational model predicted a nonlinear dependence with stimulation strength (Fig. 4d), which was supported by the TMS data (Fig. 6b). Interestingly, the behavioral impact of PFC TMS stimulation declined throughout the session, as if participants became desensitized to the TMS pulse (Extended Data Fig. 7). Importantly, we show combined results from two separate experiments of $n=10$ participants each, one being a preregistered replication (Methods; Extended Data Figs. 8 and 9). These results provide causal evidence for the involvement of the PFC in the serial bias machinery during the ITI. Furthermore, we show that TMS affects serial biases in a nonlinear manner, as predicted by model simulations that implement the bump-reactivation hypothesis via the interplay of bump attractor and activity-silent mechanisms.

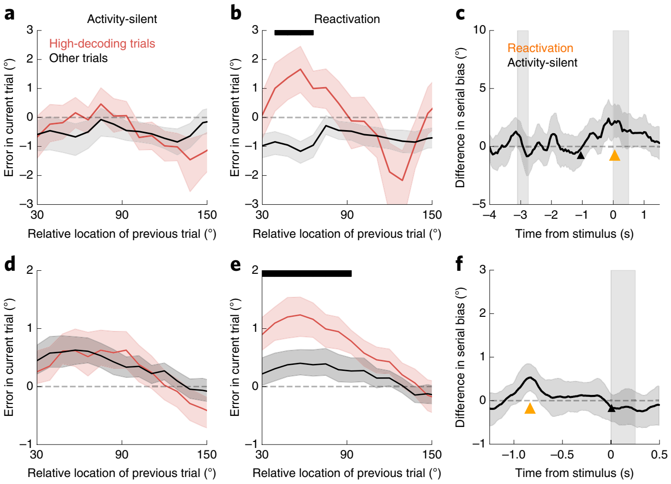


Fig. 5 | Bump reactivation from a hidden trace increases serial biases. Serial bias for trials with high previous⊠rial stimulus information (upper quartile, red) and for all other trials (black) in monkeys (**a**ñ**c**, $n=1,362$ trials) and in humans (**d**ñ**f**, $n=15$ participants, with a range of 792ñ908 trials in this analysis). See Extended Data Fig. 6 for different quantiles. **a**, Trials selected based on a decoder trained and tested early in the fixation period (black triangle in **c**), did not reveal differences in serial bias. **b**, Serial biases were markedly enhanced for high⊠ ecoding trials when training and testing the decoder at the time of reactivation (Fig. 1c, orange triangle in **c**). **c**, Differences in serial bias curves between high⊠ ecoding and other trials became significant only in late fixation, concomitant with reactivation (Fig. 1c). Triangles mark the center of decoding windows for the splits shown in **a** and **b**. **d**ñ**f**, Same analyses for human EEG ($n=15$ independent participants). Note that for humans, **d** corresponds to an activity⊠ ilent period in late fixation (black triangle in **f**), and **e** to the reactivation period in early fixation (Fig. 2c, orange triangle in **f**). **f**, As for monkeys, serial bias differences in humans were significant only during reactivation. In **c** and **f**, time courses of differences between high⊠ ecoding and other trials were smoothed in time using a 5⊠ ample (monkey) and 16⊠ ample (human) square filter. Black horizontal bars (**b** and **c**) mark significant differences between high⊠ ecoding and other trials ($P<0.05$, one⊠ ided permutation test). Error shading represents 95% CI (**c** and **f**) or ±s.e.m. (**a**, **b**, **d** and **e**).

Discussion

By studying the neural basis of serial biases, we showed how the interplay of bump-attractor dynamics and activity-silent mechanisms

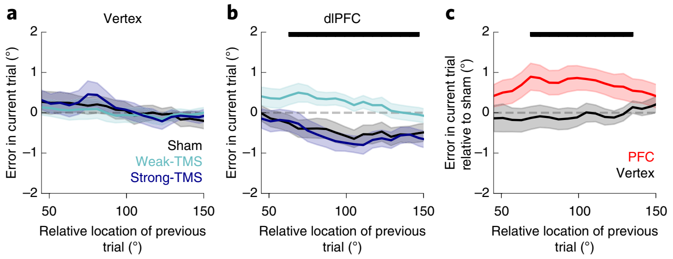


Fig. 6 | Single-pulse TMS on the dlPFC during fixation modulates serial biases nonlinearly. a,**b**, Serial bias computed in vertex (**a**) and PFC (**b**) blocks, separately for trials with strong fixation-applied TMS pulse (130% of RMT, deep blue), weak (70% RMT, cyan) and sham (0% RMT, black) for the first half of each session (225 trials, 2 sessions per participant, $n=20$ participants; Extended Data Figs. 7ñ9). Serial biases were modulated by TMS in the PFC but not in the vertex (prev-curr × TMS intensity × coil location, $t_{18,272}=2.21$, $P=0.027$. For the dlPFC: prev-curr × TMS intensity, $t_{11,087}=2.13$, $P=0.032$. For the vertex: $t_{7,166}=0.03$, $P=0.97$. Methods, linear mixed models; analysis performed on the entire session). In the PFC, serial bias modulation depended nonlinearly with the stimulation strength (ΔAIC = 4.6, relative likelihood 0.9, for the comparison of regression models with nonlinear versus linear TMS intensity factor; Methods). **c**, The difference between serial biases computed for sham and weak-TMS trials in the vertex (black) and in the PFC (red) blocks. Error bars are bootstrapped ±s.e.m. Solid black bars (**b** and **c**) mark significant differences (two-sided permutation test, $P<0.05$, $n=20$ independent participants).

in the PFC maintains and eventually reactivates information about previous stimuli in spatial working memory. In delayed-response tasks, prefrontal tuned persistent activity consistent with bump-attractor dynamics characterizes the delay period and correlates with behavioral precision[6,43]. We have now seen that this sustained activation disappears from the prefrontal network between trials, but is reactivated before the new trial (Figs. 1 and 2) and enhances behavioral serial biases (Figs. 5 and 6). This reactivation is directly linked to previous-trial activity: it emerged specifically in those neural ensembles that showed strongest persistent tuning in the delay period (Fig. 1c; Extended Data Fig. 1), it was decoded from human EEG data with decoders trained in the delay period (Fig. 2) and it exhibited the fingerprints of bump attractors as evaluated using pairwise correlations (Extended Data Fig. 2). Activity-silent mechanisms in the PFC bridge disconnected periods of persistent activity, carrying trial-specific information from one trial to the next (Fig. 3). Importantly, this latent tuning was directly associated with trial-by-trial firing rates in the preceding delay period (Fig. 3e), thus establishing a coupling between activity-based and activity-silent mechanisms in the PFC. Taken together, our results are consistent with the view that attractor-based and activity-silent mechanisms are jointly represented in the prefrontal circuit and that their tight interplay influences representations in spatial working memory. We specified this in a computational network model, whereby delay-period attractor dynamics imprint activity-silent mechanisms, which then retain information between trials and allow reactivations to recapitulate attractor states (Fig. 4).

Our data indicate that nonspecific PFC stimulation can revive subthreshold information, thus supporting the ideas put forward in computational models[9] and in previous neuroimaging and EEG studies[14,44,45]. Importantly, we obtained explicit causal evidence supporting the role of ITI reactivations in enhancing serial biases. Similarly, recent causal evidence obtained in rodents[26] showed the role of parietal activations in generating history-dependent biases. However, the absence of selective mnemonic delay activity in rat parietal neurons[26] suggests that parietal ITI representations do not emerge from trace reactivations. A directed mechanistic investigation of the rat posterior parietal cortex in this task, similar to our efforts here, would be necessary to clarify the mechanisms and origin of history biases, and potential differences between the generation of contraction and serial biases in rodents and primates. More in line with our reasoning, human TMS studies found behavioral effects of memory reactivations when applied in the delay period, but only when memories were still behaviorally relevant[14]. In contrast, we show here that fixation-period TMS enhanced the behavioral influence of previous, already irrelevant memories. Reactivations may therefore not depend on behavioral relevance but rather on the decaying dynamics of activity-silent mechanisms; a more advanced decay of irrelevant memory traces may limit memory reactivations in ref. [14]. Reactivations also offer alternative explanations to TMS effects in working memory that have previously been interpreted on the basis of network disruptions[46].

Our data support the idea that activity-silent and attractor-based mechanisms are not orthogonal, alternative mechanisms, but that they are interdependent mechanisms colocalized in the PFC. In turn, their different timescales may associate them preferentially with different types of memory processes. During active maintenance of working memory, rapid persistent attractor-based activity may encode memory, with slower activity-silent mechanisms providing a supporting, stabilizing role[11,16,17]. Note that although direct evidence of this interplay in the delay period is problematic (Extended Data Fig. 3), our approach of separately assessing delay period and ITI, and their trial-by-trial correlation, indirectly supports this interplay and may be the most direct evidence that can be accessed extracellularly without resorting to detailed intracellular measurements in awake monkeys. After the deactivation of attractor-based active maintenance in the ITI, slowly decaying activity-silent maintenance may underlie secondary, possibly involuntary memory traces, leading to serial biases in upcoming trials. Note that previous studies have also proposed a central role for activity-silent maintenance for an additional, intermediate type of memory: unattended, behaviorally relevant memories[14,44]. It was hypothesized that by resorting to different mechanisms, unattended memories may be reserved and protected while processing attended memories. Although our data do not address the mechanism of unattended memories, in our proposed framework, the close interplay between attractor-based and activity-silent mechanisms does not allow unattended memories (activity-silent memories) to be protected from intervening attended memories (attractor-based). This yields the prediction that serial-bias-like patterns of interference[39,40] between unattended and attended memories should be observed in these experiments[14,44].

Our results have implications for the functional interpretation of serial biases and their relation with the interplay of prefrontal mnemonic mechanisms. First, enhanced serial biases after reactivating latent traces from earlier memories are consistent with the view that biases are the by-product of memory-supporting processes. As previous computational studies have shown, long-lasting cellular or synaptic mechanisms can enhance the stability of working memory retention (for examples, see refs. [11,16±18]), but with the cost of across-trial interference of memories[11,16]. Along these lines, a recently found reduction in serial biases in patients with schizophrenia[41], anti-NMDA receptor encephalitis[41] or autism[28] may reflect a reduced interplay of memory-supporting mechanisms. Second, we see an active role of the PFC in generating serial biases, rather than suppressing them as proposed by the proactive interference literature[29,30]. This discrepancy could be resolved if the role of PFC was two-sided: (1) the PFC could generate biases either as a by-product of stable memory retention[11,16] or actively, in circumstances in which past memory traces are adaptive for behavior[24]; alternatively, (2) strong PFC activation would suppress maladaptive memory remnants in situations where biases are particularly detrimental to behavioral performance. This dual PFC function is supported in our modeling and TMS data by the contrasting effect of weak and strong PFC activation on serial biases.

Our TMS experiment clarified our EEG results by demonstrating the role that the PFC plays in serial biases. Because we did not concurrently acquire EEG data during the TMS study, we could not directly measure the neural reactivation induced by the TMS pulse. However, prior work has shown the reactivation of EEG memory representations with TMS[14], albeit in different conditions (pulses in the memory period targeted at parietal and occipital regions). Intriguingly, serial biases for trials without TMS stimulation in PFC-stimulation blocks were repulsive (Fig. 6b). We speculate that this was due to suppressive long-lasting physiological effects in the PFC that carried over from previous TMS-stimulated trials in the block[47] (see Extended Data Fig. 10 for a phenomenological model of this hypothesis). Future work involving more fine-grained TMS intensities and carefully controlled block designs will be necessary to further clarify these results.

