Experimental Brain Research
https://doi.org/10.1007/s00221-024-06785-z

RESEARCH ARTICLE

Taking time to compose thoughts with prefrontal schemata

Kwang Il Ryom[1] · Anindita Basu[1] · Debora Stendardi[2] · Elisa Ciaramelli[2] · Alessandro Treves[1]

Received: 24 July 2023 / Accepted: 16 January 2024

Abstract
Under what conditions can prefrontal cortex direct the composition of brain states, to generate coherent streams of thoughts? Using a simplified Potts model of cortical dynamics, crudely differentiated into two halves, we show that once activity levels are regulated, so as to disambiguate a single temporal sequence, whether the contents of the sequence are mainly determined by the frontal or by the posterior half, or by neither, depends on statistical parameters that describe its microcircuits. The frontal cortex tends to lead if it has more local attractors, longer lasting and stronger ones, in order of increasing importance. Its guidance is particularly effective to the extent that posterior cortices do not tend to transition from state to state on their own. The result may be related to prefrontal cortex enforcing its temporally-oriented schemata driving coherent sequences of brain states, unlike the atemporal "context" contributed by the hippocampus. Modelling a mild prefrontal (vs. posterior) lesion offers an account of mind-wandering and event construction deficits observed in prefrontal patients.

Keywords Associative memory · Cortical networks · Latching dynamics · Sequential retrieval · Mind-wandering · Spontaneous cognition

Constructive associative memories

Recent explorations of the mechanisms underlying *creative* forms of human cognition (Mekern et al. 2019; Benedek et al. 2023), ranging from musical improvisation (Beaty 2015) through visual creativity (Aziz-Zadeh et al. 2013) up to poetry (Stockwell 2019), or mere mind wandering (Ciaramelli and Treves 2019), have again questioned the validity of reducing the cortex to a machine operating a complex transformation of the input it currently receives. On one hand, sophisticated and massive artificial intelligence systems like ChatGPT or midJourney, with their impressive performance, have adhered to the standard operational paradigm of producing a response to a query. On the other, a simple observation of cortical circuitry, with its extensive recurrence and quantitatively limited external inputs, have long ago led to the proposal that the cortex is (largely) a machine talking to itself (Braitenberg and Schüz 1991). Likewise, when confronted with an artistic or literary creation we sometimes ask: what was the query? Was there a query?

If it is the cortex itself that takes the initiative, so to speak, is it the *entire* cortex?

Understanding the mechanisms of cortico-cortical dialogue that generate spontaneous behaviour cannot eschew their statistical character, that of a system with very many imprecisely interacting elements. Valentino Braitenberg suggested a framework for such a statistical analysis, which to a first approximation considers the cortex as a homogeneous structure, not differentiated among its areas (nor, other than quantitatively, among mammalian species) (Braitenberg 1978): the only distinction is between long-range connections and local ones—those which reach in the immediate surround of the projecting neuron and do not travel through the white matter. Importantly, by asking whether there is any computational principle other than just associative memory operating at both long-range and local synapses (Braitenberg and Schüz 1991), Braitenberg pushes the age-old debate of whether cortical activity is more like a classic orchestra led

Dedicated to our dear colleague and friend Francesca Frassinetti, who had convened the *Brainstorm Time* meeting that stimulated this study.

Communicated by Massimiliano Oliveri.

✉ Alessandro Treves
ale@sissa.it

[1] SISSA - Cognitive Neuroscience, via Bonomea 265, 34136 Trieste, Italy

[2] Dip. Psicologia Renzo Canestrari, Univ. Bologna, Viale C. Berti-Pichat 5, 40126 Bologna, Italy

Published online: 14 March 2024

by a conductor or more like a jazz jam session, beyond the limits of abstract information-processing models. In traditional box-and-arrows models of that kind, a box, whether it represents a specific part of the brain or not, can operate any *arbitrary* transformation of its input, which makes it difficult to relate it to physiological measures, and tends to leave the debate ill-defined. If at the core one is dealing solely with associative memory, instead, the issue can be approached with well-defined formal models, generating statistical insights that can be later augmented with cognitive qualifications.

Given the canonical cortical circuit (Douglas et al. 1989) as a basic wiring plan for the generic cortical plaquette, or patch, getting at the gist of how it contributes to the exchanges mediated by long-range cortico-cortical connectivity among different patches requires considering the fundamental aspects that vary, at least quantitatively, among the areas. A number of reviews (Finlay and Uchiyama 2015; Hilgetag et al. 2022) have pointed out that several prominent features align their gradients of variation, across mammals and in particular in the human brain, along a *natural cortical axis*, roughly from the back to the front of the cortex. Actual observations and measurements may be incomplete or even at variance with such a sweeping generalization, but here we take it as a convenient starting point. Anatomical measures point at more spines on the basal dendrites of pyramidal cells, indicating more local synaptic contacts in temporal and especially frontal, compared to occipital cortex (Elston et al. 2001). This may support a capacity for more and/or stronger local attractor states. More linear and prompt responses to afferent inputs in posterior cortices, e.g. visual ones (Miller et al. 1996; Rotshtein et al. 2005), also suggest reduced local feedback relative to more anterior areas.

The rapidity of the population response to an incoming input has been related to the notion of an intrinsic *timescale* that might characterize each cortical area, and that may produce highly non-trivial effects, for example when inhibiting a particular area with TMS (Cocchi et al. 2016). The timescales measured with similar methods have been shown to differ considerably, even within individual areas (Cavanagh et al. 2020), and to define distinct cortical *hierarchies*, when extracted in different behavioural states, e.g. in response to visual white noise stimuli (Chaudhuri et al. 2015) or during free foraging (Manea et al. 2023). Thus, it remains unclear whether the ambition to define a unique hierarchy of timescales can really be pursued (Gao et al. 2020), and whether they can be related to patterns of cortical lamination (Barbas and Rempel-Clower 1997) and to biophysical parameters, including the I_h current and others underlying firing rates and firing frequency adaptation (Chang et al. 2005). Still, in broad terms multiple timescale hierarchies do roughly align with the natural axis, from faster in the back to slower in the front of the brain, and ignoring a factor of, say, four (Gao et al. 2020) would appear to grossly overlook a basic principle of cortical organization.

Here, we ask what are the implications of major differences in *cortical parameters* for how basic associative memory mechanisms may express cortically initiated activity. We focus on a simple differentiation between a posterior and a frontal half of the cortex, and neglect finer distinctions, e.g., rostrocaudal hierarchies within prefrontal cortex (Koechlin et al. 2003; Badre 2008) or the undoubtedly major differences within posterior cortices.

A simply differentiated Potts model

The mathematically defined model we use is based on the abstraction of a network of $\sqrt{\mathcal{N}}$ patches of cortex (where $\mathcal{N}$ are all its pyramidal cells), interacting through long-range, associatively modified synapses, an abstraction close to that informing *connectome* research (Roe 2019). Each patch would be a densely interconnected network of $\sqrt{\mathcal{N}}$ pyramidal cells interacting through local synapses, also associatively modifiable according to some form of Hebbian plasticity. Such a local cortical network may operate as an autoassociative memory once it has acquired through learning a number S of attractor states. In the simplified Potts formulation adopted here, the local network realized in each patch is replaced by a Potts unit with S states, and the analysis can focus on the network of long-range effective interactions between Potts units, which are no more mediated by simple synaptic connections, rather the connections are mathematically expressed as tensors (Naim et al. 2018).

We refer to previous studies (Ryom et al. 2021) and to Appendix A for a description of the standard model and of its key parameters. Suffice here to note that while the number S of local attractor states measures the range of options available for the dynamics of a patch of cortex, the feedback coefficient w quantifies how deep those options are, i.e., how strongly the patch is driven to choose one of them, and the adaptation time constant τ_2 parametrizes the time it takes for it to be eventually eased out of its current attractor.

A network of Potts units can express spontaneous behaviour when it *latches*, i.e., it hops from a quasi-stationary pattern of activity to the next, in the absence of external input—of a query (Treves 2005). Latching dynamics are a form of iterated associative memory retrieval; each extended activity pattern acts briefly as a global cortical attractor and, when destabilized by the rising thresholds which model firing rate adaptation, serves as a cue for the retrieval of the next pattern. Studies with brain-lesioned patients indicate, however, that there is structure in such spontaneous behaviour. In studies of mind-wandering, for example, patients with lesions to ventromedial prefrontal cortex (vmPFC) show reduced mind-wandering, and their spontaneous thoughts

tend to be restricted, focused on the present and on the self, suggestive of a limited ability to project coherently into the future (Bertossi and Ciaramelli 2016).

