# Bump attractor dynamics underlying stimulus integration in perceptual estimation tasks

Jose M. Esnaola-Acebes[1,2], Alex Roxin[1,2], Klaus Wimmer[1,2,*]

[1] Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra (Barcelona), Spain
[2] Barcelona Graduate School of Mathematics, Barcelona, Spain
* kwimmer@crm.cat

**Perceptual decision and continuous stimulus estimation tasks involve making judgments based on accumulated sensory evidence. Network models of evidence integration usually rely on competition between neural populations each encoding a discrete categorical choice and do not maintain information that is necessary for a continuous perceptual judgement. Here, we show that a continuous attractor network can integrate a circular stimulus feature and track the stimulus average in the phase of its activity bump. We show analytically that the network can compute the running average of the stimulus almost optimally, and that the nonlinear internal dynamics affect the temporal weighting of sensory evidence. Whether the network shows early (primacy), uniform or late (recency) weighting depends on the relative strength of the stimuli compared to the bump's amplitude and initial state. The global excitatory drive, a single model parameter, modulates the specific relation between internal dynamics and sensory inputs. We show that this can account for the heterogeneity of temporal weighting profiles and reaction times observed in humans integrating a stream of oriented stimulus frames. Our findings point to continuous attractor dynamics as a plausible mechanism underlying stimulus integration in perceptual estimation tasks.**

Integrating information over time is a fundamental computation that neural systems need to perform in perceptual decision and continuous estimation tasks. While a categorical perceptual decision usually requires discriminating stimuli in order to select between several options (e.g. left vs. right motion of an ambiguous random dot stimulus[1], or clockwise vs. counterclockwise motion compared to a reference[2]), in stimulus estimation tasks participants report a continuous stimulus feature, for example the stimulus orientation in degrees. Typical estimation tasks involve accumulating evidence about the average motion direction of a random dot pattern in degrees[2–5] or the average of a sequence of oriented gratings[6,7]. In particular, we are interested here in experimental tasks that require estimating, as an analog quantity, the temporal average of a circular feature such as the orientation or direction of a potentially time-varying stimulus.

Neural network models of evidence integration usually rely on slow recurrent dynamics and competition between neural populations, each encoding one of the categorical choice options[8–11]. The architecture of these *discrete* attractor models thus directly reflects the categorical nature of the decision task. Similarly, phenomenological drift-diffusion models linearly integrate stimulus evidence for or against different choice options until one of the bounds is reached[12–17]. While the amount of accumulated evidence (i.e., the state of the decision variable) at the time of the decision is predictive of choice certainty or confidence[18], by design these models do not maintain information about continuous features of the integrated stimulus (e.g. the average orientation; but see ref.[19]). They are thus not well suited to study stimulus estimation tasks that require continuous perceptual judgements or tasks that require both the report of a continuous estimate and a categorical choice[2,20,21].

The representation of continuous stimuli in neural population codes has been studied extensively using normative models[20] or neural coding models[2,22]. These models can explain how to optimally represent a stimulus in a neural population (e.g. an orientated stimulus in area V1), how to compute an estimate from a noisy sensory input[23], how to combine different cues in a single representation[24], and how to read-out the population code in order to optimally discriminate stimuli[25]. These theoretical studies have focused on investigating the representation of stimulus features in sensory cortices rather than the computations involved in perceptual decision making and stimulus estimation. There is a lack of biophysical neural network models that can be used to study the underlying circuit mechanisms of these computations.

A natural candidate for such a neural network model is the well-known *continuous* attractor model[26,27]. Continuous bump attractor networks have a topological ring architecture and due to strong recurrent coupling they can show localized activity in a subset of neurons with a bell-shaped pattern (bump) that persists even in the absence of tuned external input. Because all states along the ring are energetically equal, the bump can be located at any location along the network. The bump location thus provides a substrate for encoding a continuous variable such as an orientation or a spatial location. Moreover, while intrinsically the bump position is stable, external inputs can drive the bump to move. Due to these features, bump attractor networks have been proposed as the neural mechanism underlying a variety of sensory and cognitive computations: the emergence of contrast-invariant orientation tuning in visual cortex[28], the impact of microstimulation on the identification of the direction of a weak motion stimulus[29], the maintenance of spatial working memory in prefrontal cortex[30,31], and the representation of spatial orientation and angular path integration in the mammalian head direction system[32–35], the fly's heading direction system[36] and in grid cells[37]. None of these previous studies has investigated whether a bump attractor network can accumulate sensory information to estimate the time-average of a continuous stimulus as required in perceptual estimation tasks.

Here, we show that the ring attractor model can accurately compute the average of a time-varying circular stimulus feature. Close to the dynamical regime where the bump of activity emerges the network approximates a *perfect vector integrator* such that the phase of the bump of activity at each time point corresponds to the running circular average of the stimulus. The temporal weighting of stimulus information is uniform in this case, as is characteristic of perfect integration. Moreover, the model can also exhibit primacy and recency regimes, i.e., it can integrate the stimulus giving higher weight to early or late parts of the stimulus in a fixed duration experiment. These integration regimes depend on the global excitatory drive to the network which is a single parameter controlling the amplitude and the dynamics of the bump. On average, the model estimates the mean stimulus orientation faithfully albeit with a higher trial-to-trial variability when the temporal weighting is not uniform. To illustrate the relevance of our theoretical results, we analyze data from psychophysical experiments in which human observers integrated a stream of eight oriented stimulus frames. We show that the continuous attractor model can quantitatively fit the observed heterogeneity of temporal weighting regimes across subjects and their relationship to the subjects' reaction times. Our work suggests continuous attractor dynamics as a potential mechanism underlying stimulus integration in perceptual estimation tasks.

# Results

## Stimulus integration with bump attractor dynamics

We start by simulating the integration of stimuli in a continuous ring attractor model, as in experimental tasks that require estimating, as an analog quantity, the average direction of motion of a noisy random dot stimulus[2,3] or of a sequence of oriented Gabor patches[6,7]. The network model has a ring architecture and Mexican-hat connectivity (Fig. 1a) and its firing rate dynamics are described by Eq. 4 (Methods). Due to strong recurrent connectivity, a bump of activity emerges in the network at a position determined by the stimulus, and this bump persists when the stimulus is removed (Fig. 1b). In this regime, a fast change of the stimulus to a new direction does not, in general, lead to an instantaneous extinction and re-appearance of the bump at the new direction, but rather initiates a continuous translation of the activity along the ring network, directed towards the new stimulus direction (Fig. 1b, Extended Data Fig. 1). This has been termed "virtual rotation"[28]. Thus, the bump position at a given time depends on the current input to the network and on the previous bump position.