We proposed a computational model that can parsimoniously explain our data using STP in the synapses of a recurrent network. STP has also been used in previous computational models of interacting activity-based and activity-silent dynamics[9,10,13] and of serial biases[16,31]. Beyond previous modeling efforts, we explored the mechanistic requirements of code reactivations before a new trial, and we derived predictions whose validation conferred plausibility to the model. Our findings do not unequivocally identify this mechanism and we could have chosen another mechanism with a long time constant to computationally implement our hypothesis (for example, calcium-activated depolarizing currents[17], depolarization-induced suppression of inhibition[11] or short-term potentiation[48]). Also, synaptic plasticity mechanisms linked to feedforward connections into the PFC[38] could conceivably play a role. Still, several lines of evidence support the involvement of STP in prefrontal function. First, there is explicit evidence for enhanced short-term facilitation and augmentation among PFC neurons in in vitro studies[49,50]. Second, extracellular recordings in behaving animals cannot directly probe activity-silent mechanisms, but indirect evidence for synaptic plasticity has been gathered from prefrontal activity correlations of rodents engaged in working memory tasks[35]. Our study also follows this approach to seek evidence for activity-silent stimulus encoding, but we applied it specifically at time periods without firing-rate codes for task stimuli, thus unambiguously decoupling activity-silent from activity-based selectivity (Fig. 3; Extended Data Fig. 3).

In summary, our data show that subthreshold traces of recent memories remain imprinted in PFC circuits and bias behavioral output in working memory in particular through network reactivations of recent experiences. Our findings suggest that the dynamic interplay between attractor and subthreshold network dynamics in the PFC supports closely associated memory storage processes: from effortful memory to occasional reactivation of fading experiences.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-0644-4.

Received: 22 August 2019; Accepted: 21 April 2020;
Published online: 22 June 2020

References

1. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
2. Kubota, K. & Niki, H. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiol.* **34**, 337–347 (1971).
3. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
4. Leavitt, M. L., Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Sustained activity encoding working memories: not fully distributed. *Trends Neurosci.* **40**, 328–346 (2017).
5. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The distributed nature of working memory. *Trends Cogn. Sci.* **21**, 111–124 (2017).
6. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
7. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).
8. Stokes, M. G. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
9. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
10. Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* **22**, 1159–1167 (2019).
11. Carter, E. & Wang, X.-J. Cannabinoid-mediated disinhibition and working memory: dynamical interplay of multiple feedback mechanisms in a continuous attractor model of prefrontal cortex. *Cereb. Cortex* **17**, i16–i26 (2007).
12. Fiebig, F. & Lansner, A. A spiking working memory model based on Hebbian short-term potentiation. *J. Neurosci.* **37**, 83–96 (2017).
13. Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).
14. Rose, N. S. et al. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
15. Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C. & Haynes, J.-D. Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* **21**, 494–496 (2018).
16. Kilpatrick, Z. P. Synaptic mechanisms of interference in working memory. *Sci. Rep.* **8**, 7879 (2018).
17. Tegnér, J., Compte, A. & Wang, X.-J. The dynamical stability of reverberatory neural circuits. *Biol. Cyber.* **87**, 471–481 (2002).
18. Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput. Biol.* **15**, e1006928 (2019).
19. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).
20. Papadimitriou, C., Ferdoash, A. & Snyder, L. H. Ghosts in the machine: memory interference from the previous trial. *J. Neurophysiol.* **113**, 567–577 (2015).
21. Fritsche, M., Mostert, P. & de Lange, F. P. Opposite effects of recent history on perception and decision. *Curr. Biol.* **27**, 590–595 (2017).
22. Bliss, D. P., Sun, J. J. & D'Esposito, M. Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* **7**, 14739 (2017).
23. Jonides, J. & Nee, D. E. Brain mechanisms of proactive interference in working memory. *Neuroscience* **139**, 181–193 (2006).
24. Kiyonaga, A., Scimeca, J. M., Bliss, D. P. & Whitney, D. Serial dependence across perception, attention, and memory. *Trends Cogn. Sci.* **21**, 493–497 (2017).
25. Barbosa, J. & Compte, A. Build-up of serial dependence in color working memory. Preprint at https://www.biorxiv.org/content/10.1101/503185v1 (2018).
26. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).
27. Hermoso-Mendizabal, A. et al. Response outcomes gate the impact of expectations on perceptual decisions. *Nat. Commun.* **11**, 1057 (2020).
28. Lieder, I. et al. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nat. Neurosci.* **22**, 256–264 (2019).
29. D'Esposito, M., Postle, B. R., Jonides, J. & Smith, E. E. The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proc. Natl Acad. Sci. USA* **96**, 7514–7519 (1999).
30. Feredoes, E., Tononi, G. & Postle, B. R. Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. *Proc. Natl Acad. Sci. USA* **103**, 19530–19534 (2006).
31. Bliss, D. P. & D'Esposito, M. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS ONE* **12**, e0188927 (2017).

32. Papadimitriou, C., White, R. L. & Snyder, L. H. Ghosts in the machine II: neural correlates of memory interference from the previous trial. *Cereb. Cortex* **27**, 2513–2527 (2017).
33. Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K. & Awh, E. The topography of alpha-band activity tracks the content of spatial working memory. *J. Neurophysiol.* **115**, 168–177 (2016).
34. Trousdale, J., Hu, Y., Shea-Brown, E. & Josić, K. Impact of network structure and cellular response on spike time correlations. *PLoS Comput. Biol.* **8**, e1002408 (2012).
35. Fujisawa, S., Amarasingham, A., Harrison, M. T. & Buzsáki, G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. Neurosci.* **11**, 823–833 (2008).
36. Barthó, P. et al. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J. Neurophysiol.* **92**, 600–608 (2004).
37. Cohen, J. Y. et al. Cooperation and competition among frontal eye field neurons during visual target selection. *J. Neurosci.* **30**, 3227–3238 (2010).
38. Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P. & Husain, M. Neural mechanisms of attending to items in working memory. *Neurosci. Biobehav. Rev.* **101**, 1–12 (2019).
39. Almeida, R., Barbosa, J. & Compte, A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J. Neurophysiol.* **114**, 1806–1818 (2015).
40. Nassar, M. R., Helmers, J. C. & Frank, M. J. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychol. Rev.* **125**, 486–511 (2018).
41. Stein, H. et al. Disrupted serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. Preprint at https://www.biorxiv.org/content/10.1101/830471v1 (2019).
42. Reinhart, R. M. G. et al. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *J. Neurosci.* **32**, 7711–7722 (2012).
43. Sajad, A., Sadeh, M., Yan, X., Wang, H. & Crawford, J. D. Transition from target to gaze coding in primate frontal eye field during memory delay and memory-motor transformation. *eNeuro* **3**, ENEURO.0040-16.2016 (2016).
44. Wolff, M. J., Jochim, J., Akyürek, E. G. & Stokes, M. G. Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20**, 864–871 (2017).
45. Bae, G.-Y. & Luck, S. J. Reactivation of previous experiences in a working memory task. *Psychol. Sci.* **30**, 587–595 (2019).
46. Zokaei, N., Manohar, S., Husain, M. & Feredoes, E. Causal evidence for a privileged working memory state in early visual cortex. *J. Neurosci.* **34**, 158–162 (2014).
47. Moliadze, V., Zhao, Y., Eysel, U. & Funke, K. Effect of transcranial magnetic stimulation on single-unit activity in the cat primary visual cortex. *J. Physiol.* **553**, 665–679 (2003).
48. Volianskis, A. et al. Long-term potentiation and the role of *N*-methyl-ᴅ-aspartate receptors. *Brain Res.* **1321**, 5–16 (2015).
49. Wang, Y. et al. Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
50. Hempel, C. M., Hartman, K. H., Wang, X. J., Turrigiano, G. G. & Nelson, S. B. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* **83**, 3031–3041 (2000).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Behavioral task and recordings. *Monkey behavioral task and recordings.* Four adult (>6 years old), male rhesus monkeys (*Macaca mulatta*) were trained in an oculomotor delayed response task requiring them to fixate, view a peripheral visual stimulus on a screen at a distance of 50 cm and make a saccadic eye movement to its location after a delay period. During execution of the task, neurophysiological recordings were obtained from the dlPFC. Detailed methods of the behavioral task, training, surgeries and recordings, as well as descriptions of neuronal responses in the task, have been previously published[6,51–54] and are only summarized briefly here. Visual stimuli were 1° squares, flashed for 500 ms at an eccentricity of either 12° or 14°, indicated as degrees of visual angle. Stimuli were randomly presented at one out of eight possible locations around the fixation point. A delay period lasting 3 s followed the presentation of the stimulus, at the end of which the fixation point turned off and a saccade terminating within 5° from the location of the remembered stimulus was reinforced with a liquid reward (5° corresponds to about 20° of arc on the circle of possible cues). Although fixation was maintained through cue and delay periods, we denote the fixation period as the interval between fixation onset and cue onset, when the only behavior expected was fixation (fixation period, Fig. 1b). A fixed ITI of 3.1 s elapsed between fixation cue extinction and the onset of the cue in the next trial (ITI, Fig. 1b). Eye position was monitored using a scleral eye coil system in two monkeys and an ISCAN camera in the other two. From two of those monkeys, we collected single-unit responses from the dlPFC using tungsten electrodes of 1–4 MΩ impedance at 1 kHz while they were performing the task[51]. Simultaneous recordings were obtained from arrays of 2–4 microelectrodes spaced 0.2–1 mm apart. A substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period of the task ($n=206$ out of 822 neurons[6,51–54]). For decoding analyses, we grouped those neurons in simultaneously recorded ensembles (total of $n=94$ neural ensembles, 1–6 neurons per ensemble, Extended Data Fig. 1a). All experiments were conducted in accordance with the guidelines set forth by the US National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee, and by the Wake Forest University Institutional Animal Care and Use Committee. Data collection and analyses were not performed blinded to the conditions of the experiments. No statistical methods were used to predetermine sample sizes, and we followed the customary practice of testing $n=2$ monkeys for electrophysiology data and $n=4$ monkeys for behavioral data. We note that the electrophysiology data were previously acquired and have been used in other publications[6,51–56].

Human participants and behavioral task. Thirty-five neurologically and psychologically healthy volunteers with normal or corrected vision (EEG experiment: $n=15$ (4 male), 21.27 ± 4.86 years (mean ± s.d.); two additional participants were tested, but aborted the EEG experiment with insufficient trials; TMS experiments: $n=20$ (6 male), 29.86 years ± 9.55 years (mean ± s.d.); one additional participant was excluded before their MRI scan due to health concerns) from the Barcelona area provided written informed consent and were monetarily compensated for their participation, as reviewed and approved by the Research Ethics Committee of the Hospital Clínic de Barcelona. During both the EEG and TMS experiments, each participant performed two sessions lasting approximately 1.5 h. To perform behavioral and EEG analyses, we concatenated the two sessions for each participant. Stimuli were presented on a 17″ HP ProBook viewed at a distance of 65 cm, and we used Psychopy (v.1.82.01) running on Python 2.7. The TMS study consisted of an initial experiment with ten participants and a preregistered replication experiment (https://osf.io/rguzn/) with ten more participants (Extended Data Figs. 7–9). For all three studies (one EEG and two TMS experiments), we recruited independent participant pools. For the fully randomized within-subjects design of our EEG task, condition-blind data collection and analyses were not a critical issue. In the TMS study, the experimenter could not be blinded to the location of the coil. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in relevant previous publications[14,33,46].