We then take our standard, homogeneous Potts network, differentiate it in two halves, and ask whether a structure of this type may reflect a basic differentiation between frontal and posterior cortices in the number or in the strength of their local attractor states, or in the time scale over which they operate, as expressed in differences, in the model, in the three relevant parameters, ΔS, Δw and $\Delta \tau_2$.

We assume that the two sub-networks store the same number p of memory patterns (with the same sparsity a), and that all the connections already encode these p patterns, as a result of a learning phase which is not modelled. We have seen in a previous study (Ryom and Treves 2023) that a differentiation ΔS has important dynamical implications during learning itself, but here we imagine learning to have already occurred. For a statistical study, we take the activity patterns to have been randomly generated with the same statistics, therefore, any correlation between pattern μ and ν is random, and randomly different if calculated over each sub-network. These restrictive and implausible assumptions—they discard for example the possibility of structured associations between frontal and posterior patterns of different numerosity, statistics and internal non-random correlations—are needed to derive solid quantitative conclusions at the level of network operation, and might be relaxed later in more qualitative studies.

Connectivity in the differentiated network

For the statistical analysis, carried out through computer simulations, to be informative, the structure of the network model and in particular its connectivity have to be chosen appropriately. First, each sub-network should have the same number of units (half the total) and each unit the same number of inputs, for the comparisons between different conditions to be unbiased by trivial factors. Second, each sub-network should be allowed to determine, to some extent, its own recurrent dynamics, which requires the inputs onto each unit from the two halves not to be equal in strength, which would lead to washing away any difference, effectively, at each recurrent reverberation.

We then set the connection between units i and j, in their tensorial states k and l, as

$$J_{ij}^{kl,\text{intra, inter}} = \frac{c_{ij}}{c_m a \sqrt{(1-\frac{a}{S_i})(1-\frac{a}{S_j})}} \sum_{\mu=1}^{p} \left(\delta_{\eta_i^\mu k} - \frac{a}{S_i}\right) \left(\delta_{\xi_j^\mu l} - \frac{a}{S_j}\right)(1-\delta_{k0})(1-\delta_{l0}), \quad (1)$$

where $\{c_{ij}\}$ is a sparsity $\{0,1\}$ matrix that ensures that Potts unit i receives c_m intra inputs from other units in the same sub-network and also receives c_m inter inputs from units of the other sub-network. Note that the number of Potts states of each unit, S, may depend on which sub-network the unit belongs to.

The partially differential dynamics is obtained by setting the strength coefficients as

$$J_{ij}^{kl} = \frac{(1+\lambda)}{2} J_{ij}^{kl,\text{intra}} + \frac{(1-\lambda)}{2} J_{ij}^{kl,\text{inter}}, \quad (2)$$

where the parameter $\lambda \in [-1,1]$ controls the relative strength of two terms. For $\lambda = 0.0$, the connectivity matrix becomes homogeneous and we cannot distinguish the two sub-networks from connectivity alone. If $\lambda = 1.0$, each sub-network is isolated from the other. For values of λ between 0 and 1, the recurrent connections within a sub-network prevail over those from the other sub-network, generating partially independent dynamics. We set $\lambda = 0.5$ as our reference value.

Results

We assume that the attractors of the frontal network have been associated one-to-one with those of the posterior network, via Hebbian plasticity, during a learning phase, which we do not model. When there is no external stimulus, e.g. when modelling creative thinking and future imaging, the network can sustain *latching* dynamics, i.e. it can hop from state to state, as in Fig. 1, provided its activity is appropriately regulated by suitable thresholds, as we have reported elsewhere (Treves 2005). Such spontaneous dynamics of the entire network might be led to a different extent by its frontal and posterior halves, depending on their characteristic parameters.

In order to quantify the relative influence of the two sub-networks on the latching sequences produced by the hybrid Potts model, we look at whether the actual occurrence of each possible transition depends on the correlations, computed separately in the frontal and posterior parts, between the two patterns before and after the transition.

For the randomly correlated patterns used here, the correlations are relatively minor, but they can be anyway quantified by two quantities, C_{as} and C_{ad} (Russo and Treves 2012; Boboeva et al. 2018), that is, the fraction of active units in one pattern that are co-active in the other and in the same, C_{as}, or in a different state, C_{ad}. In terms of these quantities, two memory patterns are highly correlated if C_{as} is larger than average and C_{ad} is smaller than average, and we can take the difference $C_{ad} - C_{as}$ as a simple

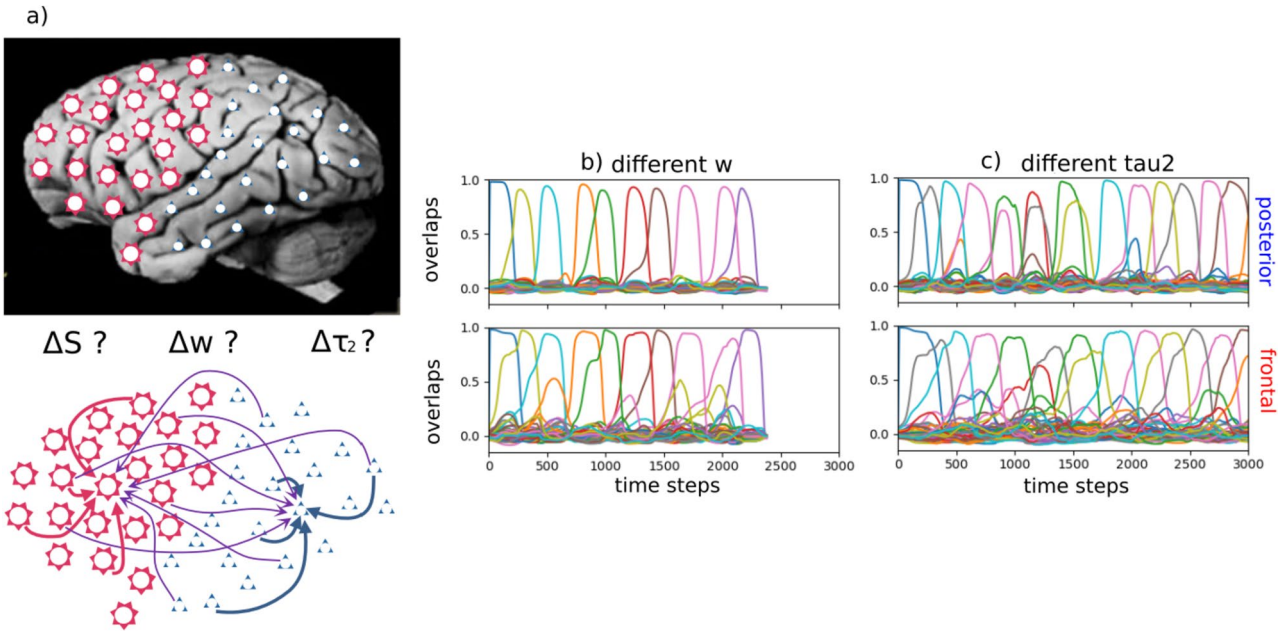

Fig. 1 The differentiated network and examples of latching sequences. **a** The differentiated network is comprised of frontal and posterior halves, in each of which units receive the same number of inputs from both halves, but not of the same average strength. **b** and **c** The latching sequences—visualized by the overlaps, i.e., by how close the state of the network at time t is to each memory pattern (assigned an arbitrary color)—are very similar if extracted from the posterior (upper panels) or the frontal sub-network (bottom panels). In **b**, parameters are set as in Fig. 2e. In **c**, parameters are set as in Fig. 3c. Close inspection reveals that in (**b**) the transitions in the frontal network appear to anticipate those in the posterior one, while in (**c**) the trend is not clear, consistent with the results described below

compact indicator (actually, a proxy) of the "distance" between the two patterns.

How strongly are transitions in a latching sequence driven by pattern correlations in each subnetwork? To measure this, we take the weighted average of C_{as} and C_{ad} with the weights given by latching sequences; that is, we compute

$$\langle C_{as}\rangle_T \equiv \sum_{(\mu,\nu)} t_{\mu\nu} C_{as}^{\mu\nu}, \tag{3}$$

(and analogously for $\langle C_{ad}\rangle_T$) where the sum $\sum_{(\mu,\nu)}$ runs over all possible pairs of memories and $t_{\mu\nu}$ is the normalized frequency of latching transitions for the pair μ, ν: $\sum_{(\mu,\nu)} t_{\mu\nu} = 1$. This average is compared with the "baseline" average, e.g.,

$$\langle C_{as}\rangle_B \equiv \frac{2}{p(p-1)} \sum_{(\mu,\nu)} C_{as}^{\mu\nu}, \tag{4}$$

independent of the transitions, where p is the number of stored memories in the network. The comparison between the two averages, $\langle C_{as(d)}\rangle_T$ and $\langle C_{as(d)}\rangle_B$, is one index of how strongly latching sequences are related to correlations between patterns in one of the two sub-networks.