We wondered whether through this mechanism the position (phase, $\psi$) of the activity bump could track the average of a time-varying stimulus. We simulated a task which required estimating the average direction of eight successive oriented stimulus frames with constant strength (e.g. stimulus contrast), and directions distributed between $-90°$ and $+90°$ (Fig. 1c). We found that the transient population response to changing stimulus input effectively computes this estimation (Fig. 1c). In the example trial,

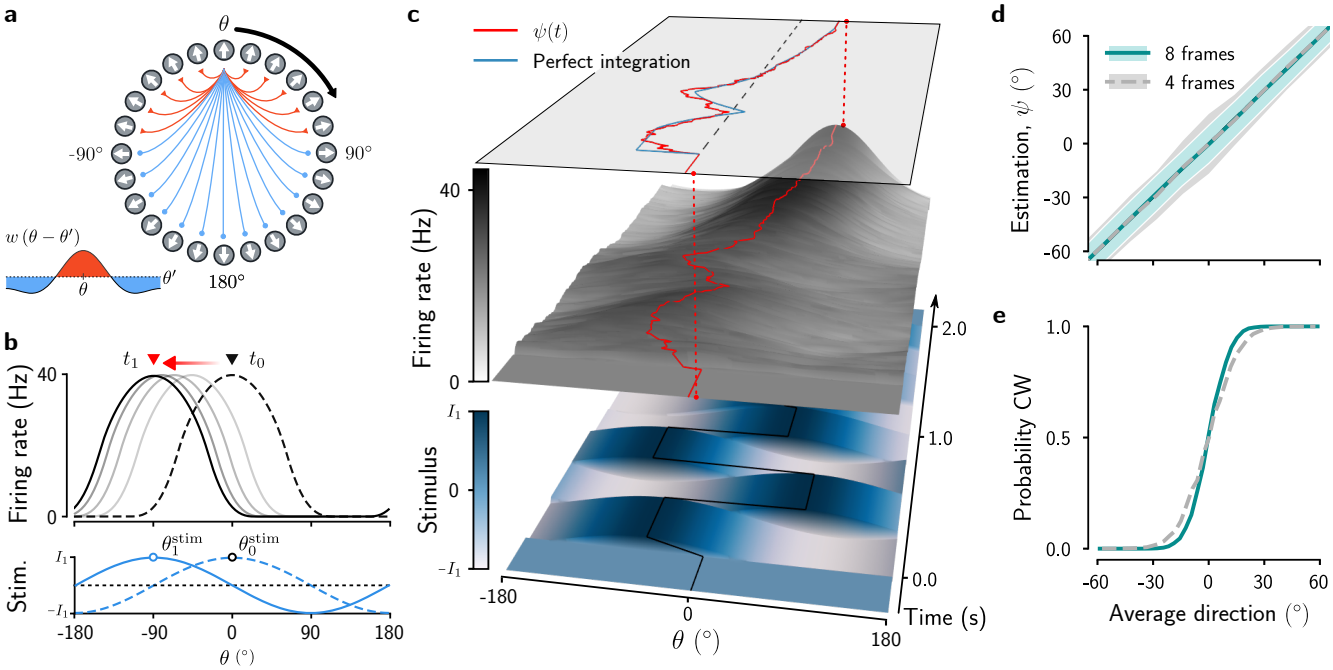



**Fig. 1 | Stimulus integration in the bump attractor network.** **a**, Ring network with Mexican-hat connectivity, that is, strong local excitation (red connections) and broader inhibitory coupling (blue). **b**, Network activity for neurons arranged according to their position in the ring (**a**). Due to strong recurrent connectivity, a bump of activity emerges in this network at a position determined by the external input and persists when the input is removed. For time-varying inputs, the activity bump (dashed line) moves towards a new position (solid black line). **c**, Network activity in a single trial. The initial activity of the network (middle panel) is spatially homogeneous and evolves into a bump while integrating a time-varying stimulus composed of eight oriented stimulus frames (bottom panel). Top: The phase of the bump (red) as a function of time closely tracks the running average of the orientations of the stimulus frames (blue). **d**, **e**, Continuous stimulus estimate (**d**) and probability of clockwise choices (**e**) as a function of the average stimulus direction for stimulus durations of 1 s (4 stimulus frames of 250 ms) and 2 s (8 stimulus frames). Categorical choices were obtained by converting positive angles to clockwise reports and negative angles to counterclockwise reports.

the evolution of the bump phase closely approximated the cumulative running average of the stimulus, that is, the time-average of the stimulus up to a given time point (Fig. 1c, top). Indeed, estimation curves, obtained by simulating many trials, show that, on average, the bump phase closely tracks a continuous estimate of the averaged sensory input (Fig. 1d). The estimation accuracy improves with the stimulus duration, as expected from an integration process (errorbars in Fig. 1d). This improvement can also be seen in psychometric curves obtained by converting the analog estimate to a categorical choice (Fig. 1e). In sum, these simulations show that the bump attractor network can integrate the stimulus over times much larger than the intrinsic time constant of the network ($\tau = 20$ ms, stimulus duration $T_{\text{stim}} = 2$ s).

## Dynamics of the bump

Motivated by the observation that the phase of the bump can track the average of the stimulus with striking precision (Fig. 1), we next sought to identify and characterize the neural network mechanisms underlying stimulus integration in the bump attractor network. To study the dynamics of the movement of the bump, we applied a standard perturbation method (Methods) and reduced the network model (Eq. 4)

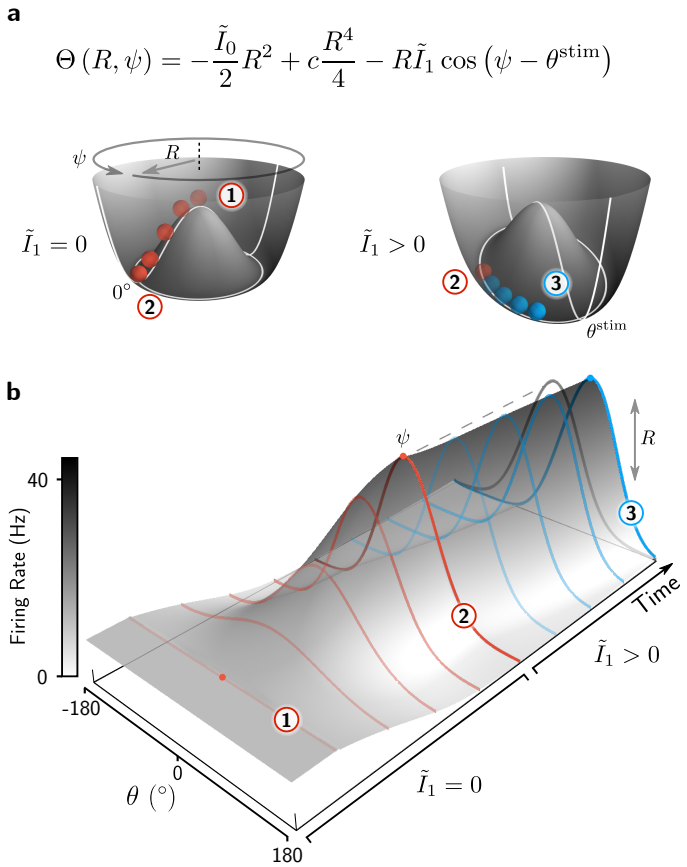



**Fig. 2 | Dynamics of the bump attractor network in the potential landscape. a**, The equation of the potential corresponding to the dynamics of the bump attractor (Eq. 1) and its geometrical representation. The movement of the red ball in each potential corresponds to the dynamics of bump formation and translation shown in **b**. **b**, Evolution of the network activity from the unstable homogeneous state ① to the stable bump state ② in the absence of external stimuli ($\tilde{I}_1 = 0$), followed by a translation of the bump (transition from ② to ③) towards the location of the stimulus $\theta^{\text{stim}}$ when the external stimulus is present ($\tilde{I}_1 > 0$). Note that for $\tilde{I}_1 = 0$, the bump forms at an arbitrary angle $\psi$, determined by noise fluctuations.