In each 1.5-h EEG session, participants completed 12 blocks of 48 trials (except for one participant, who completed 12 blocks in one session and 9 blocks in the second session). Each trial began with the presentation of a central black fixation dot (0.5 × 0.5 cm) on a gray background. After 1.1 s of fixation, a single colored circle (stimulus, diameter of 1.4 cm) appeared for 0.25 s at any of 360 circular locations at a fixed radius of 4.5 cm, randomly sampled from a uniform distribution. In 66.67% of trials (a total of 768 trials per participant), the stimulus was followed by a 1-s delay in which only the fixation dot remained visible. In the remaining trials, the delay duration was either 3 s (16.67% of trials, 192 trials per participant) or 0 s (16.67% of trials, 192 trials per participant). Trials with 0-s delay were excluded from the analyses in this study. The change in the fixation dot color (from black to the stimulus color) instructed participants to respond (response probe). Participants responded by making a mouse click at the remembered location. A transparent circle with a white border indicated the radial distance of the stimulus, so the participant was only asked to remember its angular location. After the response was given, the cursor had to be moved back to the fixation dot to self-initiate a new trial. The total length of the ITI, defined as the time between response probe and the next stimulus onset, was around 2.72 s (median, 95% confidence intervals (CIs) = [2.11 s, 4.16 s]). Participants were instructed to maintain fixation during pre-stimulus fixation, stimulus presentation and delay, and were free to move their eyes during the response and when returning the cursor to the fixation dot. Colors (one out of six colors with equal luminance) were randomly chosen with an equal probability for each trial.

Stimuli and the trial structure in the TMS task were similar to the EEG task, except for the fixation period duration (0.6 s), screen background (white), stimulus color (black) and response probe color (red). At the end of the fixation period (16.7 ms before stimulus onset), a single TMS pulse was applied in half of the vertex trials (TMS or sham trials, randomly interleaved) and in two-thirds of prefrontal trials (weak or strong TMS or sham trials, randomly interleaved). See TMS details below. Only delays of 1 s were used in this experiment. Participants completed 4 blocks of 90 (vertex) and 4 blocks of 130 (PFC) trials within each session. In the first TMS study, these eight blocks were randomly shuffled for each session. In the replication TMS study, we successively alternated vertex and PFC blocks within each session, and the two sessions of a given participant started alternately with each area in a counterbalanced design.

EEG recordings and preprocessing. We recorded EEG data from 43 electrodes attached directly to the scalp. The electrodes were located at the following modified combinatorial nomenclature sites: Fp1, Fpz, Fp2, AF7, AFz, AF8, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, A1, T7, C5, C3, Cz, C4, C6, T8, A2, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2. Sites were referenced to an average of mastoids A1 and A2 and re-referenced offline to an average of all electrodes. We further recorded horizontal electrooculography data from both eyes, vertical electrooculography data from an electrode placed below the left eye and electrocardiography data to detect cardiac artifacts. We used a Brainbox EEG-1166 EEG amplifier with a 0.017–100 Hz bandpass filter and digitized the signal at 512 Hz using Deltamed Coherence software (v.5.1).

EEG data were preprocessed using Fieldtrip (v.20171231) in Matlab R2017b and R2019a. We excluded outlier trials in which variance or kurtosis across samples exceeded four standard deviations from mean variance or kurtosis over trials, respectively. To reduce artifacts in the remaining data, we ran an independent component analysis on the trial-segmented data and corrected the signal for blinks, eye movements and electrocardiogram signals, as identified by visual inspection of all components. Data were Hilbert-transformed (using the FieldTrip function ft_freqanalysis.m) to extract frequencies in the alpha band (8–12 Hz), and total power was calculated as the squared complex magnitude of the signal. Finally, we excluded trials in which log-normal alpha power at any electrode exceeded the time-resolved trial average of log-normal alpha power by more than four standard deviations, and trials in which the time-averaged variance across electrodes exceeded the mean variance over trials by more than four standard deviations (to increase the stability of trial-wise decoding predictions for different randomly chosen training sets). In total, we rejected an average of 3.95 ± 1.07% (mean ± s.d.) of trials per participant. Excluding rejected trials and trials with 0-s delay, we used 914.33 ± 28.94 trials per participant. To concatenate data from the two sessions for the same participant, we normalized the alpha power of each session for each electrode separately.

TMS study. Stimulation was performed in the TMS study using a Magstim Rapid 2 machine with a 70-mm figure-of-eight coil. TMS target points were located using a BrainSight navigated brain stimulation system that allowed coordination of the coil position based on the structural MRI scan of each participant. A region of interest in the right dlPFC (MNI152 coordinates $x=40$, $y=34$, $z=16$) was defined using a NeuroSynth[57] term-based meta-analysis of 53 functional MRI studies associated with the key phrase 'spatial working memory' (Supplementary Fig. 1 and Supplementary Data). This mask was transformed into the structural MRI space of each participant. Vertex target points were defined using the 10–20 measurement system. Stimulator intensity, coil position and coil orientation were held constant for each participant for the duration of each session. To mask the sound of TMS coil discharge, we had participants listen to white noise through earphones for the duration of the session. White noise volume was selected based on the threshold of the participant for detecting a TMS click using the staircase method (two up, one down). Stimulation intensity was determined by the individually defined RMT. We applied two different TMS intensities at 70% RMT (weak-TMS, 24.5–41.5% (min–max) of stimulator output) and 130% RMT (strong-TMS, 45.5–76.5% of stimulator output) depending on the trial (see main text). To reduce the number of trials per session, we applied strong-TMS at the vertex in the original study, but weak-TMS for the replication study (preregistered at https://osf.io/rguzn/; Extended Data Figs. 9 and 10). The stimulation parameters were in accordance with published TMS guidelines[58]. In a post-experiment debriefing session, we collected information about the subjective experience of the participants. Many participants (13 out of 20) reported facial muscle twitching in the dlPFC blocks. This is an unlikely explanation for the effects observed in Fig. 6 because (1) twitching is expected to increase with TMS intensity, but we instead observed a nonlinear dependency in our effect (Fig. 6b), and (2) behavioral performance in our task as measured by the precision of the responses was not modulated by the TMS intensity in the dlPFC blocks (linear mixed model: $\theta_e^2 \sim \text{intensity} + (1|\text{subject})$, $P > 0.5$), which suggests that our reported intensity-dependent effect (Fig. 6b) was not the result of a general behavioral impairment caused by facial twitching.

Serial bias analysis. *Human study.* For each trial, we measured the response error (θ_e) as the angular distance between the angle of the presented stimulus and the angle of the response. To exclude responses produced by guessing or motor imprecision, we only analyzed responses within an angular distance of 1 radian and a radial distance of 2.25 cm from the stimulus. Furthermore, we excluded trials in which the time of response initiation exceeded 3 s, and trials for which the time between the response probe of the previous trial and the stimulus presentation of the current trial exceeded 5 s. On average, 2.99 ± 4.51% (mean ± s.d.) of trials per participant were rejected.

We measured serial biases as the average error in the current trial as a function of the circular distance between the target locations of the previous and the current trial (θ_d) in sliding windows with size $\pi/3$ and in steps of $\pi/20$ radians, and steps of $\pi/100$ radians for Fig. 2a (note that for easier interpretability, all figures depict values in angular degrees). To increase power and correct for global response biases, we calculated a 'folded' version of serial biases as follows[25]. We multiplied trial-wise errors by the sign of θ_d: $\theta'_e = \theta_e \times \mathrm{sign}(\theta_d)$, and used absolute values of θ_d. Positive mean folded errors should be interpreted as attraction toward the previous stimulus and negative mean folded errors as repulsion away from the previous location. For a scalar estimate of differences in serial bias curves (Fig. 5f), we averaged folded errors for close θ_d distances (between 0 and $\pi/2$ radians).

Monkey study. In contrast to the human study, the stimulus distribution was discrete for all the monkey experiments. On each trial, the subject was cued to one of eight possible cue locations equidistant on a circle. This restricted the minimal angular distance between cues in two consecutive trials to be $\pi/4$ radians. To obtain a finer resolution to calculate serial biases, we capitalized on the response variability on each trial: we computed θ_d as the distance between the stimulus of the current trial and the response of the previous trial (instead of the stimulus of the previous trial). Similar methods to the human study were used, except for Fig. 1a, where we used smaller sliding window sizes ($\pi/10$ in steps of $\pi/100$ radians), which was essential to capture the thinner attractive serial bias profile in monkeys (Fig. 1a). Specific differences in our monkey and human serial bias curves (Figs. 1a and 2a) may be due to the discrete stimulus distribution (eight possible locations) that we used for monkeys, in contrast to the continuous distribution used in our human experiments. Indeed, studies with larger samples and continuous stimulus distributions have reported behavioral biases in monkeys more consistent with the human literature[20,32]. For all our serial bias curves, x axis coordinates mark the central value of the corresponding sliding window.

Statistical methods. Data were analyzed using custom scripts in Python 2.7 (monkey and TMS data) and in Python 3.7.4 (human EEG data). Details of statistical methods are tabulated in the Nature Research Reporting Summary available online. Unless stated otherwise, all hypothesis tests were two-tailed (permutation tests or bootstrap hypothesis test, $n = 10^6$) and CI are at [2.5, 97.5] percentiles of a bootstrapped distribution. Using bootstrap distributions, we avoid assuming normality for our statistical tests. One exception was the linear model used for TMS data analyses, in which normality was assumed. Supplementary Fig. 2 shows the distribution of residuals of this model and the corresponding qqplot. There was a significant deviation from normality in extreme values. This did not compromise our statistical inference because of the large sample size ($n = 18{,}299$ trials)[59] and because the interaction of interest was confirmed by model-free analyses (Fig. 6; Extended Data Figs. 7–9).

To test the effect of TMS on serial biases, we fit a linear mixed-effects model using the R function lme[60]. In particular, we modeled trial-wise behavioral errors θ_e as a linear model with interaction terms for coil location (PFC versus vertex), TMS intensity (strong-TMS, sham and weak-TMS) and the sine of θ_d (prev-curr), which approximates the expected dependency of θ_e on θ_d in the presence of serial biases ($\theta_e \propto \sin(\theta_d)$). We incorporated the nonlinear dependency of serial bias on stimulation intensity that our model simulations predicted by using −1, 0 and 1 for strong-TMS, sham and weak-TMS, respectively. In one model, we used instead the nominal percent of RMT TMS intensity used (70, 0 and 130, respectively) for comparison (Fig. 6b). We accounted for subject-by-subject variability by including random-effect intercepts and random-effect coefficients of prev-curr. The full, three-way interaction model was as follows: $\theta_e \sim \text{coil location} \times \text{intensity} \times \text{prev-curr} + (1 + \text{prev-curr}|\text{subject})$

Decoding stimulus information. *Monkeys.* Population decoder. For each recorded ensemble, we decoded stimulus θ_j in trial j by modeling it as a linear combination of the spike counts n_{ij} ($i = 1 \ldots k$) of k simultaneously recorded neurons, computed in sliding windows of 0.5 s and steps of 0.1 s during that trial (in all decoding time courses depicted in figures (monkeys and humans), time (x axis) coordinates mark the central value of the corresponding sliding window):

$$\cos(\theta_j) \sim 1 + \sum_i^k \beta_i n_{ij} \quad \text{and} \quad \sin(\theta_j) \sim 1 + \sum_i^k \omega_i n_{ij}$$

For each set of neurons, we trained two sets of weights $\{\beta_i\}$ and $\{\omega_i\}$ on 80% of randomly selected trials and tested in the remaining trials. We applied Monte–Carlo cross-validation with 50 random splits to obtain angle estimates $\hat{\theta}_j$. We obtained a measure of error (err) by averaging across splits the mean absolute error ($|\hat{\theta}_j - \theta_j|$) in each split.