Second, based on the hypothesis that the frequency of transitions tends to decrease exponentially with the distance between the two patterns, as defined above, we look for the linear regression between the logarithm of the normalized transition frequency, $\log(t)$, and the proxy of the distance, $C_{ad} - C_{as}$.

We first consider a case when all the macroscopic parameters are equal between the two sub-networks, while the connection parameter is set as $\lambda = 0.5$. In this case, the intra-connections (within each sub-network) are 3 times, on average, as strong as the inter-connections (between the two sub-networks), but the two halves are fully equivalent, or Not Differentiated (ND). With the appropriate parameters, in particular the feedback w, we find that the network as a whole shows robust latching and that latching sequences in each sub-network are well synchronized with each other: the two sub-networks essentially latch as one. Comparing latching dynamics in two sub-networks, we find that latching is largely driven by correlations between patterns, in either half or in both, as found previously (Russo and Treves 2012). This can be seen, leftmost bars of Fig. 2a and b, by the higher value of $\langle C_{as}\rangle_T$ relative to $\langle C_{as}\rangle_B$, and vice versa for C_{ad}, in the ND case. Correlations in the two sub-networks appear to contribute equally to determine latching sequences, as expected. This is confirmed by the similar negative slopes in the two scatterplots of Fig. 2c.

Different S. We now examine a case in which the two networks share the same values of all but one parameter: the number of Potts states, S. When the posterior network has fewer states ($S = 3$ instead of the reference value, 7), the

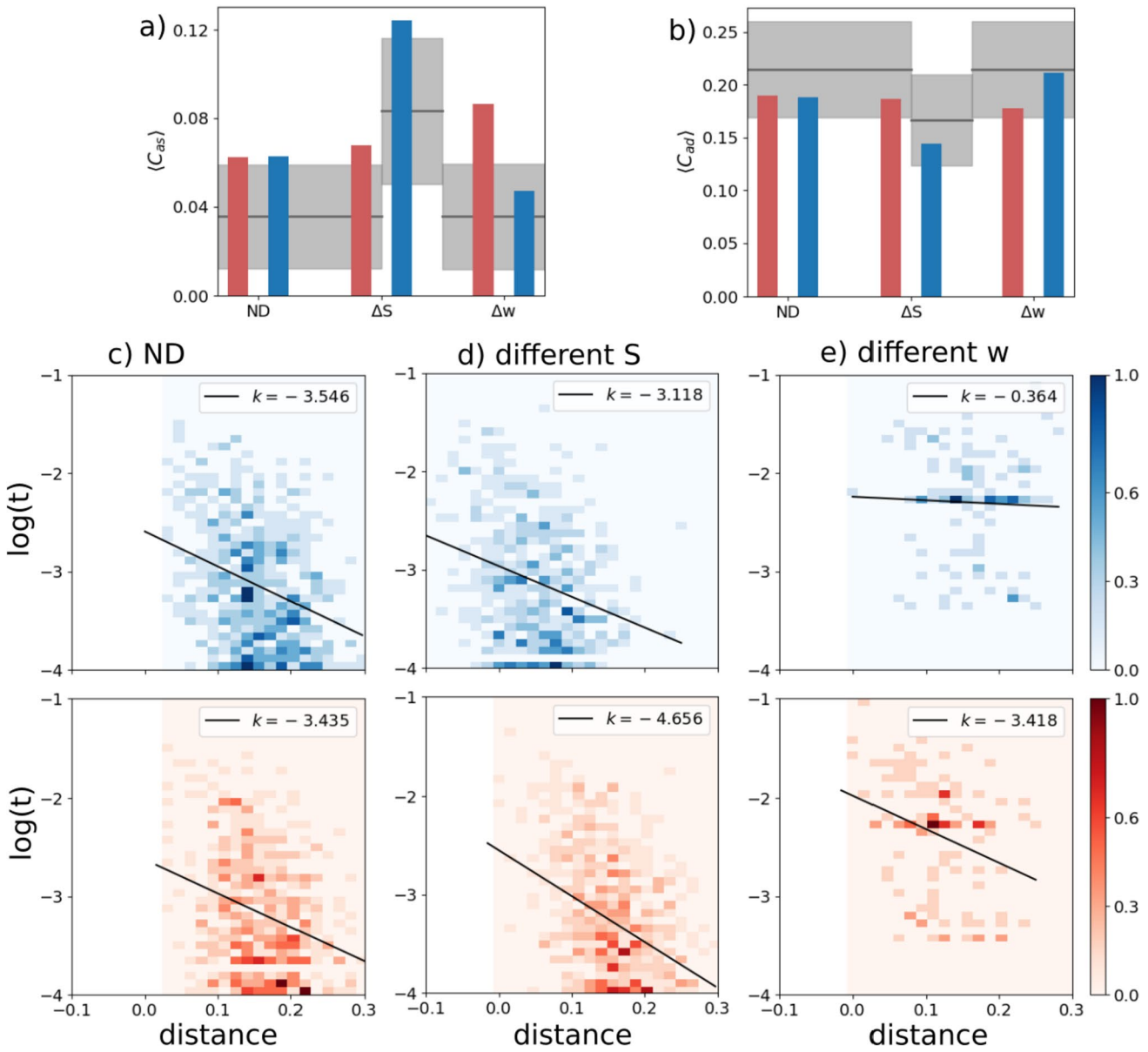

Fig. 2 A latching frontal network leads a non-latching posterior network. Red indicates the frontal and blue the posterior network in this and other figures. **a** and **b** The transition-weighted averages of C_{as} and C_{ad} are compared to their baseline values for three cases: no difference between the two networks (ND, leftmost bars), a difference in S (ΔS, middle bars) and a difference in w (Δw, rightmost bars). The gray horizontal line and shaded area indicate the baseline average and its standard deviation. **c**–**e** Scatterplots of (log) transition frequencies between individual patterns pairs versus their "distance", for the three conditions. The darkness of color indicates the number of pairs at each combination of abscissa and ordinate. For the ND condition, parameters are set as $w_p = w_f = 1.1$, $S_p = S_f = 7$. For the other conditions, the parameters of the frontal network are kept the same as in the ND condition, while the parameters of the posterior sub-network are set as $S_p = 3$ and $w_p = 0.6$, respectively, in (**d**) and (**e**). Note the negative values on the x-axis, particularly in panel (**d**) upper, due to using just a proxy of a proper distance measure, a proxy which reaches in the negative range when $S = 3$

baselines for both C_{as} and C_{ad} are shifted, above and below, respectively, but their transition-weighted values are similarly positioned, above and below the respective baselines, as in the frontal network. Also in terms of the second indicator, the scatterplot of Fig. 2d shows rather similar slopes, with only a modest quantitative "advantage" for the frontal network (in red), which can be said to lead the latching sequence somewhat more than the posterior one. One should note that, with these parameters, both sub-networks would latch if isolated.

Different w. In contrast to the two cases above, ND and ΔS, we see a major difference between the two sub-networks if it is the w parameter which is lower for the posterior network (the rightmost bars of Figs. 2a, b). In this case, it is

obviously the correlation structure of the frontal patterns, not of the posterior ones, that dominates in determining latching sequences. This is also evident from the very different slopes, k, in the scatterplot of Fig. 2e. With the lower value $w = 0.6$ chosen for the posterior sub-network, this time it would not latch, if isolated. Note that to preserve its latching, and for it to be a clear single sequence, we would have to set w at almost the same value as for the frontal sub-network, unlike the case with the S parameter.

And/or different τ_2. We now allow the adaptation timescale, τ_2, to differ between the two sub-networks. We first note that latching sequences between the sub-networks are remarkably well synchronized despite their different adaptation timescales (Fig. 1c). If isolated, the two sub-networks would each latch at a pace set by its own τ_2. Their synchronization thus shows that, even with this relativity weaker connectivity coupling (inter-connections 1/3 of the average strength of the intra-connections) the two halves are willing to compromise, and latch at some intermediate pace, close to the one they sustained when τ_2 was not differentiated.

Furthermore, latching sequences are affected predominantly by frontal correlations rather than posterior ones. In Fig. 3, we show two cases: the two sub-networks have two different adaptation timescales; and in the second case also different w. We see a moderate effect if τ_2 is the only parameter that differs between the two. Note that in this case the posterior sub-network, if isolated, would latch.

The effect is most pronounced if w is also lowered to $w = 0.6$ for the posterior sub-network, as is evident from the weak positive slope k it shows, see Fig. 3d. In this case it would not latch if isolated.

We have also inverted the τ_2 difference, making the posterior sub-network, still with a lower w, slower in terms of firing rate adaptation. In this case (not shown) latching is virtually abolished, showing that the parameter manipulations do not simply add up linearly.