to a 2-dimensional differential equation for the amplitude, $R(t)$, and the phase of the bump, $\psi(t)$:

$$\tau\frac{\partial R}{\partial t} = \tilde{I}_1\cos\left(\psi - \theta^{\text{stim}}\right) + \tilde{I}_0 R - cR^3 + \xi_1(t) \tag{1a}$$

$$\tau\frac{\partial \psi}{\partial t} = -\frac{\tilde{I}_1}{R}\sin\left(\psi - \theta^{\text{stim}}\right) + \frac{\xi_2(t)}{R}, \tag{1b}$$

where $\tau$ is the neural time constant and $\tilde{I}_0$ is proportional to the global excitatory drive to each neuron, relative to its critical value at the onset of the bump ($\tilde{I}_0 \propto I_0 = I_{\text{exc}} - I_{\text{crit}}$; Methods). The constant $c$ depends on the synaptic connectivity profile and the nonlinear neural transfer function (Eq. 5). The time-varying stimulus input is characterized by its strength $\tilde{I}_1 \propto I_1$, and its orientation $\theta^{\text{stim}}(t)$. The noise terms $\xi_1(t)$ and $\xi_2(t)$ are related to the internal stochasticity and to fluctuations in the stimulus realized as independent noise inputs to each neuron in the full model (Eq. 4). The reduced model allows us to study the dynamics of the bump analytically. From Eq. 1b it directly follows that the rate of change of the phase (the angular speed $d\psi/dt$) depends on the ratio of the stimulus strength $I_1$ and the bump amplitude $R$. Thus, for a given stimulus strength $I_1$, the higher the amplitude of the bump the smaller will be the angular speed, that is, larger bumps are more sluggish (Extended Data Fig. 1). The steady-state bump amplitude is determined by the global excitatory drive $I_0$, ($R_\infty \propto \sqrt{I_0}$; Methods), indicating the key role of this parameter for the integration dynamics.

A deeper understanding about how a time-varying stimulus impacts the phase of the bump can be gained if we visualize the dynamics of the system, Eq. 1, by means of the physical idea of the potential energy. We can interpret the dynamics of the bump as a heavily damped particle sliding down the walls of a 3-dimensional potential well, where a potential $\Theta(R, \psi)$ of the system described in Eq. 1 is given by

$$\Theta(R, \psi) = -R\tilde{I}_1 \cos\left(\psi - \theta^{\text{stim}}\right) - \frac{\tilde{I}_0}{2}R^2 + \frac{c}{4}R^4. \tag{2}$$

The surface of this potential represents the dynamical manifold of Eq. 1, such that a point on this surface is a possible state of the activity bump, characterized by its amplitude $R$ and phase $\psi$ (Fig. 2). This interpretation allows us to describe the mechanism underlying the evolution of the bump graphically (Fig. 2).

In the absence of stimulus input ($I_1 = 0$), the potential $\Theta(R, \psi)$ resembles a juice-squeezer-shaped surface with an unstable fixed point at the center, corresponding to a bump with zero amplitude ($R = 0$), i.e. the unstable spatially homogeneous network state (Fig. 2; ①). The circular well of the manifold represents bump attractor states ($R > 0$) that are neutrally stable along the angular dimension $\psi$. The formation of the bump is described by the movement of the particle that, starting at the center of the juice squeezer, rolls downwards towards the stable manifold while increasing its distance to the center, $R$ (Fig. 2a, transition from ① to ②). The corresponding evolution of the network activity is shown in Fig. 2b, where the initially homogeneous network activity evolves into a bump state, reaching its steady-state amplitude at point ②.

When presenting a stimulus to the network ($I_1 > 0$), the radial symmetry of the potential is broken and a deeper region arises at the location of the stimulus $\theta^{\text{stim}}$ (Fig. 2a, right). If the particle has already reached the stable circular manifold at the bottom of the potential, the stimulus will force it to move towards this deeper state along the manifold (Fig. 2a, transition from ② to ③). In terms of the network activity, this corresponds to a movement of the bump towards $\theta^{\text{stim}}$, accompanied by only slight changes in bump amplitude (Fig. 2b). Solving Eq. 1 for small stimulus strengths, we found that the evolution of the phase of the bump depends approximately linearly on time and on the difference between the bump phase and the stimulus direction (Extended Data Fig. 2 and Methods). In contrast, if the particle has not yet reached the stable manifold, i.e., when a stimulus is applied during the initial bump formation, it will cause both a change in bump phase and, together with the internal dynamics, a change in bump amplitude.

In general, successive stimulus frames of a time-varying stimulus drag the bump towards their respective orientations in a continuous fashion (Fig. 1c), with an angular displacement that depends on the current bump amplitude (Extended Data Fig. 1). In sum, we can distinguish three distinct stages of the evolution of the bump during which it becomes increasingly difficult for the stimulus to impact the phase of the bump. Initially, starting from homogeneous network activity, the stimulus orientation strongly impacts the phase of the emerging bump. Second, the bump grows in amplitude, driven by the internal dynamics and the stimulus, and thus successive stimulus frames become less and less effective in moving the bump. Third, when the bump amplitude reaches its steady state, it cannot grow further and the impact of subsequent stimuli becomes constant over time. Equipped with these theoretical results, we can now study different stimulus integration regimes in the network.

## The ring model shows a variety of temporal evidence integration regimes

Our theoretical results showed that the changes in bump phase caused by a time-varying stimulus critically depend on the bump amplitude, suggesting its key role in the integration of sensory stimuli. To investigate this, we ran simulations in which networks with different steady-state bump amplitudes had to estimate the average direction of a sequence of oriented frames, distributed between $-90°$ and $+90°$ as in Fig. 1c. To quantify the dynamics of evidence integration, we used the psychophysical kernel (PK) which measures the influence of sensory evidence during the course of the stimulus (see Methods).