Accuracy of ensembles: distance from shuffle. To establish the significance of decoding accuracy (z), we compared the decoding error (err) for each ensemble to the distribution of decoding errors in 1,000 shuffled stimulus sequences (err_s). By shuffling the list of stimuli presented in the particular recording of each ensemble, we maintained the characteristics of the distribution (for example, unbalanced distribution of stimuli), but effectively destroyed correlations between stimuli and neural activity.

$$z = -\frac{\mathrm{err} - \mathrm{mean}(\mathrm{err}_s)}{\mathrm{s.d.}(\mathrm{err}_s)}$$

In Fig. 1c and Extended Data Fig. 1b, we separately tested ensembles that had the strongest and weakest decoding accuracy in the delay period by obtaining z from spike counts in the delay period and classifying the ensembles based on z: ensembles within the top tertile (high-decoding delay ensembles) and those in the bottom tertile (low-decoding delay ensembles).

Accuracy of single trials: leave-one-out decoder. To measure stimulus information on a trial-by-trial basis, we used leave-one-out cross-validation (Fig. 5a–c). We regressed the β_i and ω_i weights in all trials, except the one left out for testing. For these analyses we computed spike counts in windows of 1 s in steps of 50 ms.

Humans. Linear decoder. EEG alpha power is known to decrease in occipital sites contralateral to attended locations and for locations being actively maintained in working memory[33,61–63]. We used this feature to decode the angular position of the stimulus from the distribution of alpha power over all 43 electrodes. We trained the decoder on the stimulus label of the previous trial and decoded this information throughout the previous and current trial. Trial-wise alpha power for each electrode was modeled as a linear combination of a set of regressors representing the stimulus location in the corresponding trial, $U = WM$, where U is a $J \times K$ matrix of alpha power measured at electrode j in trial k, M is the $N \times K$ design matrix of values for regressor n in trial k, and W is the $J \times N$ weight matrix, mapping the weight for regressor n to electrode j. U and M were given by the experiment, while W was fitted using least squares.

The design matrix M is a set of eight regressors M_n representing expected "feature activations"[64] for feature n in trial k. The value of regressor M_n in trial k was determined as $\left|\sin(n\pi/8 - s_k\pi/8 + \pi/2)^7\right|$, where $s_k = [0 \ldots 7]$ indicates which one of eight angular location bins (width $\pi/8$ radians) included the stimulus shown in trial k.

As in the monkey analyses, we measured single-trial stimulus representations using leave-one-out cross-validation, ensuring an equal number of trials from each location bin in the training set (U_t and M_t). We estimated the weight matrix $\hat{W}$ and the design matrix $\hat{M}_k$ for the left-out trial k, as follows:

$$\hat{W} = U_\mathrm{t} M_\mathrm{t}^T \left(M_\mathrm{t} M_\mathrm{t}^T\right)^{-1}$$

$$\hat{M}_k = \left(\hat{W}^T \hat{W}\right)^{-1} \hat{W}^T U_k$$

For each trial and time point, we repeated this analysis 100 times with randomly chosen training sets (except for the temporal generalization matrix, for which ten repetitions were run, Fig. 2b), and averaged $\hat{M}$ over all repetitions. Finally, we estimated the predicted angle $\hat{\theta}_k$ as the direction of the vector sum of feature vectors with length $\hat{M}_{nk}$ pointing at angular location bin centers $\mathrm{b}_n = n\pi/8$ ($n = 0 \ldots 7$). Trial-wise decoding strength was then defined as $\cos\left(\hat{\theta}_k - \theta_k\right)$. To correlate the decoding strength with behavioral biases (Fig. 5d–f), we increased the stability of trial-wise measures by temporal averaging over moving 200-ms windows (x axis ticks in Fig. 5f are centered at window centers).

Cross-temporal decoding. To explore the temporal generalization of the mnemonic and the response code over time, we trained decoders in independent time windows of the previous and current trial, and tested them in all time points of consecutive trials (from 0.25 s to 1.25 s after previous stimulus onset (Fig. 2c, left), −0.25 s to 0.25 s after previous response (Fig. 2c, middle), and −1.25 s to 0.25 s after the stimulus onset of the current trial (Fig. 2c, right)). For the temporal generalization matrix (Fig. 2b), we averaged training and test data over independent windows of 50 samples (~97.77 ms). High-resolution time courses of mnemonic and response code (Fig. 2c) were obtained by training the decoder on averaged data from 0.5 s to 1 s after previous stimulus onset and −0.25 s to 0.25 s relative to the response time (dashed lines in Fig. 2b), respectively, and by testing on averaged data from five samples (~9.77 ms) through consecutive trials.

Preferred location. We computed the preferred locations of each neuron. Similar to ref. [6], the preferred location was determined by computing the circular mean of the cue angles (0–315°, in steps of 45°) weighted by the mean spike count of the neuron over the delay period (3 s) following each cue presentation.

Cross-correlations. *Dataset.* For the estimation of functional connectivity, we estimated cross-correlations by computing the jittered cross-covariances[65] of spike counts from simultaneously recorded neuron pairs, whose preferred locations were separated by a maximum of 60° ($n = 67$). We included pairs of neurons recorded from the same electrode ($n = 21$) and pairs recorded from different electrodes ($n = 46$). For each pair, we selected those trials in which the presented cue fell within the preferred range (pref, within 40° from either preferred locations) or outside the preferred range (anti-pref, all the other trials). We discarded those trials without at least one spike for each neuron in the pair.

Jittered cross-covariance. We used the Python function scipy.signal.correlate to compute cross-covariances between spike trains of simultaneously recorded pairs. Spikes were counted in independent windows of 10 ms[37,66]. For each trial, 1,000 jittered cross-covariances were computed as follows[65]. We shuffled the spike counts within non-overlapping windows of 50 ms and computed cross-covariance for each of these jittered spike counts. This captured all the cross-covariance caused by slow dynamics (>50 ms) but destroyed any faster dynamics. Finally, we removed the mean of these jittered cross-covariances from the cross-covariance of each trial, ending up with correlations due to faster dynamics (≤50 ms). We considered the magnitude of the central peak of the cross-covariance in our analyses by averaging 3 bins (±1 bin from the zero-lag bin). For the time-resolved cross-correlation function (Fig. 3c,d), we repeated this process for sliding windows of 1 s and steps of 50 ms, and averaged across trials and neuronal pairs.

Putative exc and inh interaction. Because changes in connectivity strength (our hypothesis for activity-silent mechanisms) affect inversely exc peaks and inh troughs of cross-correlations[34], we separately analyzed these two types of interactions. Similar to refs. [35,36], based on the average central peak of the cross-correlation function in the entire trial [−4.5 s, 2.5 s], we classified each pair into three subgroups: (1) those with a positive peak for both pref and anti-pref trials were classified as putative exc interactions, (2) those with a negative peak for both pre and anti-pref trials were classified as putative inh interactions and (3) we discarded those with an inconsistent peak sign between pref and anti-pref trials. In total, we analyzed the cross-correlation time course of $n = 47$ pairs of neurons ($n = 27$ exc and $n = 20$ inh; from different electrodes $n = 20$ exc and $n = 13$ inh). We confirmed that our results held when analyzing only pairs from different electrodes (Fig. 3c; exc: $P = 0.01$, $n = 20$; inh: $P = 0.04$, $n = 13$, one-sided permutation test).

Delay rate versus ITI cross-correlation analyses. As shown in Fig. 3e, we sought evidence for an interplay between attractor and subthreshold network dynamics in the PFC. To this end, we computed the trial-by-trial correlation between the cross-covariance peak (see above) in the ITI—at a time point when there was no firing-rate tuning (activity-silent period, Fig. 3d)—and the mean firing rate of the two neurons at the end of the preceding delay period (last 2 s, delay-fr, Fig. 3e) for exc interaction pairs under the pref and anti-pref condition (see above). For each pair, we obtained demeaned values for each trial by subtracting the mean firing rate and the mean cross-covariance peak across all trials, respectively. This allowed us to compute the correlation based on trial-by-trial measurements of all pairs together ($n = 27$) to increase statistical power. Error bars were then computed based on a bootstrap approach on all trials for all pairs. A local activity-dependent subthreshold mechanism for ITI memory traces predicts that for pref trials, but not for anti-pref trials, firing-rate variations in the delay period determines the degree of latent variable loading (cross-covariance peak) in the ITI (Fig. 3e).

Simulating bump reactivation. We used a previously proposed computational model[39,67,68] to study serial dependence between two consecutive trials. The model consists of a network of interconnected 2,048 excitatory and 512 inhibitory leaky integrate-and-fire neurons[69]. This network was organized according to a ring structure: excitatory and inhibitory neurons were spatially distributed on a ring so that nearby neurons encoded nearby spatial locations. All connections were all-to-all and spatially tuned, so that nearby neurons with similar preferred directions had stronger than average connections, while distant neurons had weaker connections. Inhibitory-to-inhibitory connections were untuned. Network parameters were taken from ref. [67] except for the following:

$$G_{\mathrm{EE,\,AMPA}} = 0.1\,\mathrm{nS},\ G_{\mathrm{EI,\,AMPA}} = 0.192\,\mathrm{nS}$$

$$G_{\mathrm{EE,\,NMDA}} = 0.42\,\mathrm{nS},\ G_{\mathrm{EI,\,NMDA}} = 0.49\,\mathrm{nS}$$

$$G_{\mathrm{II,\,GABA}} = 0.7413\,\mathrm{nS}, G_{\mathrm{IE,\,GABA}} = 0.9163\,\mathrm{nS}$$

$$g_{\mathrm{ext,\,I}} = 5.8\,\mathrm{nS},\ g_{\mathrm{ext,\,E}} = 5.915\,\mathrm{nS}$$

$$J^{+}_{\mathrm{EE}} = 7.1, \sigma_{\mathrm{EE}} = 18^{\circ}, J^{+}_{\mathrm{EI}} = J^{+}_{\mathrm{IE}} = 2.2, \sigma_{\mathrm{EI}} = \sigma_{\mathrm{IE}} = 32^{\circ}$$

where G values are the maximum conductances of the corresponding connections (e.g., $G_{\mathrm{EE,\,AMPA}}$ is the total maximum conductance of AMPAR-mediated local excitation onto an excitatory neuron), $g_{\mathrm{ext,E}}$ and $g_{\mathrm{ext,I}}$ are the maximum conductance of external Poisson inputs to an excitatory or inhibitory neuron, respectively, and J^{+} and σ values define the amplitude and width of corresponding connectivity footprints, respectively. See ref. [67] for more details.