Lesioning the network

To model lesions in either sub-network, we define a procedure that still allows us to compare quantities based on the same number of inputs per unit, etc. The procedure acts only on the relative weights of the connections (through λ), which are modulated while keeping their average for each receiving unit always to 1/2. Other parameters of the network are set in such a way that the frontal sub-network leads the latching sequences and that lesions do not push the network into a no-latching phase: the self-reinforcement parameter is set as $w = 0.7$ for the posterior sub-network and $w = 1.2$ for the frontal one, while S and τ_2 are set as specified in Table 1 and thus take the same value for both sub-networks. For "healthy" networks, we use $\lambda = 0.5$ in Eq. (2), meaning the intra-connections (within the frontal and within the posterior half) are 3 times, on average, as strong as the inter-connections (between frontal and posterior halves). For lesioned networks, we use smaller values of λ than 0.5 for their input connections: the smaller the value is, the stronger the lesion is. So, for example, a frontal lesion with $\lambda = 0.2$ implies that its recurrent weights are weighted by a factor 0.6 (instead of 0.75) and the weights from the posterior sub-network by a factor 0.4 (rather than 0.25), i.e. the internal weights are only 1.5 times those of the interconnections. The posterior sub-network in this case has the same weights as the control case.

We then quantify the effect of the lesions with the slopes in the scatterplots as before, but also with an entropy measure. The entropy at position z in a latching sequence measures the variability of transitions encountered at that position, across all sequences with the same starting point. It is computed as

$$S(z) = \Big\langle -\sum_{\mu\neq\nu} P_\gamma^{\mu\nu}(z)\log_2 P_\gamma^{\mu\nu}(z)\Big\rangle_\gamma, \tag{5}$$

where $P_\gamma^{\mu\nu}(z)$ is the joint probability of having two patterns μ and ν at two consecutive positions z and $z+1$ relative to the cued pattern γ in a latching sequence, and $\langle\cdot\rangle_\gamma$ means that we average the entropy across all the p patterns that are used as a cue. Note that if all transitions were incurred equally, asymptotically for large z, the entropy would reach its maximum value $S_\infty = \log_2[p(p-1)]$ (with p patterns stored in memory and available for latching). Therefore $\exp\{[S(z) - S_\infty]\ln(2)\}$ is an effective measure of the fraction of all possible transitions that the network has explored at position z, on average.

In terms of the slopes in the scatterplots, we see that posterior lesions do not have a major effect, while frontal lesions reduce the relation between the probability of individual transitions and the correlation between the two patterns, particularly in the frontal sub-network where it was strong in the "healthy" case (Fig. 4).

In terms of entropy, we see that lesions in the posterior sub-network do not affect the entropy curve, relative to that for the healthy network (Fig. 5). Lesions in the frontal sub-network, however, tend to restrict the sequences to a limited set of transitions, leading to a marked reduction in the fraction of possibilities explored by the lesioned network.

Simulated frontal lesions, therefore, produce in our model two effects that, while not opposite, are not fully congruent either. The first, manifested in the reduced slope of Fig. 4a, is suggestive of a loss of coherence in individual transitions between brain states; the second, seen in the limited entropy of Fig. 5, indicates a restriction in the space spanned by the trajectories of spontaneous thought.

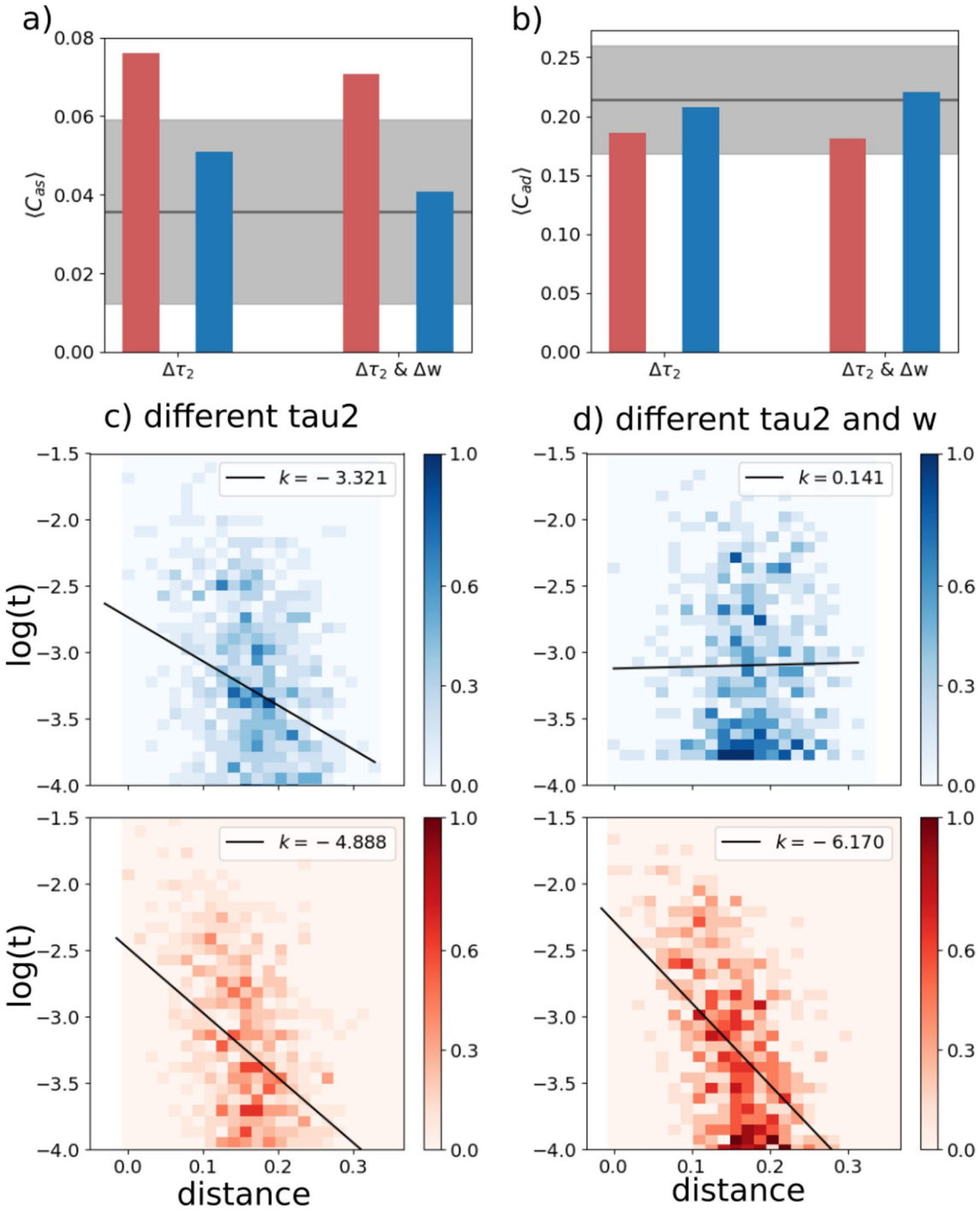

Fig. 3 The frontal sub-network is even more dominant with slower adaptation. Color code and meaning are the same as in Fig. 2. **a** and **b** Transition-weighted averages of C_{as} and C_{ad} versus their baselines are shown for two conditions: only τ_2 is different and both w and τ_2 are different. In both conditions, τ_2 is 100 for the posterior network and 400 for the frontal network. In the Δw condition, w is 0.6 for the posterior network and 1.1 for the frontal network. **c** and **d** Log-transformed transition frequencies between individual patterns pairs versus their distance

To reconcile the two outcomes, we have to conclude that while less dependent on the similarity between the two patterns, or states, individual transitions are not really random, and some become in the lesioned network much more frequent than others, gradually veering from creative towards obsessive (or perseverative) thought.