We first considered the case where a trial starts with an initial bump (i.e., the particle is at a position on the circular attractor well of the potential; point ② in Fig. 2), whose amplitude we set by varying

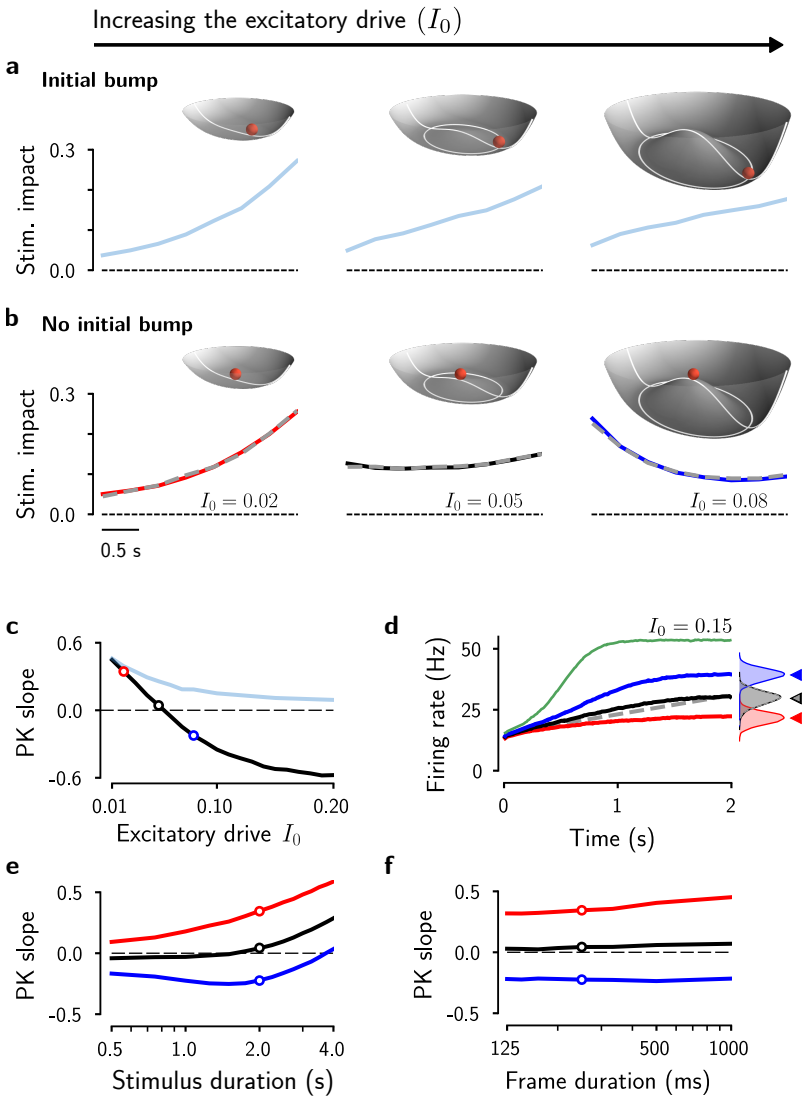



**Fig. 3 | The global excitatory drive $I_0$ determines the integration regime of the network.** **a**, Psychophysical kernels (PKs) for increasing values of $I_0$ (bottom, from left to right, $I_0 = 0.02$, $0.05$ and $0.08$) with a fully formed bump as the initial condition. The PKs describe the impact of each stimulus frame on the estimated average orientation. Top: For small values of $I_0$ the potential takes the form of a steep paraboloid where the bump has little space to grow (top left). Intermediate values of $I_0$ transform the potential into a wide surface similar to a dish with a relative flat central area (top middle). Increasing $I_0$ widens the potential and gives rise to an increasingly deeper circular manifold at the bottom of the juice squeezer (top right). **b**, PKs as in **a** but with spatially homogeneous initial conditions (flat network activity; particle at the center of the potential). PKs were obtained from numerical simulations of the ring attractor network (colored lines) and from simulating the amplitude equation (dashed gray lines). **c**, PK slope as a function of the excitatory drive $I_0$, for the spatially homogeneous (black line) and inhomogeneous (blue line) initial conditions. **d**, Evolution of the bump amplitude for different values of $I_0$. Colored lines correspond to values of $I_0$ used in **b**. The evolution of the vector length in the PVI is shown for comparison (dotted gray line). Inset: distribution of the bump amplitude at the end of the trial. The histograms for $I_0 = 0.05$ and for the PVI (in black) perfectly overlap. **e**, PK slope as a function of the stimulus duration, with stimuli composed of frames with a fixed duration of 250 ms. PKs are shown in Extended Data Fig. 5a. **f**, PK slope for a fixed stimulus duration of $T_{\text{stim}} = 2$ s as a function of the duration of a stimulus frame. The number of frames ranges from 2 (frame duration 1000 ms) to 16 (frame duration 125 ms).

the global excitatory drive to the network $I_0$ and keeping all other parameters fixed. Spontaneous bump states have been observed in prefrontal and visual cortex[38–40], and we show further below how they can give rise to biases in the estimation process (Fig. 5). The PKs we obtained for different values of $I_0$ are qualitatively similar (Fig. 3a): they all rise, indicating that the stimulus frames later in the trial have more impact on the final estimate (i.e., the bump phase $\psi$ at the end of the trial) than early stimulus frames. This commonly called "recency" effect becomes weaker as $I_0$ increases, yielding a decrease of the PK slope (Fig. 3c, blue line). The reason for the general over-weighting of late stimulus information is not trivial and deserves a closer inspection. Essentially, it occurs because the bump amplitude is fixed and each successive stimulus frame leads to a partly discounting of the previous estimate of the stimulus average (leaky integration).

To fully understand the recency effect, it is necessary to consider how the running average of the stimulus direction can be tracked optimally. Mathematically, we want to keep track of the circular mean of the time-varying stimulus direction $\theta^{\mathrm{stim}}(t)$, which requires computing the vector sum of the stimulus vectors, each defined by their strength $I_1$ and their direction $\theta^{\mathrm{stim}}$ (Extended Data Fig. 3a,c; Methods). To compare this optimal solution to the bump attractor network, we derived a dynamical system that continuously tracks the cumulative circular average, the *perfect vector integrator (PVI)*. The corresponding two-dimensional equation (Eq. 22 in Methods) turned out to be nearly identical to the amplitude equation but without the intrinsic dynamics of the bump amplitude (the two terms depending on $R$ in Eq. 1a). This remarkable equivalence indicates that any deviation from uniform evidence weighting (as in the PVI) must be due to the non-linear amplitude dynamics of the bump attractor model. To gain an intuition on how the amplitude dynamics affects evidence integration, we can again consider the movement of a particle in a potential which for the PVI takes the form of a plane going through the origin and tilted towards $\theta^{\mathrm{stim}}$

$$\Theta^{\mathrm{PVI}}(R,\psi) = -RI\cos\left(\psi - \theta^{\mathrm{stim}}\right). \tag{3}$$