STP dynamics. Simulation of activity-silent mechanisms during the inter-trial period was done by adding two more variables x and u, as described in refs. [9,70], to excitatory presynaptic neurons as follows:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1 - x}{\tau_x} - u\ x\ \delta\left(t - t_{\mathrm{sp}}\right)$$

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{U - u}{\tau_u} + U\left(1 - u\right)\ \delta\left(t - t_{\mathrm{sp}}\right)$$

With t_{sp} marking all spike times and $\delta(t)$ being the Dirac delta function. We used the parameters $U = 0.2, \tau_x = 200\,\mathrm{ms}, \tau_u = 1,500\,\mathrm{ms}$. The effective conductance of each excitatory synapse was then $g \times u \times x$, with g being the corresponding maximum conductance parameter (see above). These STP dynamics affected only AMPA-receptor-mediated recurrent connections in the network. In a separate set of network simulations (not shown), we also included STP in inhibitory connections in the network (same parameters as indicated above) and we found that we could obtain a similar pattern of serial bias modulations as shown in Fig. 4d. This shows that our results are not specifically dependent on whether inhibitory connections present facilitation dynamics or not.

Stimulation and behavioral readout. External stimuli were fed into the circuit as weak inputs (0.25 nA) to neurons selective to the stimulus as previously described[67]. Each simulation of our computational model consisted of two trials run in sequence: a first stimulus of 250 ms, a first delay period of 1,000 ms, a network resetting input (nonspecific current −0.261 nA, duration 300 ms), an ITI of 1,300 ms, a second stimulus (250 ms) and a second delay period of 1,000 ms. The first and second cue stimuli were independently drawn randomly from 360 uniformly distributed angular values, and only the network readout of the second trial was analyzed to obtain a 'behavioral' readout. The readout was obtained with a bump-tracking procedure: starting at cue presentation, the instantaneous network readout was derived as the angular direction of the population vector of single-neuron firing rates (computed in windows of 250 ms, sliding by 100 ms) considering the ±100 neurons surrounding the readout estimated in the previous time step. The instantaneous readout was iteratively derived to track the center of the bump (thus ignoring possible elevated activity extending from the fixation period), and the final behavioral output was defined as the readout in the last 250 ms of the trial. Serial bias was calculated by measuring single-trial errors (behavioral readout minus target location) in relation to the angular distance θ_{d} between the first and second stimulus locations, as described above for experimental data.

Consecutive trials and bump reactivation. Reactivation of the previous-trial stimulus during the reactivation period (300 ms before the second stimulus onset) was accomplished by stimulating all excitatory neurons with a nonspecific external stimulus[9,38]. This stimulus exponentially increased with a rate of $\alpha = 10\,\mathrm{s}^{-1}$ as $\beta(1 - e^{-\alpha(t - t_0)})$, with β being the reactivation strength and t_0 the time of onset of the stimulus. The reactivation strength was weak ($\beta = 0.17\,\mathrm{nA}$) or strong ($\beta = 2.9\,\mathrm{nA}$).

Rate and synaptic tuning. For each simulation shown in Fig. 3a,b, we computed the firing rate (r) and synaptic ($s = u \times x$) tuning by computing the difference between neurons within (±50°) and outside (180 ± 50°) the previous bump location for both measures.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data that support the findings of this study are available at https://github.com/comptelab/interplayPFC.

Code availability

The custom code used in this study is publicly available at https://github.com/comptelab/interplayPFC.

References

51. Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *J. Neurosci.* **21**, 3646–3655 (2001).

52. Compte, A. et al. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *J. Neurophysiol.* **90**, 3441–3454 (2003).
53. Constantinidis, C., Williams, G. V. & Goldman-Rakic, P. S. A role for inhibition in shaping the temporal flow of information in prefrontal cortex. *Nat. Neurosci.* **5**, 175–180 (2002).
54. Constantinidis, C. & Goldman-Rakic, P. S. Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *J. Neurophysiol.* **88**, 3487–3497 (2002).
55. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).
56. Wang, X. J., Tegnér, J., Constantinidis, C. & Goldman-Rakic, P. S. Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proc. Natl Acad. Sci. USA* **101**, 1368–1373 (2004).
57. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
58. Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A. & The Safety of TMS Consensus Group. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin. Neurophysiol.* **120**, 2008–2039 (2009).
59. Lumley, T., Diehr, P., Emerson, S. & Chen, L. The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* **23**, 151–169 (2002).
60. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-147 (2019).
61. Worden, M. S., Foxe, J. J., Wang, N. & Simpson, G. V. Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *J. Neurosci.* **20**, RC63 (2000).
62. Kelly, S. P., Lalor, E. C., Reilly, R. B. & Foxe, J. J. Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *J. Neurophysiol.* **95**, 3844–3851 (2006).
63. Medendorp, W. P. et al. Oscillatory activity in human parietal and occipital cortex shows hemispheric lateralization and memory effects in a delayed double-step saccade task. *Cereb. Cortex* **17**, 2364–2374 (2007).
64. Brouwer, G. J. & Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
65. Amarasingham, A., Harrison, M. T., Hatsopoulos, N. G. & Geman, S. Conditional modeling and the jitter method of spike resampling. *J. Neurophysiol.* **107**, 517–531 (2012).
66. Nougaret, S. & Genovesio, A. Learning the meaning of new stimuli increases the cross-correlated activity of prefrontal neurons. *Sci. Rep.* **8**, 11680 (2018).
67. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
68. Edin, F. et al. Mechanism for top-down control of working memory capacity. *Proc. Natl Acad. Sci. USA* **106**, 6802–6807 (2009).
69. Tuckell, H. C. *Introduction to Theoretical Neurobiology: Volume 2, Nonlinear and Stochastic Theories* (Cambridge Univ. Press, 1988).
70. Markram, H., Wang, Y. & Tsodyks, M. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl Acad. Sci. USA* **95**, 5323–5328 (1998).

Acknowledgements

This work was funded by the Spanish Ministry of Science and Innovation and the European Regional Development Fund (references BFU2015-65315-R and RTI2018-094190-B-I00); by the Institute Carlos III, Spain (grant PIE 16/00014); by the Cellex Foundation; by the "La Caixa" Banking Foundation (reference LCF/PR/HR17/52150001); by the Safra Foundation; by the Generalitat de Catalunya (AGAUR 2014SGR1265 and 2017SGR01565); and by the CERCA Programme/Generalitat de Catalunya. C.C. was supported by NIH grant R01 EY017077. J.B. was supported by the Spanish Ministry of Economy and Competitiveness (FPI program, reference BES-2013-062654) and by the Bial Foundation (reference 356/18). H.S. was supported by the "La Caixa" Banking Foundation (reference LCF/BQ/IN17/11620008) and the European Union's Horizon 2020 Marie Skłodowska–Curie grant (reference 713673). K.C.S.A. was supported by NIH grant T32-MH020002. We thank the Barcelona Supercomputing Center (BSC) for providing computing resources, and the Neurology Department of the Hospital Clínic de Barcelona for granting access to EEG, TMS and neuronavigation equipment. This work was developed at the building Centro Esther Koplowitz, Barcelona. We thank A. Morató and D. Lozano-Soldevilla for assistance with EEG analyses, L. C. García del Molino for valuable insights during the development of early versions of the model, and A. Renart and J. de la Rocha for their comments on the manuscript.

Author contributions

J.B. and A.C. performed the monkey data analyses. J.B. and A.C. developed the model. H.S. and A.C. designed the human EEG research. H.S. and A.G.-G. performed the human EEG experiments. H.S., J.B. and A.C. performed the human data analyses. A.C. and J.D. obtained the funding used for the human EEG research. K.C.S.A. performed the preliminary human EEG data analyses. J.B., R.L.M., J.V.-S. and A.C. designed the TMS experiments. R.L.M. performed the TMS experiments and performed the data analyses. S.L. performed the monkey experiments. C.C. designed the monkey research. J.B., H.S. and A.C. discussed the results and wrote the manuscript. All authors revised the manuscript and gave critical comments.

Competing interests

J.D. receives royalties from Athena Diagnostics for the use of Ma2 as an autoantibody test and from Euroimmun for the use of NMDA as an antibody test. He received a licensing fee from Euroimmun for the use of $GABA_B$ receptor, $GABA_A$ receptor, DPPX and IgLON5 as autoantibody tests; he has received a research grant from Sage Therapeutics.

Additional information

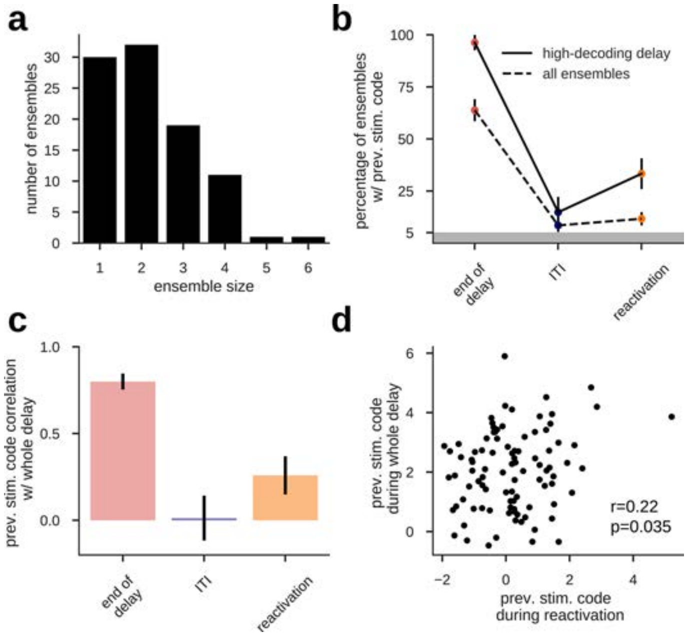
Extended data is available for this paper at https://doi.org/10.1038/s41593-020-0644-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-0644-4.

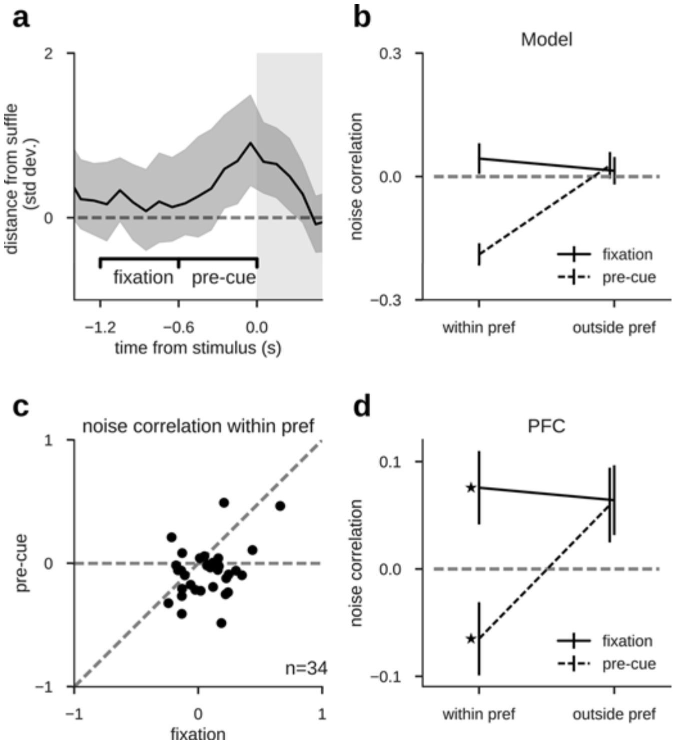
Correspondence and requests for materials should be addressed to A.C.