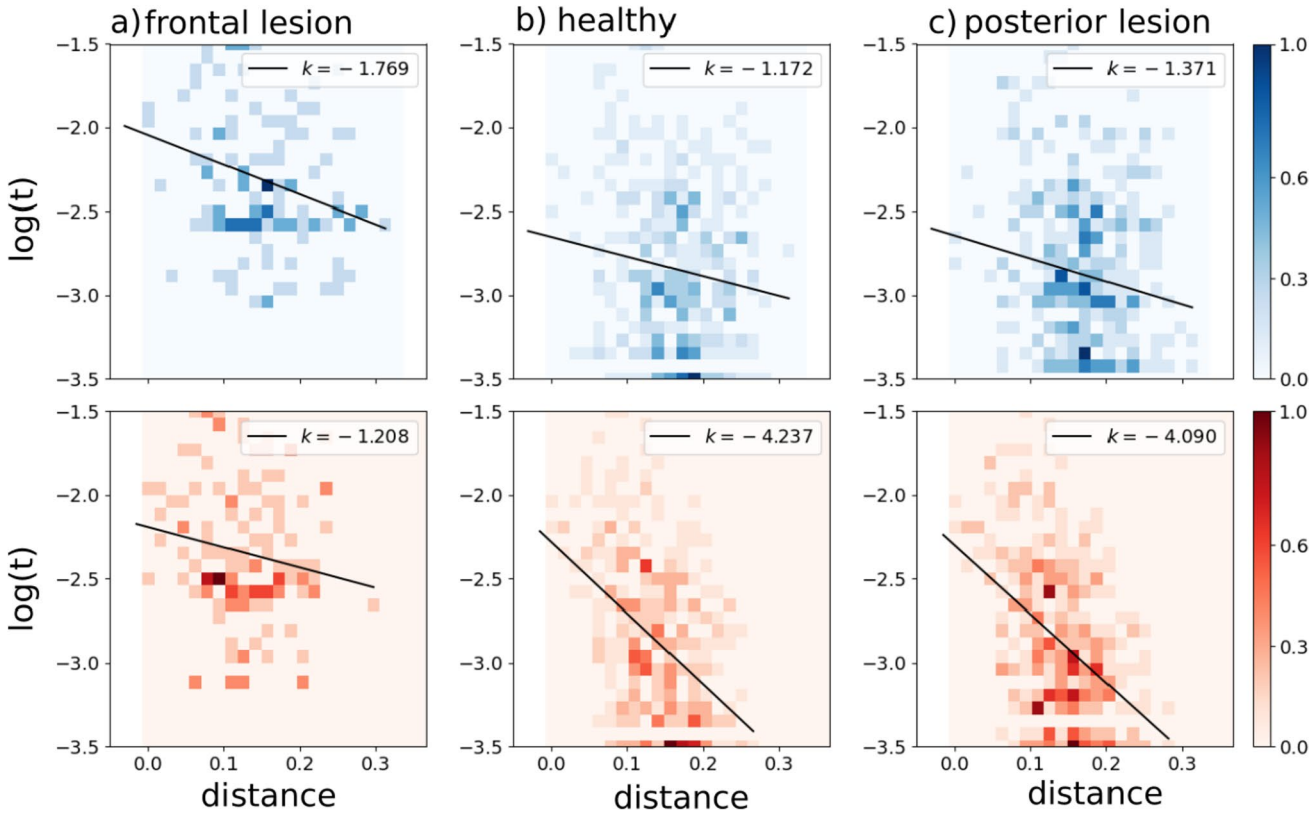

Fig. 4 Correlations between transition frequency and pattern distance are shown for a network with frontal lesions (**a**), for a healthy network (**b**) and for a network with posterior lesions (**c**). Lesions are modelled by setting $\lambda = 0.2$ (see main text). The self-reinforcement parameter is set as $w = 1.2$ for the frontal sub-network and $w = 0.7$ for the posterior one

Discussion

Simulating our model provides some insight about the conditions that may enable frontal cortex to determine the sequence of states in spontaneous thought dynamics. It is important, in assessing the computational findings, to distinguish what has gone into defining the model from what the model gives out in return. For example, much cognitive neuroscience research has been devoted to understanding the process of segmenting our ongoing experience into separate sub-events, or event segmentation (Kurby and Zacks 2008). Baldassano and colleagues (Baldassano et al. 2017) have recently demonstrated how brain activity within sub-events resembles temporarily stable activity patterns, dubbed "neural states" (Geerligs et al. 2022), which may be identified with those long posited to occur in the cortex of primates (Abeles et al. 1995) and other species (Jones et al. 2007), from analyses of single-unit activity. This notion is conceptually similar to the Potts states in a latching sequence, but finding evidence that a continuous input flow is segmented into discrete or quasi-discrete states in the brain is a major achievement, whereas in the Potts network it is a straightforward outcome of the ingredients used to define the model in the first place. Interestingly, these neural states were found to occur on different timescales across regions, with more but short-lasting transitions in low-level (posterior) sensory cortices and fewer but longer-lasting transitions in higher-level (frontal/parietal) regions. Strikingly, for some of the higher order brain regions, neural state transitions appeared to overlap with behavioural measures of event boundary perception (Baldassano et al. 2018).

In our study, the central question is which portion of the differentiated model network controls the sequence of discrete event states. We have seen that three types of differentiation, each capturing some aspect of caudo-rostral cortical variation, bias sequence control towards the "frontal" half of the network, albeit with different effectiveness. A comparison across the three types of differentiation is inherently ill-defined and somewhat arbitrary, because ΔS, Δw and $\Delta \tau_2$ are all measured on different scales, but it is apparent that the first type has a much milder effect than the second, and the third is somewhere in between. The major effect seen with Δw is likely due to the posterior network being unable to latch on its own, with the lower w value we have used. The lower S and τ_2 values do not have much of an effect on latching *per se*. The three types of

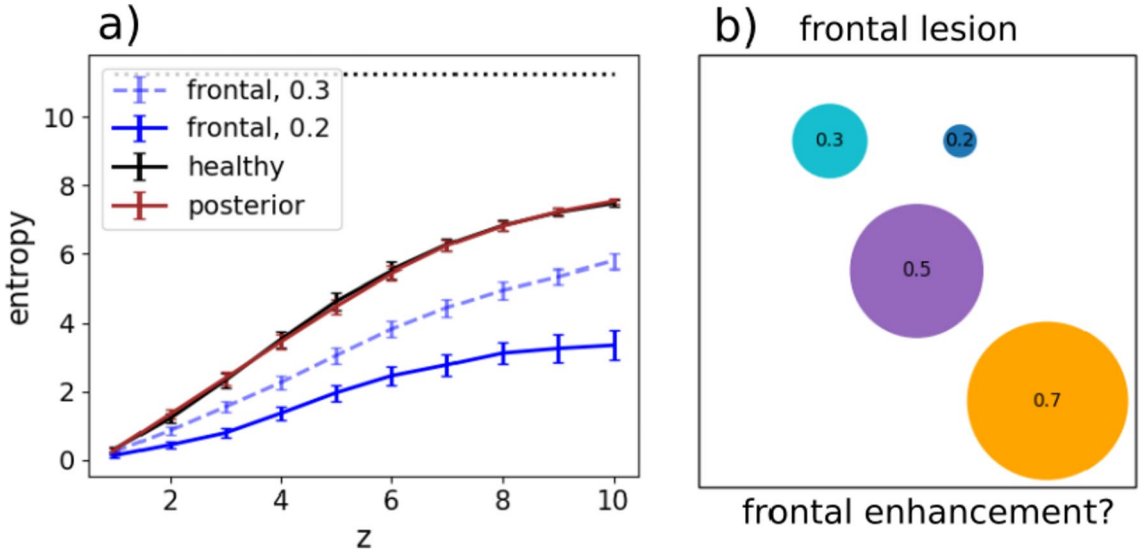

Fig. 5 **a** The entropy $S(z)$ and its standard error of the mean are shown for healthy (black), frontal-lesioned (blue) and posterior-lesioned (red) networks. Lesions are implemented by setting $\lambda = 0.2$ for solid curves, whereas the dashed blue curve is for a milder lesion in the frontal network ($\lambda = 0.3$). The black horizontal line indicates the asymptotic entropy value for a completely random sequence generated from a set of $p = 50$ patterns. The self-reinforcement parameter is set as $w = 1.2$ for the frontal network and $w = 0.7$ for the posterior network. **b** A schematic view of the diversity of transitions expressed by latching sequences. Circles are centered around an arbitrary position, while their areas extend over a fraction $2^{S(10)-S_\infty}$ of the area of the square (which would correspond to an even exploration of all possible transitions, asymptotically). The large orange circle is obtained by setting $\lambda = 0.7$, thus modelling a sort of cognitive *frontal enhancement*, perhaps obtained with psychoactive substances

differentiation are of course not mutually exclusive, and it is plausible that in the real brain, if the model makes sense, their effect would be cumulative. They do not appear to add up linearly, though: we have mentioned that inverting the τ_2 difference with respect to the w difference (i.e., making firing rate adaptation faster in the frontal sub-network) tends to abolish latching altogether, rather than reduce the frontal advantage in leading it.