In response to each stimulus the particle moves in both the radial $R$ direction and the angular $\psi$ direction, thus integrating stimulus vectors. How $R$ evolves over time in the PVI depends on the particular combination of the directions of stimulus frames in each trial, but, on average, $R$ will grow over time because of a net movement in the direction of the mean of the stimulus distribution $\bar{\theta}^{\mathrm{stim}}$ (Extended Data Fig. 3b,d show how the growth rate depends on the width of the stimulus distribution). Because $R$ is growing, the angular updates become smaller with each stimulus frame (in other words, as the number of samples grows, the running average direction increasingly relies on the already accumulated information). This is in contrast to the integration process in the bump attractor network with a formed bump in which the bump amplitude is fixed (determined by the shape of the potential; Fig. 2a) and integration occurs only in the angular direction. Because of this, the network gives equal weight to the angular updates over time, thus partly erasing the accumulated stimulus information while over-weighting late stimulus information. The strength of the recency effect depends on the bump amplitude. As described in the previous section, the smaller is $I_0$ and thus the bump amplitude, the larger is the angular displacement for a given stimulus strength $I_1$ and the faster the bump can track the orientation of incoming stimuli, resulting in a PK with increasing recency effect (Fig. 3a,c; Extended Data Fig. 1).

Strikingly, when we ran simulations in which the network had to estimate the average stimulus direction starting the trials in the homogeneous network state, with no bump formed (i.e., the particle is at the unstable center position on the potential; point ① in Fig. 2), we observed three qualitatively distinct integration regimes depending on the global excitatory drive to the network $I_0$ (Fig. 3b; Supplementary Videos 1 to 3). In the first regime, a small $I_0$ yields again a recency PK (Fig. 3b, left). Here, the steady-state bump amplitude is relatively small, and the bump can only grow very slightly over time (Fig. 3d, red line). Thus, stimulus integration is limited by the small bump amplitude and the network behaves similarly to the previously considered case with the bump formed from the beginning of the trial and with constant amplitude throughout the trial (Fig. 3a, left).

Second, for large $I_0$, the networks showed a "primacy" effect, that is, early stimulus information now had a higher impact on the final estimation than later stimulus information (Fig. 3b, right; Supplementary Video 3). This can be understood by considering that in this regime the bump amplitude is small at the

beginning of the trial but then grows to a large amplitude (Fig. 3d, blue line). Thus, the bump can quickly follow the stimulus direction early in the trial during the formation of the bump, but while growing it becomes increasingly sluggish and stimulus information becomes comparably less effective in displacing the bump. Compared to the PVI, the bump amplitude grows too quickly because it is driven by the internal dynamics in addition to the stimulus (i.e., the potential is not flat as for the PVI but has a peak in the center; Fig. 3b, right inset). The resulting down-weighting of later stimulus frames only occurs while the bump is growing (transition from ① to ② in Fig. 2). Once the steady-state bump amplitude has been reached (corresponding to the attractor well of the potential), the model starts to overweight incoming evidence and, as a consequence, for very long stimulus durations the model will eventually show a recency PK (Fig. 3e, blue line).

Third, for intermediate values of $I_0$, primacy and recency effects balance out and we obtained an almost perfectly uniform PK (Fig. 3b, middle; Supplementary Video 1). Thus, in this regime the directions of all stimulus frames had the same impact on the final estimate over the whole trial (PK slope $= 0$; Fig. 3c). In this case, the potential has only a relatively small peak at the center, approximating the flat potential of the PVI (Fig. 3b, middle inset). The growth rate of the bump amplitude is mainly driven by the stimulus as in the PVI (Extended Data Fig. 3d). The value of $I_0$ in which the model best approximates the PVI depends on the stimulus strength and the distribution of the stimulus directions (Extended Data Fig. 3e). Approximately perfect integration is only possible as long as the bump amplitude can increase (before the well of the potential is reached), such that for long stimulus durations integration becomes again leaky and we obtain again a recency PK (Fig. 3e, black line). Thus, both primacy and uniform PKs are robustly observed for small and intermediate stimulus durations (Fig. 3e; $T_{\mathrm{stim}} \leq 2$ s) while for long durations the recency regime is recovered. The results in Fig. 3 were obtained with time-varying stimuli that change their orientation every 250 ms (frame duration). The PKs remain qualitatively the same when keeping the stimulus duration constant and presenting a variable number of stimulus frames of different duration (Fig. 3f).

## Accuracy of stimulus estimation and categorization

We next quantified the estimation and choice accuracy in the different temporal integration regimes of the bump attractor model when starting the integration without an initial bump of activity. We computed stimulus estimation curves by reading out the bump phase from the network activity at the end of the trial, averaging the estimates across trials, and plotting them as a function of the actual mean orientation $\bar{\theta}^{\mathrm{stim}}$ (Fig. 4a). We noticed that independent of the bump amplitude, that was controlled by varying the excitatory drive $I_0$, the average of the estimates was always close to the true mean direction. We quantified this by computing the estimation bias for $\bar{\theta} = 45°$ and obtained $b_{\mathrm{est}}(45°) < 1°$ for all values of $I_0$. Consistent with this, the slopes of the estimation curve obtained from regression analysis (Methods) were $k_1 \approx 1$ in all cases (Fig. 5d, red line). The estimates computed by the ring attractor network were thus approximately unbiased, independent of the temporal weighting of stimulus evidence. The unbiased estimates are a direct consequence of starting the integration process in the neutral center of the potential and the symmetry of the potential (Fig. 2, point ①). The direction of the first stimulus frame determines the initial phase of the forming bump in an unbiased manner. Subsequently, the bump will move, but as long as the directions of the stimulus frames are drawn i.i.d. from some underlying distribution, clockwise and counterclockwise bump displacements are equally likely and the estimate always remains unbiased irrespective of the integration regime.

However, the trial-to-trial variability of the estimates (Fig. 4a, shaded areas and Fig. 4c, black line) depended on the integration regime. Several factors could give rise to estimation errors in the attractor model and may thus contribute to this dependence: (1) internal noise and stimulus fluctuations of magnitude $\sigma_{\mathrm{OU}}$ (Methods), (2) sub-optimal evidence integration due to non-uniform evidence integration (i.e., non-uniform PKs), and (3) additional impact of the non-linear internal dynamics on evidence integration not captured by the PK. We first considered the impact of the noise. We ran simulations of the network without noise ($\sigma_{\mathrm{OU}} = 0$) and obtained, as expected, a general decrease of the estimation errors (Fig. 4c, orange line). We next tested to what degree the remaining estimation errors can be accounted for