Peer review information *Nature Neuroscience* thanks Bradley Postle and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

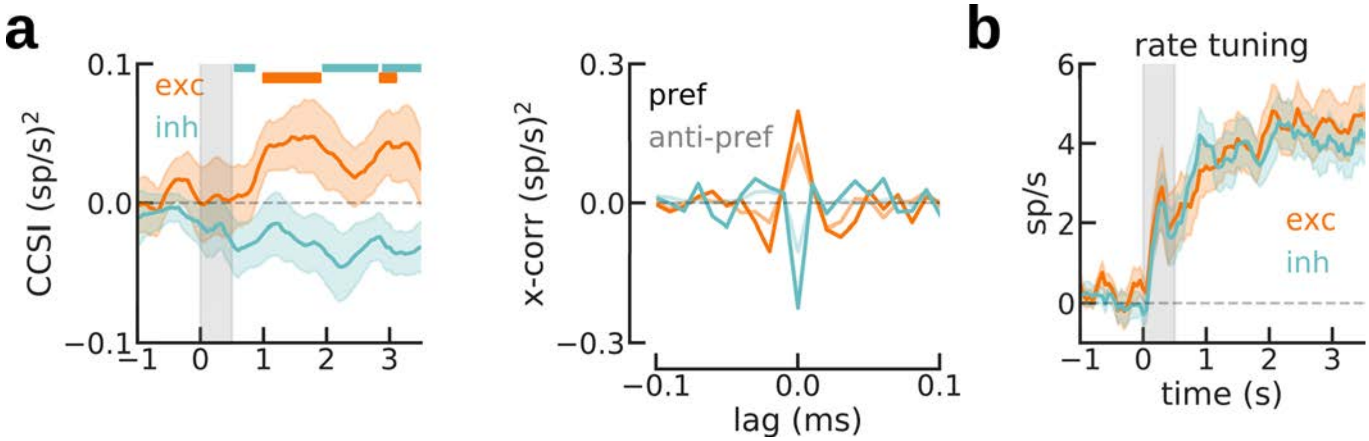
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Consistent decoding accuracy in delay and reactivation links these two representations at the neural ensemble level. a, The size of n=94 independent ensembles of simultaneously recorded neurons varies between 1-6. **b**, Fraction of neural ensembles with significant previous stimulus decoding accuracy ($z > 1.96$, see Methods) computed for all ensembles (dashed line) and only for those ensembles with strongest previous stimulus code averaged across the whole delay (see Methods). The incidence of stimulus decoding was significant in delay and reactivation, but not at ITI (two-sided binomial test at p=0.05, with n=94 and n=27 ensembles, for 'all ensembles' and 'highest delay code', respectively). Error bars are bootstrapped ±s.e.m. **c**, across-ensemble Pearson correlation between delay decoding accuracy (averaged in the entire delay) and decoding accuracy at different time points (two-sided p-values: 6.5e-30, 0.87, 0.035, n=94 ensembles). The ensembles with strongest delay code also had stronger decoding during reactivation, demonstrating the neural association between delay representations and reactivations despite absent code in the ITI. Error bars denote ±s.e.m. computed with a bootstrap procedure. **d**, Individual ensemble values from c, orange (Pearson correlation, two-sided p=0.035, n=94 ensembles).

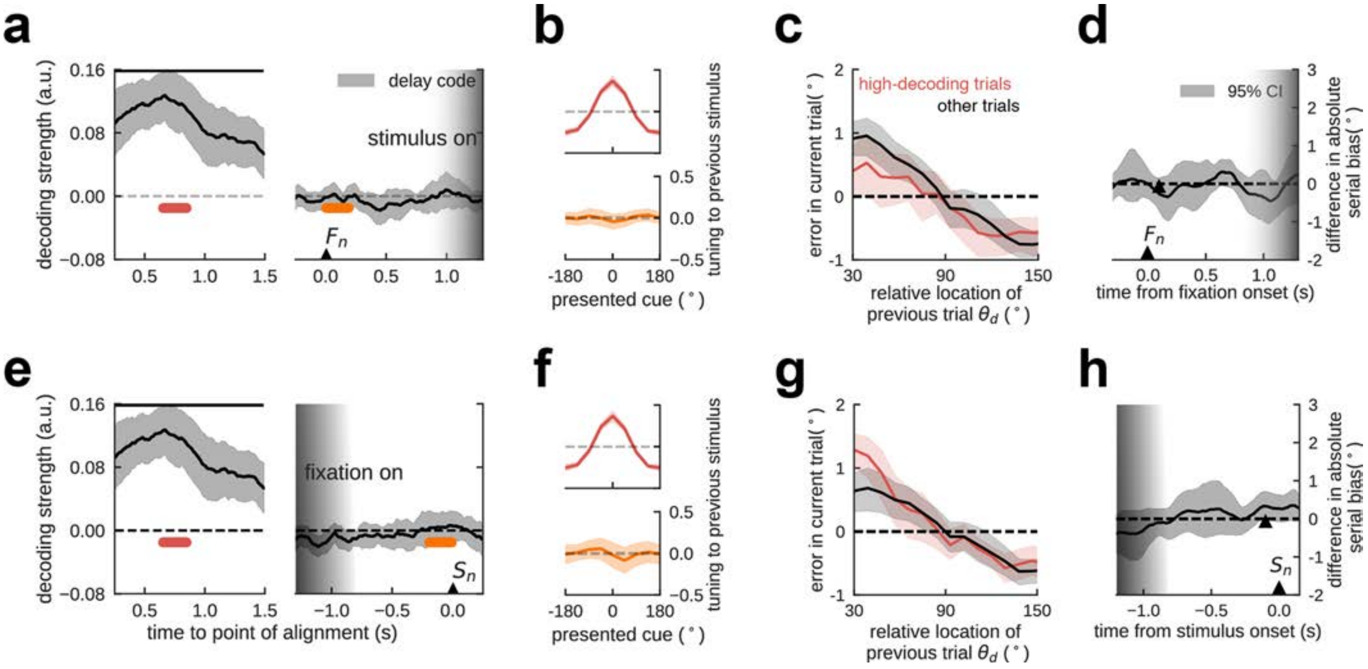


Extended Data Fig. 2 | Noise correlation between pairs of neurons is negative at reactivation, as predicted by the attractor model. Bump-attractor dynamics are characterized by negative pairwise noise correlations for cues presented between the preferred locations (*within pref*) of the two neurons, but not for other cues (*outside pref*) 6. **a**, Periods used in noise correlation analyses: early (*activity-silent*), and late fixation (*reactivation*; n=94 ensembles, zoom-in of Fig. 1c). Error shading, bootstrapped 95% C.I. **b**, In the computational model (n=1,000 independent simulations), bump reactivations from subthreshold traces are characterized by negative noise correlations only during reactivation for *within-pref* trials, following the nonspecific input drive (Fig. 4). **c**, Noise correlations of PFC pairs with dissimilar preferred angles (60° < Δθ < 120°, n=34 pairs) were lower in late than in early fixation for *within-pref* trials (bootstrap test, p=0.0001, n=34, Cohen's d=0.61). **d**, On average, lower noise correlations occurred only during reactivation and in *within-pref* trials (ANOVA *trial condition x time point*, F(4)=2.5, p=0.06, n=34). For *within-pref* trials, noise correlations differed between early and late fixation (bootstrap test, p=0.0001, Cohen's d=0.61, n=34), being negative in late (bootstrap test, p=0.035, Cohen's d=-0.32, n=34), but positive in early fixation (bootstrap test, p=0.018, Cohen's d=0.37, n=34). Correlations were positive in *outside-pref* trials both during late and early fixation (bootstrap test, p=0.024 and p=0.06, respectively), with no significant difference (two-sided bootstrap test, p=0.93, n=34). In addition, negative noise correlations diminished when using the previous saccade location rather than the previous stimulus as reference (paired bootstrap test, p=0.005, Cohen's d=-0.47, n=34), suggesting that the bump diffused only during the delay period, but not after the saccade 6. Unless stated otherwise, all bootstrap tests were one-tailed in the direction of the model predictions in b. All error bars indicate ±s.e.m.

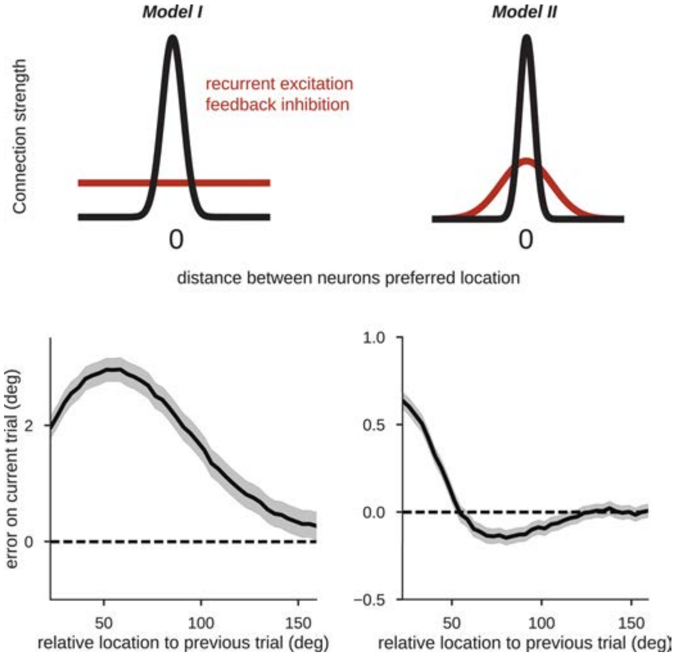


Extended Data Fig. 3 | Stimulus selectivity in both cross-correlation peaks and firing rates during the delay period prevents the isolation of activity-based and activity-silent processes. Same analysis as in Fig. 3, but performed during the current delay period (instead of ITI, Fig. 3) and selecting pref and anti-pref trials based on current stimulus (instead of previous, Fig. 3). Note that these are different trials (no need to be consecutive), so *exc* (n=33 pairs) and *inh* (n=21 pairs) might differ from Fig. 3. **a**, Left, cross-correlation peak selectivity emerged and was sustained in the delay period (left, CCSI as in Fig. 3, computed in centered 500-ms windows sliding in steps of 50 ms) and consisted in enhanced central peaks (troughs) for *exc* (*inh*) following a preferred stimulus. Color bars mark the periods where the average CCSI is different from 0 (bootstraped 95% C.I.) Right, cross-correlation averaged over 0.5-3.5 s. Zero-lag correlation for pref and anti-pref are different in exc (p=0.03, n=33, two-sided paired bootstrap test) and inh (p=0.01, n=21, two-sided bootstrap test) conditions. **b**, Firing rate selectivity (pref - anti-pref) also emerges robustly in the delay period for neurons in *exc* and *inh* pairs. The selectivity in cross-correlation peaks (CCSI) can therefore be confounded with firing rate selectivity[71] when analyzing data in the delay period. This prevents the unambiguous identification of activity-silent mechanisms in this task period. Our approach of analyzing data in the inter-trial interval, when there is no firing rate selectivity (Fig. 3f), gets around this problem. Gray shading marks the stimulus presentation. In all panels, error-bar shadings indicate ±s.e.m.