A limitation of our study is that to compare the sub-networks on an even footing we have considered an artificial scenario in which activity patterns are only randomly correlated, and also there are p in each half network and they have been paired one-to-one during learning. Obviously in this scenario there is no benefit whatsoever if the network follows a frontally rather than a posteriorly-generated sequence: they are equivalent, and both devoid of content. It will be therefore important, in future work, to understand whether the insights derived under these assumptions are applicable also to more plausible conditions, in which the frontal and posterior patterns are not paired one-to-one, and can take distinct roles, for example along the lines of the classic operator/filler (also denoted as role/filler) distinction (Do and Hasselmo 2021). In this more complex scenario, the frontal patterns, if they have to serve as operators, would "take" or be paired in certain cases to a single filler and in others to multiple fillers (and possibly to other operators, in a hierarchical scheme); but even if just to one, it would be one among several options, so the pairing scheme in long-term-memory would be considerably more complex than the one considered here.

A relevant cognitive construct we mention, only partially overlapping with that of operator, is that of a temporally-oriented *schema*. A schema is a regularity extracted from multiple experience, in which B follows A and is then followed by C, although the particular instantiation of A, B and C will be different every time (Gilboa and Marlatte 2017). Note that to be implemented in our network, the skeleton of the ABC representation would have to stay activated while the specific filling items A, B and C are specified, in succession, in the posterior cortex. Alternatively, ABC could be conceptualized as a short tight latching sequence. Clearly, more attention has to be paid to the possibility of formalizing these constructs in a future well-defined network model.

Mind wandering and creativity

Within its present limitations, still our approach may offer insights relevant to the dynamics of state transitions in spontaneous cognition, such as those underlying mind wandering. Mind wandering occurs when attention drifts away from ongoing activities and towards our inner world, focusing for example on memories, thoughts, plans, which typically follow one another in a rapid, unconstrained fashion (Smallwood and Schooler 2015; Christoff et al. 2016).

The dynamics governing the flow of thoughts can indeed be described as latching (see also (Ciaramelli and Treves 2019)).

Mind wandering is known to engage the Default Mode Network (DMN), a set of interconnected brain regions, spanning from posterior, temporal, and frontal cortices (Buckner et al. 2008; Andrews-Hanna et al. 2014; Raichle 2015; Smallwood 2013; Christoff et al. 2016; Stawarczyk et al. 2011), underlying introspection and spontaneous (endogenously triggered) cognition. Ciaramelli and Treves (2019) and McCormick et al. (2018) have proposed that the prefrontal cortex, especially in its ventral-medial sectors (vmPFC) might support the initiation (internal triggering) of mind-wandering events. Indeed, recent MEG findings show that activity in the vmPFC precedes (presumably drives) hippocampal activity during (voluntary) scene construction and autobiographical memory retrieval [(Barry et al. 2019); see also (Monk et al. 2020, 2021)], and this region may play a similar role during spontaneous cognition. Indeed, damage (Bertossi and Ciaramelli 2016; Philippi et al. 2021) or inhibition (Bertossi et al. 2017; Giordani et al. 2023) of the vmPFC [but not the hippocampus; McCormick et al. (2018)] reduce the frequency of mind-wandering.

On one view, vmPFC initiates event construction by activating schemata (about the self, or common events) that help collect relevant details that the hippocampus then binds in coherent, envisioned scenes [(Ciaramelli et al. 2019); see also Benoit et al. (2014); Moscovitch et al. (2016); Rolls (2022)]. Consistent with the schema hypothesis, vmPFC (but not hippocampal) patients are particularly impaired in event construction when the task benefits from the activation of the self schema (Verfaellie et al. 2019; Stendardi et al. 2021), and are not impaired when the need for self-initiation is minimized (De Luca et al. 2019). vmPFC may also govern schema-congruent transitions between successive scenes of constructed events based on event schemata (scripts) (Stawarczyk et al. 2011; Lieberman et al. 2019), which may explain why vmPFC patients are particularly poor at simulating extended events as opposed to single moments selected from events (Bertossi and Ciaramelli 2016; Kurczek et al. 2015). The results from our computational simulations accord with and complement this view. Lesioning the frontal (but not the posterior) sector of the network led to more random state transitions, less dependent on the correlation between patterns, and also led to shorter-lasting sequences, that fade out after fewer state transitions. This pattern of findings is expected if transitions in thought states were not guided by schematic knowledge, making them less coherent in content and self-exhausting.

A second effect we observed is a reduced entropy following lesions in the frontal (but not posterior) half of the network, which indicates that the trajectories of state transitions were confined in a limited space, as if mind wandering lost its 'wandering' nature to become more constrained, with recurring thoughts characteristic of the perseverative responses long observed in prefrontal patients; suggesting that vmPFC patients, in addition to an impaired activation of relevant schemata, also fail in flexibly *de*activating current but no longer relevant ones (Gilboa and Marlatte 2017).

The most characteristic memory deficit following vmPFC damage is confabulation, the spontaneous production of false memories. Confabulations often involve an inability to inhibit previously reinforced memory traces (Schnider 2003). For example, confabulators can falsely endorse personal events as true because these were true in the past (e.g., that they just played football while in fact they used to play football during childhood). If presented with modified versions of famous fairy tales to study, confabulators tend to revert to the original versions of the stories in a later recall phase (Attali et al. 2009). Similarly, during navigation, confabulators may get lost because they head to locations they have attended frequently in the past, instead of the currently specified goal destination (Ciaramelli 2008).

The inability to flexibly switch between relevant time schemata and memory traces has been linked to reduced future thinking and reduced generation of novel scenarios in prefrontal patients ((de Vito et al. 2012); see also (Bertossi and Ciaramelli 2016)), who admitted they found themselves bound to recast past memories while trying to imagine future events. More in general, prefrontal lesions impair creativity. There is interaction between the DMN and the fronto-parietal control network while generating (DMN) and revising (fronto-parietal network) creative ideas (Beaty et al. 2014; Bendetowicz et al. 2017). Bendetowicz et al. found that damage to the right medial prefrontal regions of the DMN affected the ability to generate remote ideas, whereas damage to left rostrolateral prefrontal region of the fronto-parietal control network spared the ability to generate remote ideas but impaired the ability to appropriately combine them.

Note, however, that the originality associated with creative ideas can be conceived as disrupting the automatic progression from a thought to the one most correlated to it. Fan et al. (2023) had participants perform a creative writing task, and indeed found the *semantic distance* between adjacent sentences to be positively correlated with the story originality. Also, semantic distance was predicted by connectivity features of the salience network (e.g., the insula and anterior cingulate cortex) and the DMN. Green et al. (2006) have also reported a putative role of mPFC (BA 9/10) in connecting semantically distant concepts during abstract relational integration. In a following study (Green et al. 2010), mPFC activity was found to vary monotonically with increasing semantic distance between abstract concepts, even when controlling for task difficulty. Indeed, preliminary evidence from patients with vmPFC lesions is indicative of

a greater global semantic coherence in speech compared to healthy participants (Stendardi et al., in preparation). These results align with our finding that a lesion of the frontal component of the network produces a reduction in entropy, making latching dynamics "less creative"; but not, *prima facie*, with the reduced slope in Fig. 4a, which indicates that the lesion would produce more random transitions, frequent also among distant patterns. The apparent contradiction can be reconciled by noting that, as seen above, *individual* random transitions can still result in reduced entropy, if they tend to recur perseveratively within a sequence; and also that *semantic* coherence may reflect pattern correlation in posterior rather than frontal cortices, whereas it is logical/syntactic consequentiality that is expected to be impaired by random frontal transitions. In fact, in our model lesion, the decreased slope in the frontal sub-network seen in Fig. 4a (more random transitions) is accompanied by a slightly increased slope, suggestive of more semantic coherence, posteriorly.

Clearly, a major refinement of our approach is required, before these suggestions can be taken seriously, and articulated in a more nuanced and anatomy-informed view (Rolls et al. 2023) of how operating along the time dimension may be coordinated across cortical areas.

Appendix A: Potts model details

A Potts neural network is an autoassociative memory network comprised of N Potts units, which model patches of cortex as they contribute to retrieve distributed long-term memory traces addressed by their contents (Treves 2005). Each Potts unit has S active states, indexed as $1, 2, \cdots, S$, representing local attractors in that patch, and one quiet state, the 0 state. The N units interact with each other via tensor connections, that represent associative long-range interactions through axons that travel through the white matter (Braitenberg and Schüz 1991), while local, within-gray-matter interactions are assumed to be governed by attractor dynamics in each patch. The values of the tensor components are pre-determined by the Hebbian learning rule, which can be construed as derived from Hebbian plasticity at the synaptic level (Naim et al. 2018)

$$J_{ij}^{kl} = \frac{c_{ij}}{c_m a(1-\frac{a}{S})} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right) \left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{6}$$

where c_{ij} is either 1 if unit j gives input to unit i or 0 otherwise, allowing for asymmetric connections between units, and the δ's are the Kronecker symbols. The number of input connections per unit is c_m. The p distributed activity patterns which represent memory items are assigned, in the simplest model, as composition of local attractor states $\{\xi_i^\mu\}$ ($i = 1, 2, \cdots, N$ and $\mu = 1, 2, \cdots, p$). The variable ξ_i^μ indicates the state of unit i in pattern μ and is randomly sampled, independently on the unit index i and the pattern index μ, from $\{0, 1, 2, \cdots, S\}$ with probability

$$P(\xi_i^\mu = k) = \frac{a}{S}(1-\delta_{k,0}) + (1-a)\delta_{k,0}. \tag{7}$$

Constructed in this way, patterns are randomly correlated with each other. We use these randomly correlated memory patterns $\{\xi_i^\mu\}_{\mu=1,\ldots,p}$ in this study. The parameter a is the sparsity of patterns—fraction of active units in each pattern; the average number of active units in any pattern μ is therefore given by Na.