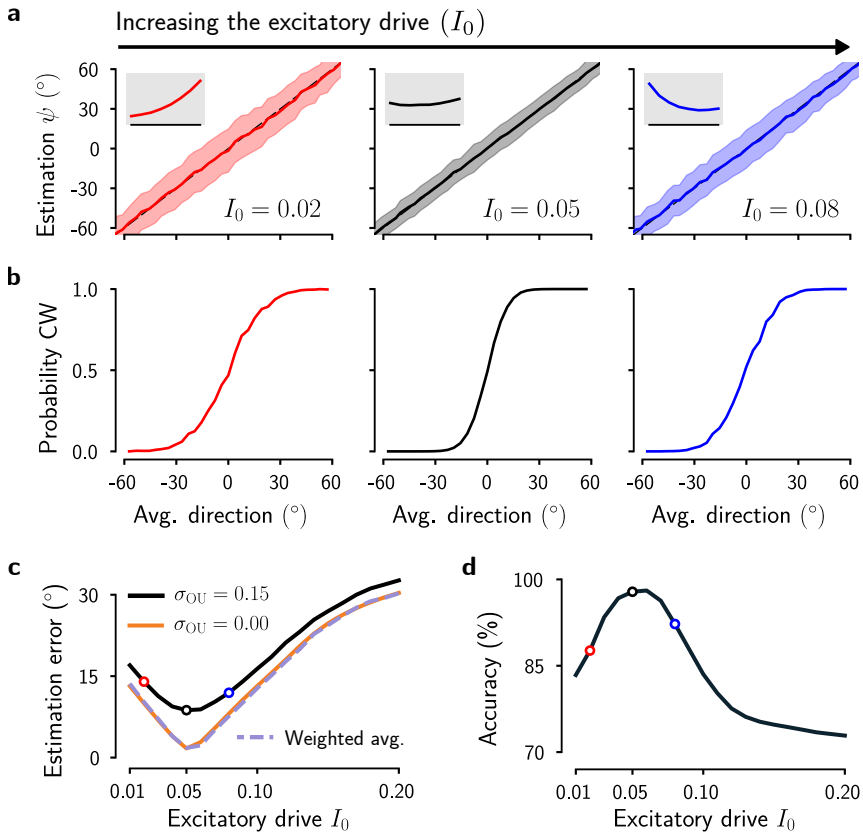



**Fig. 4 | Dependence of estimation and categorization accuracy on the global excitatory drive $\mathbf{I_0}$.** **a**, Estimated average direction as a function of the true average of a stimulus composed of 8 oriented frames, for increasing values of $I_0$ (from left to right, $I_0 = 0.02, 0.05$ and $0.08$). Solid lines indicate the average across trials and the shadings indicate s.d. Insets: corresponding PKs from Fig. 3b. **b**, Probability of a clockwise choice versus average stimulus orientation, for increasing values of $I_0$. A clockwise/counterclockwise choice corresponds to a positive/negative phase of the bump at the end of the trial. **c**, Root-mean-square error (Methods) of the ring attractor network with noise ($\sigma_{\mathrm{OU}} = 0.15$) and without noise ($\sigma_{\mathrm{OU}} = 0$) as a function of the global excitatory drive $I_0$. The estimation error for $\sigma_{\mathrm{OU}} = 0.15$ (black line) corresponds to the estimations shown in **a**. The estimation error for a perfect weighted average of the stimulus directions (with the same temporal weighting as the network model but otherwise optimal; Methods) is shown for comparison. Note that the curve corresponding to $\sigma_{\mathrm{OU}} = 0$ has been shifted such that for both conditions the critical value of $I_{\mathrm{exc}}$ at the bifurcation point is aligned (Methods). **d**, Categorization accuracy as a function of the excitatory drive $I_0$, for a fixed average stimulus orientation $\bar{\theta}^{\mathrm{stim}} = 23°$.

by the non-uniform evidence weighting in the primacy and recency regimes. To show this we compared the dependency of the estimation error on $I_0$ of the attractor model with the estimation error that was obtained when computing the true average of the stimulus directions with the PVI but weighted with the PKs from the bump attractor model (Fig. 4c, purple dashed line). The obtained estimation errors are almost indistinguishable from the errors of the noise-less bump attractor network (Fig. 4c). Thus, the increase of estimation errors for non-optimal $I_0$ (corresponding to values of $I_0$ different from $0.05$ in Fig. 4c) can be attributed to the suboptimal, non-uniform, temporal weighting captured by the PK. Because the non-linear internal dynamics of the ring model can only yield an approximately uniform PK, the estimation errors are non-zero even for the best $I_0$ for a given stimulus distribution, in particular when the stimulus distribution is wide (Extended Data Fig. 3e). Moreover, the model PKs change with the duration of the stimulus and thus the estimation error also shows a time-dependence (Extended Data Fig. 5). Together, these analyses show that the estimation accuracy in the bump attractor network is limited by noise and by the non-linear internal dynamics causing non-uniform evidence integration, whereas further contributions of non-linearities (that are not captured by the PK) are negligible.

A direct consequence of the trial-to-trial variability of the estimates is that they limit the accuracy of a categorical decision that is based on the read-out of the bump phase. This can be seen when computing psychometric curves by interpreting a positive bump phase at the end of the trial as a clockwise choice and a negative bump phase as a counterclockwise choice, respectively (Fig. 4b). As expected from the stimulus estimation results, the psychometric curve for the uniform integration regime had the highest slope (Fig. 4b), and decision accuracy is maximal for the $I_0$ that gives rise to a uniform PK (Fig. 4d).

## Initial bump states cause an estimation bias

We next investigated in more detail the accuracy of stimulus estimation in the network with an initially formed bump (Fig. 3a), in particular the impact of the initial bump phase on the estimation process. In psychophysical experiments this initial bump position could be determined by a reference line shown before the onset of the stimulus[2,4] or it could be related to the subject's prior expectation before the start of the stimulus[5]. Single trials are illustrative of the effect of the different initial conditions (Fig. 5a,b). When starting a trial in the homogeneous network state, the first stimulus frame determines the phase where the bump emerges (Fig. 5a), and the bump tracks an estimate of the average direction as the trial progresses. As described above, the network's estimates were unbiased in this case (Fig. 5c,d; red line). In contrast, when starting with a fully formed bump at a particular position as the initial condition, the bump phase was biased towards the initial phase of the bump (Fig. 5b,c). As the network integrates the stimulus, the bump moves successively closer to the true average orientation, leading thus to a decrease of the systematic deviations from the true average with stimulus duration (Fig. 5c, blue lines). The magnitude of the estimation bias for a fixed stimulus duration of 2 s increased with the global excitatory drive $I_0$ to the network (Fig. 5d). The reason is that a higher $I_0$ yields larger bumps that are more sluggish and thus need more time to overcome the bias caused by the initial bump position. While the variability of the estimates in the biased case is generally lower than in the unbiased case, the overall estimation error (bias + variance) is higher (Fig. 5e). To confirm that the bias is actually caused by the pre-determined phase of the activity bump at the beginning of the trial and not a consequence of starting the trials with a fully formed bump, we ran simulations with a fully formed initial bump with a phase determined by the first stimulus frame and indeed observed no bias (Extended Data Fig. 4).

In sum, starting with an initial bump led to recency PKs (Fig. 3c) and, when setting the initial phase to some pre-determined value it yielded estimates that were biased towards the initial bump phase. It is worth noting that here we considered two extreme cases, either starting in the neutral condition (no initial bump) or with a fully formed initial bump. Intermediate scenarios (e.g. starting with a partially formed bump of lower amplitude or starting with a fully formed bump in only a fraction of trials) would lead to PKs, biases and estimation errors that lie between the extreme cases.