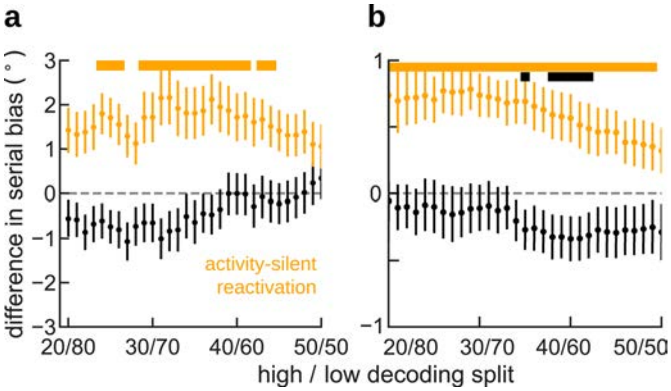
71. de la Rocha, J., Doiron, B., Shea-Brown, E., Josić, K. & Reyes, A. Correlation between neural spike trains increases with firing rate. *Nature* **448**, 802–806 (2007).



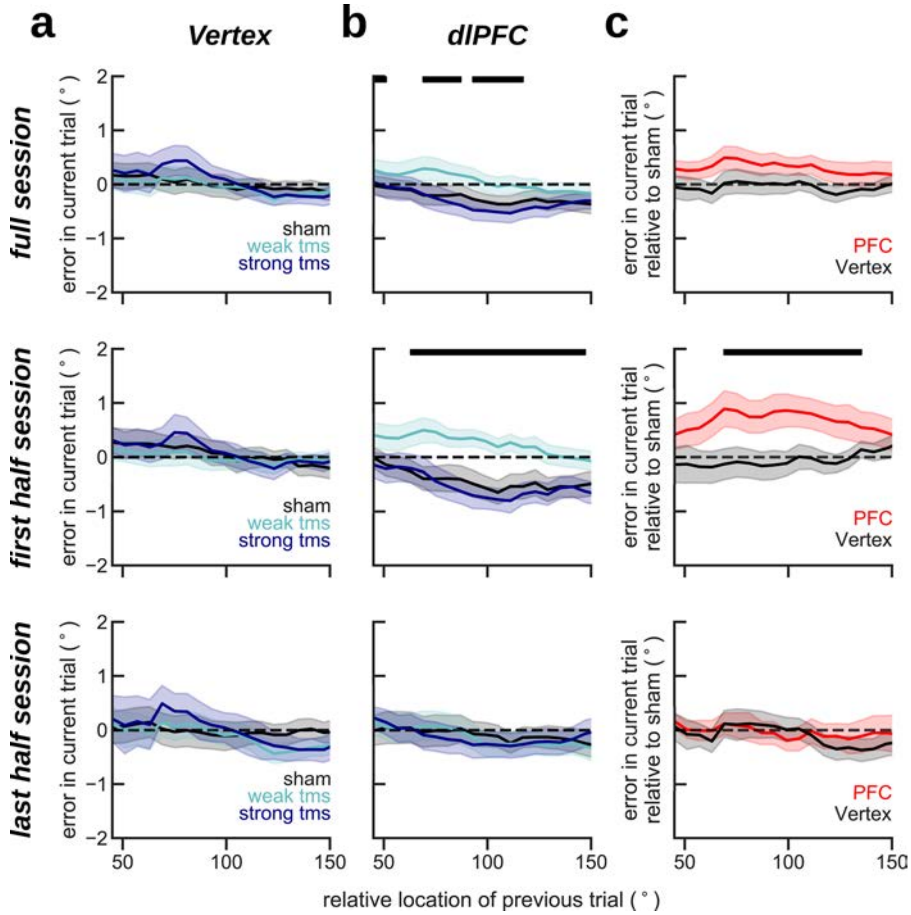
Extended Data Fig. 4 | In a dataset with unpredictable stimulus-onset time, previous item representations were not reactivated in the pre-stimulus period. We conducted the same analysis as in human EEG (Fig. 2) in a previously published dataset (n=15 independent subjects for all panels; for experimental details, please refer to the original publication, ref. [33]) with unpredictable fixation period durations (range 0.7 s-1.3 s). Decoding analyses were applied separately for data aligned to the onset of fixation (*Fn*, graded shading indicates range of possible stimulus onset times, upper panels) and aligned to the onset of the stimulus (*Sn*, graded shading indicates possible fixation onset times, lower panels). **a**, Tuning to previous-trial location (decoder trained in delay, 0.5s - 1.0s after stimulus onset) during previous-trial delay (left, stimulus aligned) vanishes in current-trial fixation (right, fixation onset aligned). No reactivation occurs. **b**, Average tuning reconstruction at different epochs for the delay decoder, indicated in **a**. **c**, Serial dependence separating trials with high (red curve, top quartile) from all other trials' (black curve) decoding accuracy in early fixation (orange in **a**). Unlike in an experiment with predictable stimulus onset (Fig. 5), serial bias did not differ as a function of decoding strength. **d**, Difference in serial biases (Methods) between *high-decoding* and *other* trials were not significant at any time point in fixation. The black triangle marks the center of 0.2 s decoding window for the split in **c**. **e**-**h**, Parallel results were obtained when the analyses of panels **a**-**d** were run on data aligned to the time of stimulus onset instead of fixation onset. In **d** and **h**, time courses were smoothed using a squared filter of 5 samples. Periods with significant decoding in **a**,**e** are marked with black horizontal bars, indicating p<.001 in a two-sided bootstrap test. Shading indicates 95% C.I. in **a**,**d**,**e**,**h**, and ±s.e.m. in **b**,**c**,**f**,**g**.



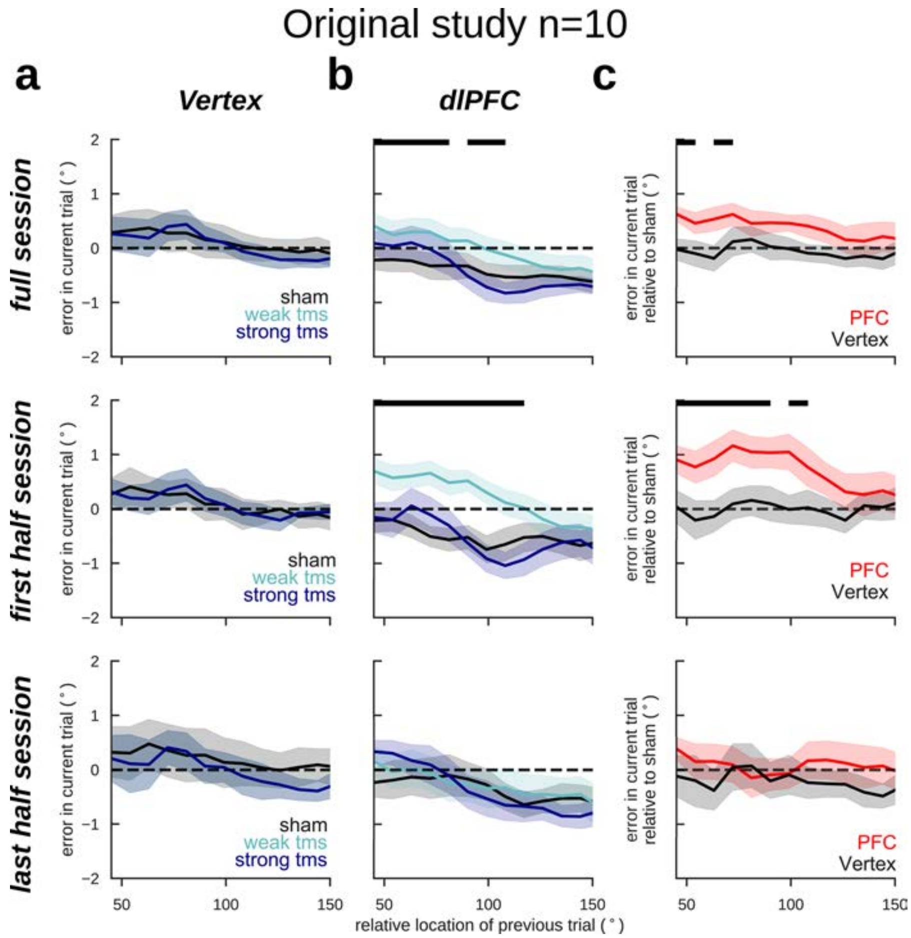
Extended Data Fig. 5 | Structured inhibition is necessary for repulsive serial biases at far distances. Top panel, illustration of two different models that have different inhibitory connectivity profiles. On the left, inhibitory connectivity strength from inhibitory to excitatory neurons is similar for all distances between their preferred locations. On the right, inhibition is structured such that similarly tuned neurons have stronger feedback inhibition. This shows that repulsive biases are caused by repulsive interactions between simultaneously active bumps in the network[39,40], and are absent when there is no reignited bump that recruits localized inhibition at the flanks of the pre-cue bump of activity.



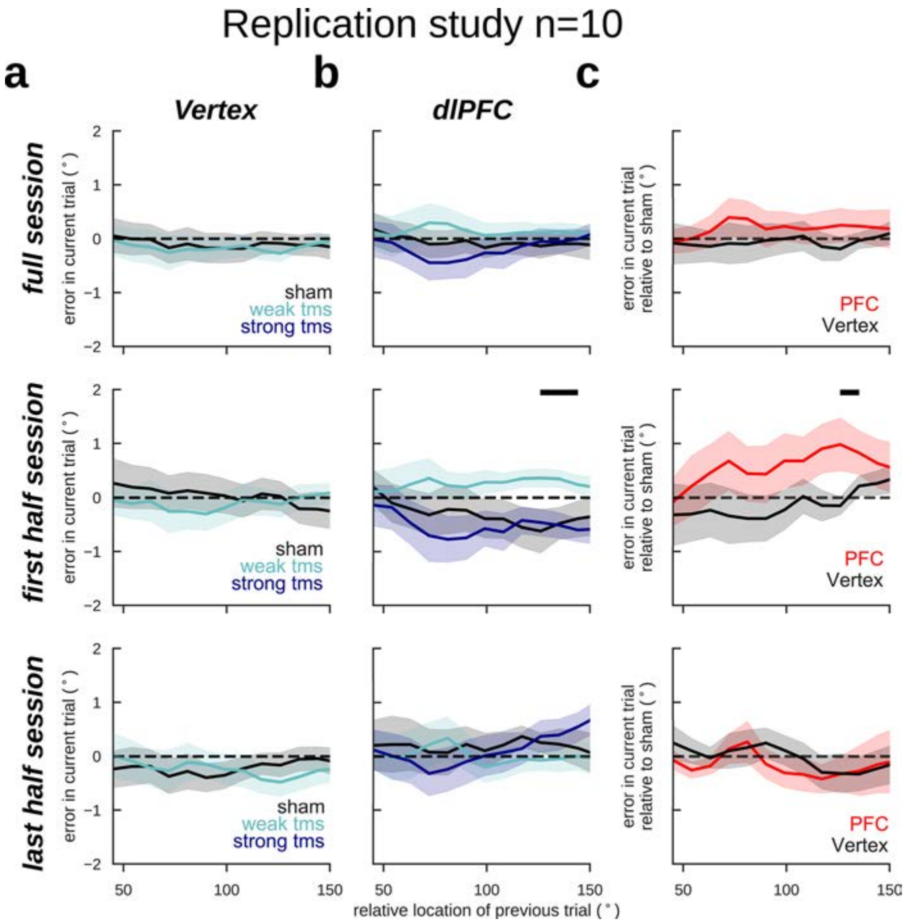
Extended Data Fig. 6 | Serial bias split between high-decoding and other trials (Fig. 5) is robust to the choice of different percentiles. a, In monkey behavior **b**, In human behavior. X-axis indicates quantiles used for the split in high- and low-decoding trials (Fig. 5), from a total of n=1362 trials in **a**, and a range of 792-908 trials per subject in **b**. Error bars are ±s.e.m. (over n=1362 trials in **a**, and over n=15 subjects in **b**) and colored bars mark where corresponding difference in serial biases is different than zero (p<0.05, two-sided bootstrap test).