Local network dynamics within a patch are taken to be driven by the "current" that the unit i in state k receives

$$h_i^k(t) = \sum_{j\neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l(t) + w\left[\sigma_i^k(t) - \frac{1}{S}\sum_{l=1}^{S} \sigma_i^l(t)\right], \tag{8}$$

where the local feedback w, introduced in Russo and Treves (2012), models the depth of attractors in a patch, as shown in Naim et al. (2018)—it helps the corresponding Potts unit converge to its most active state. The activation along each state for a given Potts unit is updated with a *soft max* rule

$$\sigma_i^k(t) = \frac{\exp[\beta r_i^k(t)]}{\sum_{k=1}^{S}\exp[\beta r_i^k(t)] + \exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}} \quad \text{if } k > 0,$$
$$\sigma_i^0(t) = \frac{\exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}}{\sum_{k=1}^{S}\exp[\beta r_i^k(t)] + \exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}} \quad \text{if } k = 0, \tag{9}$$

where U is a fixed threshold common for all units and β is an effective inverse "temperature" (noise level). Note that σ_i^k takes continuous values in (0, 1) and that $\sum_{k=0}^{S}\sigma_i^k = 1$ for any i. The variables r_i^k, θ_i^A and θ_i^B parameterize, respectively, the state-specific potential, fast inhibition and slow inhibition in patch i. The state-specific potential r_i^k integrates the state-specific current h_i^k by

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t), \tag{10}$$

where the variable θ_i^k is a specific threshold for unit i and for state k.

Taking the threshold θ_i^k to vary in time to model adaptation, i.e. synaptic or neural fatigue selectively affecting the neurons active in state k, and not all neurons subsumed by Potts unit i

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t), \tag{11}$$

the Potts network additionally expresses latching dynamics, the key to its possible role in modelling temporal schemata.

The unit-specific thresholds θ_i^A and θ_i^B describe local inhibition, which in the cortex is relayed by at least 3 main classes of inhibitory interneurons (Tremblay et al. 2016) acting on $GABA_A$ and $GABA_B$ receptors, with widely different time courses, from very short to very long. Formally in our model, θ_i^A denotes fast, $GABA_A$ inhibition and θ_i^B denotes slow, $GABA_B$ inhibition and they vary in time in the following way:

$$\tau_A \frac{d\theta_i^A(t)}{dt} = \gamma_A \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^A(t), \tag{12}$$

$$\tau_B \frac{d\theta_i^B(t)}{dt} = (1-\gamma_A) \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^B(t), \tag{13}$$

where one sets $\tau_A < \tau_1 \ll \tau_2 \ll \tau_B$ and the parameter γ_A sets the balance of fast and slow inhibition. Specifically in this work, we set these parameters as $\tau_A = 10$, $\tau_B = 10^5$, $\tau_1 = 20$ and $\gamma_A = 0.5$.

Appendix B: Simulation details

We have used an asynchronous updating, where one unit is updated at a time with a random order. Updating all Potts units in the network once is our measuring unit of simulation time: all timescales of the model are measured with this unit. We stop the simulation after updating the entire network 10,000 times (except for Fig. 5, see next paragraph). Then, we cut out the first 3 patterns in the sequence to remove the effect of initialization. Every stored memory is used as a cue with its full representation.

In order to compute the probability $P_\gamma^{\mu\nu}(z)$ in Eq. (5), we have run $p \times 1000$ simulations for each condition. For each memory pattern, we take 40% of its active units and flip them into different states. We prepare 1000 corrupted versions of each memory by repeating this procedure 1000 times. Each of these corrupted versions is used as a cue in each simulation, which is terminated after 12 transitions.

Unless specified explicitly, parameters of the Potts model are set as in Table 1.

Table 1 Parameters of the network

Symbol	Meaning	Default value
N	Number of Potts units	256
c_m	Number of presynaptic units	50
S	Number of states per unit	7
p	Number of memory patterns	50
a	Sparsity of patterns	0.25
λ	Relative coupling strength	0.5
U	Global threshold	0.1
τ_2	Adaptation timescale	200
w	Self-reinforcement term	1.1
β	Inverse "temperature"	11

Acknowledgements This work was supported by PRIN Grant 20174TPEFJ "TRIPS" to EC and AT. We are grateful for early discussions with Massimiliano Trippa and with Edmund Rolls.

Author contributions KIR and AT conceived and developed the computational project as a means of articulating the neuropsychological perspective put forward by EC and DS. KIR run most of the simulations, which were complemented by others developed by AB. DS spurred the analysis of coherence. KIR drafted this report and all coauthors contributed to refine and complete it.

Funding Open access funding provided by Scuola Internazionale Superiore di Studi Avanzati - SISSA within the CRUI-CARE Agreement.

Data availability The simulation code is available in the OSF repository at the link (to be made public upon acceptance) https://osf.io/fk5uz/?view_only=55fd1a67077846c8b44dd6e1d3933831.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abeles M, Bergman H, Gat I, Meilijson I, Seidemann E, Tishby N, Vaadia E (1995) Cortical activity flips among quasi-stationary states. Proc Natl Acad Sci 92(19):8616–8620

Andrews-Hanna JR, Smallwood J, Spreng RN (2014) The default network and self-generated thought: component processes, dynamic control, and clinical relevance. Ann N Y Acad Sci 1316(1):29–52

Attali E, De Anna F, Dubois B, Barba GD (2009) Confabulation in Alzheimer's disease: poor encoding and retrieval of overlearned information. Brain 132(1):204–212

Aziz-Zadeh L, Liew S-L, Dandekar F (2013) Exploring the neural correlates of visual creativity. Soc Cogn Affect Neurosci 8(4):475–480

Badre D (2008) Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn Sci 12(5):193–200

Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman Kenneth A (2017) Discovering event structure in continuous narrative perception and memory. Neuron 95(3):709–721

Baldassano C, Hasson U, Norman KA (2018) Representation of real-world event schemas during narrative perception. J Neurosci 38(45):9689–9699

Barbas H, Rempel-Clower N (1997) Cortical structure predicts the pattern of corticocortical connections. Cereb Cortex 7(7):635–646

Barry DN, Barnes GR, Clark IA, Maguire EA (2019) The neural dynamics of novel scene imagery. J Neurosci 39(22):4375–4386

Beaty RE (2015) The neuroscience of musical improvisation. Neurosci Biobehav Rev 51:108–117

Beaty RE, Benedek M, Wilkins RW, Jauk E, Fink A, Silvia PJ, Hodges DA, Koschutnig K, Neubauer AC (2014) Creativity and the default network: a functional connectivity analysis of the creative brain at rest. Neuropsychologia 64:92–98

Bendetowicz D, Urbanski M, Aichelburg C, Levy R, Volle E (2017) Brain morphometry predicts individual creative potential and the ability to combine remote ideas. Cortex 86:216–229

Benedek M, Beaty RE, Schacter DL, Kenett YN (2023) The role of memory in creative ideation. Nat Rev Psychol 2(4):246–257

Benoit RG, Szpunar KK, Schacter DL (2014) Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. Proc Natl Acad Sci 111(46):16550–16555

Bertossi E, Ciaramelli E (2016) Ventromedial prefrontal damage reduces mind-wandering and biases its temporal focus. Soc Cogn Affect Neurosci 11(11):1783–1791

Bertossi E, Peccenini L, Solmi A, Avenanti A, Ciaramelli E (2017) Transcranial direct current stimulation of the medial prefrontal cortex dampens mind-wandering in men. Sci Rep 7(1):16962

Boboeva V, Brasselet R, Treves A (2018) The capacity for correlated semantic memories in the cortex. Entropy 20(11):824

Buckner RL, Andrews-Hanna JR, Schacter DL (2008) The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124(1):1–38

Cavanagh SE, Hunt LT, Kennerley SW (2020) A diversity of intrinsic timescales underlie neural computations. Front Neural Circuits 14:615626