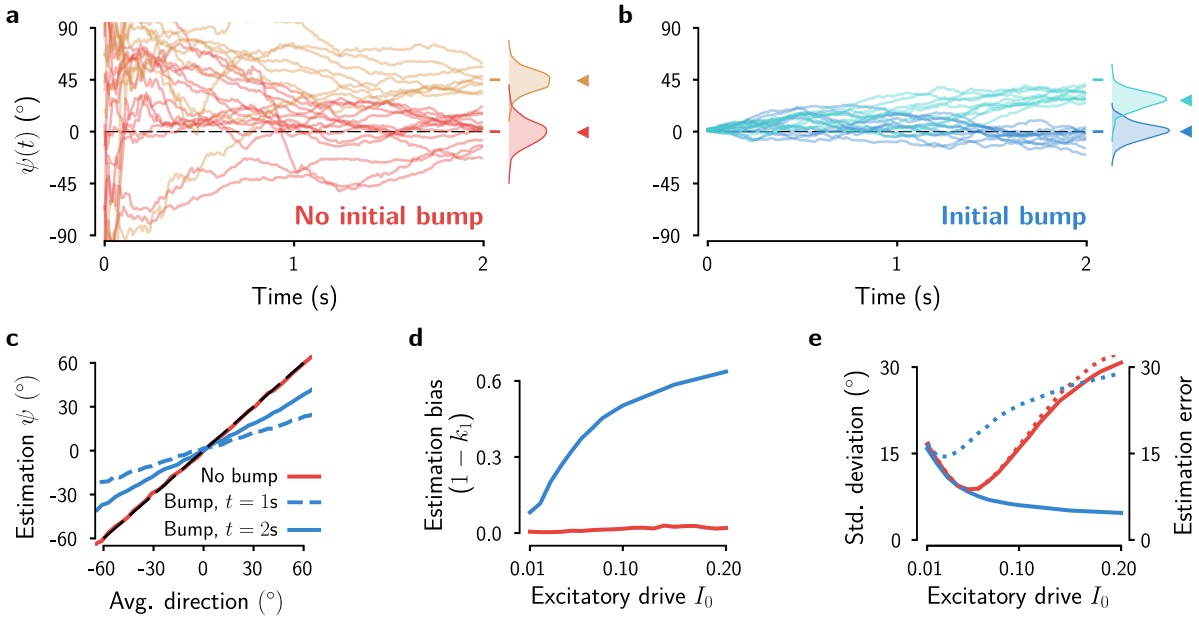



**Fig. 5 | Impact of the initial condition on the bias and variance of the stimulus estimates. a**, Bump phase in example trials with 0° and 45° average stimulus directions ($N = 8$ stimulus frames, stimulus duration $T_{\text{stim}} = 2$ s) when starting the trial with homogeneous network activity (i.e., no bump formed). Inset: distribution of the phase of the bump at the end of the trial. **b**, Bump phase for the same network and the same stimuli as in **a**, but starting the trial with a fully formed bump centered at 0°. **c**, Continuous estimations as a function of the mean stimulus direction $\bar{\theta}^{\text{stim}}$ for two different stimulus durations (no bump condition in red, bump condition in blue). For the no-bump condition, the average estimation curve does not depend on stimulus duration (red line) and it overlaps with the diagonal (dashed black line). **d**, Systematic bias of the estimates as a function of the global excitatory drive $I_0$ to the network (stimulus duration $T_{\text{stim}} = 2$ s). The estimation bias is defined as one minus the slope of the estimation curve ($k_1$ in Eq. 32, Methods). **e**, Standard deviation (solid lines, left axis) and root-mean-square estimation error (dotted lines, right axis) of the final bump position as a function of the global excitatory drive $I_0$ for the same trials as in **d**.

## The ring model explains heterogeneity in integration dynamics in humans

We tested whether human subjects show variations in their integration dynamics in a stimulus averaging task and whether the ring model could parsimoniously account for those variations (Fig. 6). In this category-level averaging task, a visual stream of eight oriented Gabor patters was presented to participants[6] (Fig. 6a). Each frame had an orientation between $-90°$ and $90°$ and a fixed duration of 250 ms, conforming a stimulus with total duration of $T_{\text{stim}} = 2$ s. At the end of the stream participants reported whether, on average, the tilt of the eight frames fell closer to the cardinal $\{0°, 90°\}$ or diagonal $\{-45°, 45°\}$ axes. On average, participants were able to combine the evidence favouring either option improving their accuracy in trials with higher evidence[6,7,41].

We merged the data from three studies[6,7,41] carried out using very similar experimental conditions ($n = 61$ subjects, Methods). We computed their individual PKs (Methods) and classified the subjects according to their temporal weighting behavior (primacy, uniform, recency). As previously reported[7], across participants there was a tendency for late temporal weighting (recency). However, subject-by-subject analysis revealed a broad range of integration dynamics (Fig. 6b,c; individual PKs are shown in Extended Data Fig. 6). The majority of participants weighted more the sensory evidence during the late periods of the stimulus (33 out of 61 participants), yet a substantial proportion of the participants weighted the evidence approximately uniformly (20 out of 61 participants). Finally, a minority of participants (8 out of 61 participants) weighted early frames of the stimulus more that late frames, showing a primacy PK. Qualitatively, the ring model could capture these different temporal integration dynamics observed in the