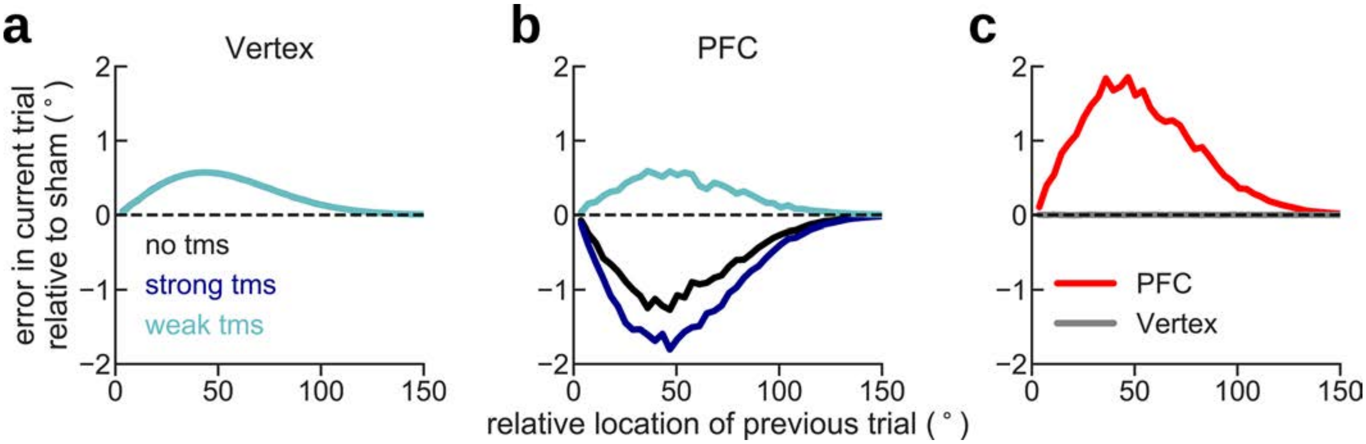
Extended Data Fig. 7 | The effect on serial biases of targeting dlPFC with TMS diminishes in the course of the experimental session. Serial bias plots averaged across n=20 independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same analyses as in Fig. 6, but (top) analyzing trials from the full session, (middle) first half session (225 trials, replication of Fig. 6) and (bottom) last half session (225 trials). The behavioral impact of PFC TMS stimulation declined through the session, as if subjects desensitized (*prev-curr* × *TMS intensity* × *session-half* $t_{11083} = -2.38$, p = 0.017. Methods, *Linear Mixed Models*). Serial biases were modulated by TMS in PFC, but not in Vertex (*prev-curr* × *TMS intensity* × *coil location*, $t_{18272} = 2.21$, p = 0.027. For dlPFC: *prev-curr* × *TMS intensity*, $t_{11087} = 2.13$, p = 0.032. For Vertex: $t_{7166} = 0.03$, p = 0.97. Methods, *Linear mixed models*) when analyzing the full session, and analyzing only the first half session ($t_{9133} = 2.51$, p = 0.011). x-axis coordinates mark the central value of windows (π/2 radians, sliding by π/30 radians) used to calculate behavioral biases.



Extended Data Fig. 8 | Consistent fixation-period single-pulse TMS effects on serial biases: first experiment. Serial bias plots averaged across n=20 independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same as Extended Data Fig. 6, but only analyzing data from the original study (n=10 subjects). Similarly to when pooling both the original and replication studies together, the behavioral impact of PFC TMS stimulation declined throughout the session, however not significantly (*prev-curr* × *TMS intensity* × *session-half* $t_{5701} = -1.73$, $p = 0.08$. Methods, *Linear Mixed Models*). Serial biases were modulated by TMS in PFC, but not in Vertex ($t_{5705} = 1.92$, $p = 0.05$) when analyzing the full session, and analyzing only the first half session ($t_{3059} = 2.59$, $p = 0.009$, Methods). x-axis coordinates mark the central value of windows (π/2 radians, sliding by π/30 radians) used to calculate behavioral biases.



Extended Data Fig. 9 | Consistent fixation-period single-pulse TMS effects on serial biases: replication experiment. Serial bias plots averaged across n=20 independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same as Extended Data Fig. 6 and 7, but only analyzing data from the pre-registered (https://osf.io/rguzn/) replication study (n=10 subjects). Similarly to the original experiment, the behavioral impact of PFC TMS stimulation declined throughout the session, however not significantly (prev-curr × TMS intensity × session-half $t_{5375} = -1.63$, p = 0.1. Methods, *Linear Mixed Models*). Similarly to the original study, serial biases were more strongly modulated by TMS in PFC than in Vertex, however not significantly ($t_{5379} = 1.12$, p = 0.25) when analyzing the full session and the effect was stronger when analyzing only the first half-session ($t_{2675} = 1.91$, p = 0.06, Methods). x-axis coordinates mark the central value of windows (π/2 radians, sliding by π/30 radians) used to calculate behavioral biases.



Extended Data Fig. 10 | A phenomenological model of our hypothesis on how long-term physiological effects of single TMS pulses affect serial bias curves in event-related experimental sessions. Our TMS results show a difference between the effects of sham stimulation at the vertex and sham stimulation over dlPFC (Fig. 6). We interpret this baseline difference as the possible effect of long-term physiological alterations by single pulses 58 (but see ref. [72]) that carry over from "strong-tms" trials to "no-tms" trials. We explicitly implemented this interpretation in the following way: we generated trial-by-trial responses biased depending on the sequence of stimuli according to a given baseline serial bias curve (**a**, "Vertex" condition where TMS is ineffective). In the "PFC" condition the serial bias strength changed depending on TMS conditions: in "weak-tms" trials the pulse had the acute effect of increasing the bias strength momentarily by an additive factor (3 times the baseline bias strength), in "strong-tms" trials the effect of the pulse was chronic: the bias changed with a negative additive component (equal in magnitude to the baseline strength), which decayed slowly through subsequent trials (10% decay/trial). When collapsing together "responses" obtained on the basis of this model through a sequence of randomly selected "no-tms", "weak-tms" and "strong-tms" trials, serial bias curves showed the pattern observed experimentally, where sham ("no-tms") trials show repulsion in the "PFC" condition (panel **b**) and not in the "Vertex" condition (panel **a**). The difference of serial bias curves for "weak-tms" and "no-tms" then showed the modulation clearly in "PFC" and not in "Vertex" (panel **c**), as seen in the data (Fig. 6).

72. Romero, M. C., Davare, M., Armendariz, M. & Janssen, P. Neural effects of transcranial magnetic stimulation at the single-cell level. *Nat. Commun.* **10**, 2642 (2019).

nature research

Corresponding author(s): Albert Compte

Last updated by author(s): Apr 3, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	EEG recordings used Deltamed Coherence Software version 5.1. Behavioral experiments with humans were programmed in Python 2.7 using Psychopy version 1.82.01.
Data analysis	Data were analyzed using custom scripts in Python 2.7 (monkey and TMS data) and in Python 3.7.4 (human EEG data). EEG data was pre-processed using Fieldtrip (version 20171231) in MATLAB R2017b and R2019a. The custom code used in this study is publicly available at https://github.com/comptelab/interplayPFC.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data that support the findings of this study are available at https://github.com/comptelab/interplayPFC.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the analysis of monkey data we were not able to predetermine sample sizes because this was data acquired previously (Constantinidis et al 2001). For human data, sample sizes were based on relevant prior literature. In the case of the EEG study, we matched the sample size (n=15) to the one used in a previous study that successfully decoded memory contents from EEG in an identical task (Foster et al. 2015). In the case of the TMS study, we predetermined the sample size (n=10) considering that TMS-induced memory reactivations had been shown in a previous study with 6 participants (Rose et al. 2016). We validated the results in a replication experiment with the same sample size (n=10).
Data exclusions	* No monkeys were excluded from the analysis. In the EEG study, one participant aborted because of physical discomfort. Another participant repeated the session on a different day because they aborted their first session with too few trial blocks. For this participant we only analyzed session 2. In the TMS study, one participant dropped the study when acquiring her MRI because she suspected pregnancy. * For neural data analyses, we excluded neurons without significant tuned delay activity. This was because of the hypothesis of our study (we wanted to explore the interaction between persistent and activity-silent mechanisms) and was predetermined in this study, as in other previous studies with this dataset (Constantinidis et al 2001; Compte et al. 2003; Wimmer et al. 2014). * For behavioral analyses, we excluded trials where behavioral reports were too far from the target to remove guess trials that may have not engaged working memory. For monkeys, this was done directly at acquisition time and could not be predetermined for this study (criterion report more than 20 degrees away from target). For humans, we excluded trials with responses further than 1 radian from targets in the angular direction and further than half the radius (2.25cm) in the radial direction. * For EEG analyses, we excluded outlier trials based on the voltage trace variance and alpha-power variance over each session. This is customary practice to remove EEG artifacts. Specific thresholds were set at the time of pre-processing of the data prior to final analyses.
Replication	We designed a replication study for the TMS experiment, to test the bias-enhancing effects of weak TMS stimulation and the disappearance of the effects as the session progressed. The methods, hypotheses and even the analysis codes for this replication study were pre-registered (https://osf.io/rguzn) prior to acquiring the data. Methods were applied as literally pre-determined and the results were parallel to our previous findings, validating our results. In the manuscript we report the aggregated data (participants were independent for the 2 studies), as well as the individual data for each experiment (supplementary data).
Randomization	Our study had a within-subject design, so randomization of participants across groups is not relevant for the study. Conditions of interest were typically randomized in our design: cue locations were pseudo-randomly chosen in monkey studies, and both cue locations and delay lengths were random in human EEG studies. For TMS experiments, cue locations and TMS intensity were random during experimental blocks, and TMS coil location was kept constant in each block and alternated from block to block, the order being counterbalanced in the 2 sessions of the same participant.
Blinding	Blinding was not necessary in regard to participants because this was a within-subject design with randomized task contingencies. For the TMS study, the experimenter could not be blind to the location of the coil.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Monkey subjects were four adult male rhesus macaques. Two of the animals were tested 20 years ago, when age reporting was

Laboratory animals	not customary. From their reported weights (Constantinidis et al. J. Neurosci. 21:3646, 2001) they were fully grown adults, so we can estimate the age at more than 6 years old. The ages of the other two animals reported in the study (with only behavioral data) were both 9 years old.
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All experiments were conducted in accordance with the guidelines set forth by the US National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee, and by the Wake Forest University Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	We studied healthy controls. The study does not address any specific covariate of interest across individuals, but within-subject comparisons between trial types.
Recruitment	Participants were recruited from a volunteer database, mostly including people associated with the research institute and hospital, in all cases naïve to this study.
Ethics oversight	Research Ethics Committee of Hospital Clínic (Barcelona)

Note that full information on the approval of the study protocol must also be provided in the manuscript.