Chang Y-M, Rosene DL, Killiany RJ, Mangiamele LA, Luebke JI (2005) Increased action potential firing rates of layer 2/3 pyramidal cells in the prefrontal cortex are significantly related to cognitive performance in aged monkeys. Cereb Cortex 15(4):409–418

Chaudhuri R, Knoblauch K, Gariel M-A, Kennedy H, Wang X-J (2015) A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. Neuron 88(2):419–431

Christoff K, Irving ZC, Fox KCR, Spreng RN, Andrews-Hanna JR (2016) Mind-wandering as spontaneous thought: a dynamic framework. Nat Rev Neurosci 17(11):718–731

Ciaramelli E (2008) The role of ventromedial prefrontal cortex in navigation: a case of impaired wayfinding and rehabilitation. Neuropsychologia 46(7):2099–2105

Ciaramelli E, Treves A (2019) A mind free to wander: neural and computational constraints on spontaneous thought. Front Psychol 10:39

Ciaramelli E, De Luca F, Monk AM, McCormick C, Maguire EA (2019) What "wins'' in vmPFC: scenes, situations, or schema? Neurosci Biobehav Rev 100:208–210

Cocchi L, Sale MV, Gollo LL, Bell PT, Nguyen VT, Zalesky A, Breakspear M, Mattingley JB (2016) A hierarchy of timescales explains distinct effects of local inhibition of primary visual cortex and frontal eye fields. Elife 5:e15252

De Luca F, McCormick C, Ciaramelli E, Maguire EA (2019) Scene processing following damage to the ventromedial prefrontal cortex. Neuroreport 30(12):828

de Vito S, Gamboz N, Brandimonte MA, Barone P, Amboni M, Della Sala S (2012) Future thinking in Parkinson's disease: an executive function? Neuropsychologia 50(7):1494–1501

Do Q, Hasselmo ME (2021) Neural circuits and symbolic processing. Neurobiol Learn Mem 186:107552

Douglas RJ, Martin KAC, Whitteridge D (1989) A canonical microcircuit for neocortex. Neural Comput 1(4):480–488

Elston GN, Benavides-Piccione R, DeFelipe J (2001) The pyramidal cell in cognition: a comparative study in human and monkey. J Neurosci 21(17):RC163–RC163

Fan L, Zhuang K, Wang X, Zhang J, Liu C, Jing G, Qiu J (2023) Exploring the behavioral and neural correlates of semantic distance in creative writing. Psychophysiology 60(5):e14239

Finlay BL, Uchiyama R (2015) Developmental mechanisms channeling cortical evolution. Trends Neurosci 38(2):69–76

Gao R, van den Brink RL, Pfeffer T, Voytek B (2020) Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. Elife 9:e61277

Geerligs L, Gözükara D, Oetringer D, Campbell KL, van Gerven M, Güçlü Umut (2022) A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. ELife 11:e77430

Giacometti GL, Andrea C, Giovanni C, Alessio A, Elisa C (2023) The role of posterior parietal cortex and medial prefrontal cortex in distraction and mind-wandering. Neuropsychologia 188:108639

Gilboa A, Marlatte H (2017) Neurobiology of schemas and schema-mediated memory. Trends Cogn Sci 21(8):618–631

Green AE, Fugelsang JA, Kraemer DJM, Shamosh NA, Dunbar KN (2006) Frontopolar cortex mediates abstract integration in analogy. Brain Res 1096(1):125–137

Green AE, Kraemer DJM, Fugelsang JA, Gray JR, Dunbar KN (2010) Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. Cereb Cortex 20(1):70–76

Hilgetag CC, Goulas A, Changeux J-P (2022) A natural cortical axis connecting the outside and inside of the human brain. Netw Neurosci 6(4):950–959

Jones LM, Fontanini A, Sadacca BF, Miller P, Katz DB (2007) Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. Proc Natl Acad Sci 104(47):18772–18777

Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. Science 302(5648):1181–1185

Kurby CA, Zacks JM (2008) Segmentation in the perception and memory of events. Trends Cogn Sci 12(2):72–79

Kurczek J, Wechsler E, Ahuja S, Jensen U, Cohen NJ, Tranel D, Duff M (2015) Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. Neuropsychologia 73:116–126

Lieberman MD, Straccia MA, Meyer ML, Meng D, Tan KM (2019) Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): causal, multivariate, and reverse inference evidence. Neurosci Biobehav Rev 99:311–328

Manea AMG, Zilverstand A, Hayden B, Zimmermann J (2023) Neural timescales reflect behavioral demands in freely moving rhesus macaques. bioRxiv, p 2023–03

McCormick C, Ciaramelli E, De Luca F, Maguire EA (2018) Comparing and contrasting the cognitive effects of hippocampal and ventromedial prefrontal cortex damage: a review of human lesion studies. Neuroscience 374:295–318

Mekern V, Hommel B, Sjoerds Z (2019) Computational models of creativity: a review of single-process and multi-process recent approaches to demystify creative cognition. Curr Opin Behav Sci 27:47–54

Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J Neurosci 16(16):5154–5167

Monk AM, Barnes GR, Maguire EA (2020) The effect of object type on building scene imagery-an meg study. Front Hum Neurosci 14:592175

Monk AM, Dalton MA, Barnes GR, Maguire EA (2021) The role of hippocampal-ventromedial prefrontal cortex neural dynamics in building mental representations. J Cogn Neurosci 33(1):89–103
Moscovitch M, Cabeza R, Winocur G, Nadel L (2016) Episodic memory and beyond: the hippocampus and neocortex in transformation. Ann Rev Psychol 67:105–134
Naim M, Boboeva V, Kang CJ, Treves A (2018) Reducing a cortical network to a Potts model yields storage capacity estimates. J Stat Mech Theor Exp 2018(4):043304
Peter S (2019) Cognitive poetics: an introduction. Routledge, Milton Park
Philippi CL, Bruss J, Boes AD, Albazron FM, Deifelt Streese C, Ciaramelli E, Rudrauf D, Tranel D (2021) Lesion network mapping demonstrates that mind-wandering is associated with the default mode network. J Neurosci Res 99(1):361–373
Raichle ME (2015) The brain's default mode network. Ann Rev Neurosci 38:433–447
Roe AW (2019) Columnar connectome: toward a mathematics of brain function. Netw Neurosci 3(3):779–791
Rolls Edmund T (2022) The hippocampus, ventromedial prefrontal cortex, and episodic and semantic memory. Prog Neurobiol 217:102334
Rolls ET, Deco G, Huang C-C, Feng J (2023) The human orbitofrontal cortex, vmPFC, and anterior cingulate cortex effective connectome: emotion, memory, and action. Cereb Cortex 33(2):330–356
Rotshtein P, Henson RNA, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat Neurosci 8(1):107–113
Russo E, Treves A (2012) Cortical free-association dynamics: distinct phases of a latching network. Phys Rev E 85(5):051920
Ryom KI, Treves A (2023) Speed inversion in a potts glass model of cortical dynamics. PRX Life 1(1):013005
Ryom KI, Boboeva V, Soldatkina O, Treves A (2021) Latching dynamics as a basis for short-term recall. PLoS Comput Biol 17(9):e1008809
Schnider A (2003) Spontaneous confabulation and the adaptation of thought to ongoing reality. Nat Rev Neurosci 4(8):662–671
Smallwood J (2013) Distinguishing how from why the mind wanders: a process-occurrence framework for self-generated mental activity. Psychol Bull 139(3):519
Smallwood J, Schooler JW (2015) The science of mind wandering: empirically navigating the stream of consciousness. Ann Rev Psychol 66:487–518
Stawarczyk D, Majerus S, Maquet P, D'Argembeau A (2011) Neural correlates of ongoing conscious experience: both task-unrelatedness and stimulus-independence are related to default network activity. PLoS ONE 6(2):e16997
Stendardi D, Biscotto F, Bertossi E, Ciaramelli E (2021) Present and future self in memory: the role of vmPFC in the self-reference effect. Soc Cogn Affect Neurosci 16(12):1205–1213
Tremblay R, Lee S, Rudy B (2016) GABAergic interneurons in the neocortex: from cellular properties to circuits. Neuron 91(2):260–292
Treves A (2005) Frontal latching networks: a possible neural basis for infinite recursion. Cogn Neuropsychol 22(3–4):276–291
Valentino B (1978) Cortical architectonics: general and areal. In: Brazier M, Petsche H (eds) Architectonics of the cerebral cortex. Raven Press, New York, pp 443–465
Valentino B, Almut S (1991) Anatomy of the cortex: statistics and geometry, vol 18. Springer Verlag, Berlin
Verfaellie M, Wank AA, Reid AG, Race E, Keane MM (2019) Self-related processing and future thinking: distinct contributions of ventromedial prefrontal cortex and the medial temporal lobes. Cortex 115:159–171