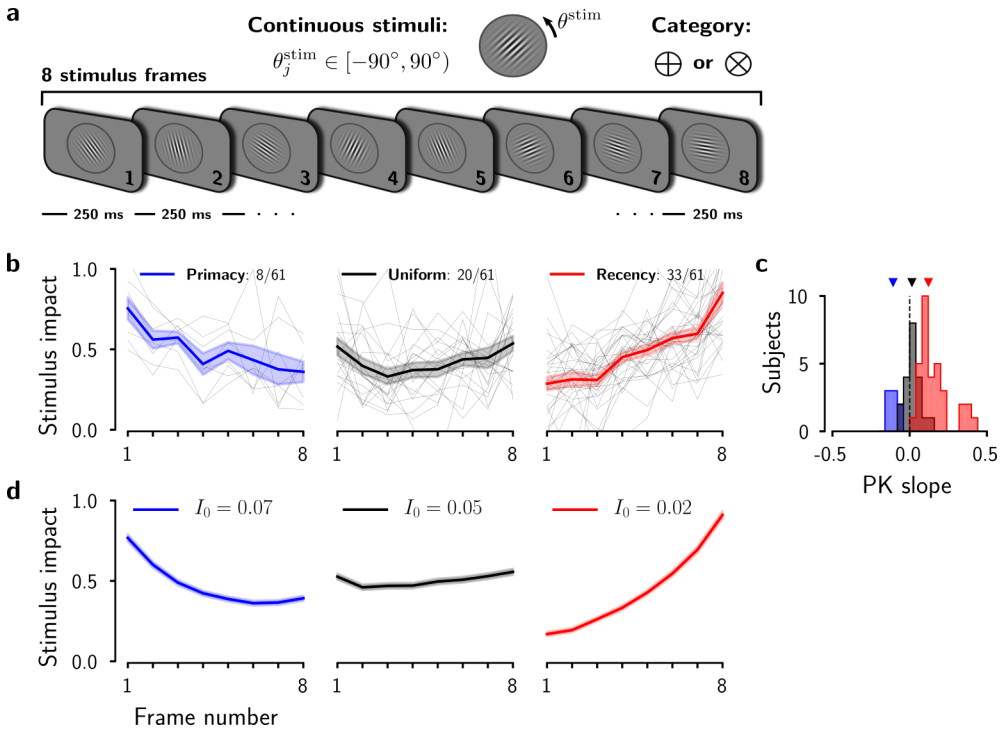



**Fig. 6 | The ring attractor model accounts for experimentally observed psychophysical kernels. a**, Category-level averaging task[6,7,41]. Participants reported whether, on average, the tilt of eight oriented Gabor patterns fell closer to the cardinal or diagonal axes. **b**, PKs obtained from human subjects were heterogeneous. Thin lines represent the PKs of individual subjects and colored lines the group averages. **c**, Distributions of the slopes of the PKs for each type of integration dynamics. The top triangle indicates the median for each case. **d**, PKs obtained by fitting the ring model to the experimental data by varying the excitatory drive $I_0$.

psychophysical experiments (Fig. 3). To assess whether the model could explain the data in a quantitative manner, we performed simulations with the same stimulus statistics as in the behavioral experiment. We systematically adjusted both the global excitatory drive to the network, $I_0$, and the noise level, $\sigma_{\text{OU}}$, such that the ring network reproduced, on average, the experimental results (Fig. 6b,c), characterized by the slope of the across-subjects PKs (Fig. 6b) and the average performance (not shown) for each integration regime (primacy, uniform and recency). The inter-subject heterogeneity in temporal evidence weighting could thus parsimoniously be explained by varying the the overall excitatory drive that determines the amplitude of the bump, while adjusting the noise fluctuations to match the performance of subjects.

It is conceivable that the global excitatory drive does not only differ between individuals but that it could also be modulated on a trial-to-trial basis, such that it may therefore change the time-course of evidence integration depending on the current task conditions. Such a task-driven change has previously been observed in a similar task as in Fig. 6 that studied the mechanisms of perceptual choices under focused and divided attention[41]. In that task, two spatially separated stimulus streams were presented simultaneously and subjects had to integrate either a single or both streams in a trail-to-trial basis. We built a network model consisting of two ring circuits (Methods) and found that the results of this experiment can be explained by a change in the excitatory drive that controlled the allocation of a fixed amount of resources between the two circuits (Extended Data Fig. 7). In sum, our model suggests that the overall excitatory drive determines the integration regime. Control signals like top-down attention signals or neuromodulatory gain changes can impact the dynamics of evidence integration and the model makes specific predictions about the underlying changes in neural activity.

## Bump attractor dynamics links integration dynamics and reaction times

Finally, to derive specific predictions from the model that go beyond the shape of the PK, we extended the model by coupling the ring integration circuit to a canonical decision circuit and studied the decision build-up in the different evidence integration regimes (Fig. 7). In this two-circuit model (Fig. 7a; Methods), evidence about the average stimulus direction is integrated in the phase of the activity bump in the integration circuit and then converted to a categorical choice in the decision circuit. The decision build-up is triggered by an urgency (or choice commitment) signal at the offset of the fixed-duration stimulus.

We observed that the reaction time (RT), from stimulus offset to reaching a categorical choice, depends on the trial outcome (correct or incorrect) and on the accumulated stimulus evidence (Fig. 7b, bottom). RTs for correct trials are shorter and they decrease with stimulus strength, whereas RTs for incorrect trials are longer and they increase with stimulus strength. This can be understood by considering how the decision build-up in the decision circuit depends on the position of the activity bump in the integration circuit. For a given average stimulus direction, the activity bump moves further away from the decision boundary in correct trials than in incorrect trials (Fig. 7b, middle). Thus, one of excitatory decision populations (**L** or **R**) receives a stronger input signal than the other population and, due to winner-take-all competition, its firing rate rapidly increases, reaching the decision threshold quickly (Fig. 7b, top). In incorrect trials, the resulting input to the **L** and **R** populations is more similar and the decision build-up takes longer. This effect gets more pronounced for larger average orientations (i.e., the bump moves further away from the decision boundary in correct trials and less so in incorrect trials). We found that both signatures, the decrease of RTs with accumulated stimulus evidence for correct trials and the increase of RTs for incorrect trials, are also present in the experimental data obtained from human subjects (Fig. 7e).

Moreover, the model predicts a direct relationship of RTs and the shape of the psychophysical kernel (Fig. 7c). Different psychophysical kernels in the model are obtained by varying the excitatory drive to the network (Fig. 6). Increasing the excitatory drive leads to an increase of the bump amplitude and a transition from recency over uniform to primacy weighting (Fig. 3b,c). The bump amplitude, in turn, affects the RT in the two-circuit model (Fig. 7c). RTs are larger for small bump amplitudes because the decision circuit receives weaker input and takes therefore longer to reach the decision threshold. Taken together, the model predicts that primacy temporal weighting should be accompanied by shorter RTs and recency weighting by longer RTs, i.e., RT should increase with the slope of the PK (Fig. 7c, bottom). We tested this non-trivial prediction in the experimental data and indeed found a significant correlation between the subject's RTs and PK slopes (Fig. 7f; $r = 0.39$, $P = 0.006$, $n = 47$, $t$-test). It is unlikely that the differences in RTs could be explained by a different fraction of correct trials depending on the PK slope because we did not find a correlation between PK slope and the fraction of correct trials (Fig. 7d; $r = 0.04$, $P = 0.774$, $n = 47$, $t$-test). In sum, the two-circuit model provides a mechanistic link between neural population activity, evidence integration dynamics and reaction times.

# Discussion

We investigated how the classical continuous ring attractor model[28,30,33] integrates a time-varying stimulus in the amplitude and the phase of the network's activity bump. We found that the model can approximately compute the time integral of stimulus vectors defined by the strength and the direction of the stimulus. This operation can be understood in terms of a population vector[42]. But while a population vector is typically used to read out direction information from a neural population (by weighting the activity of individual neurons according to their preferred direction), the population vector encoded in the amplitude and the phase of the activity bump results from a temporal integration process and is continuously updated in response to incoming stimulus evidence.

We studied the neural network mechanism underlying this integration process and identified how the model integrates a sequence of stimulus frames of varying direction and constant strength, e.g. a sequence of visual gratings with different orientation but the same contrast. Analytically and through

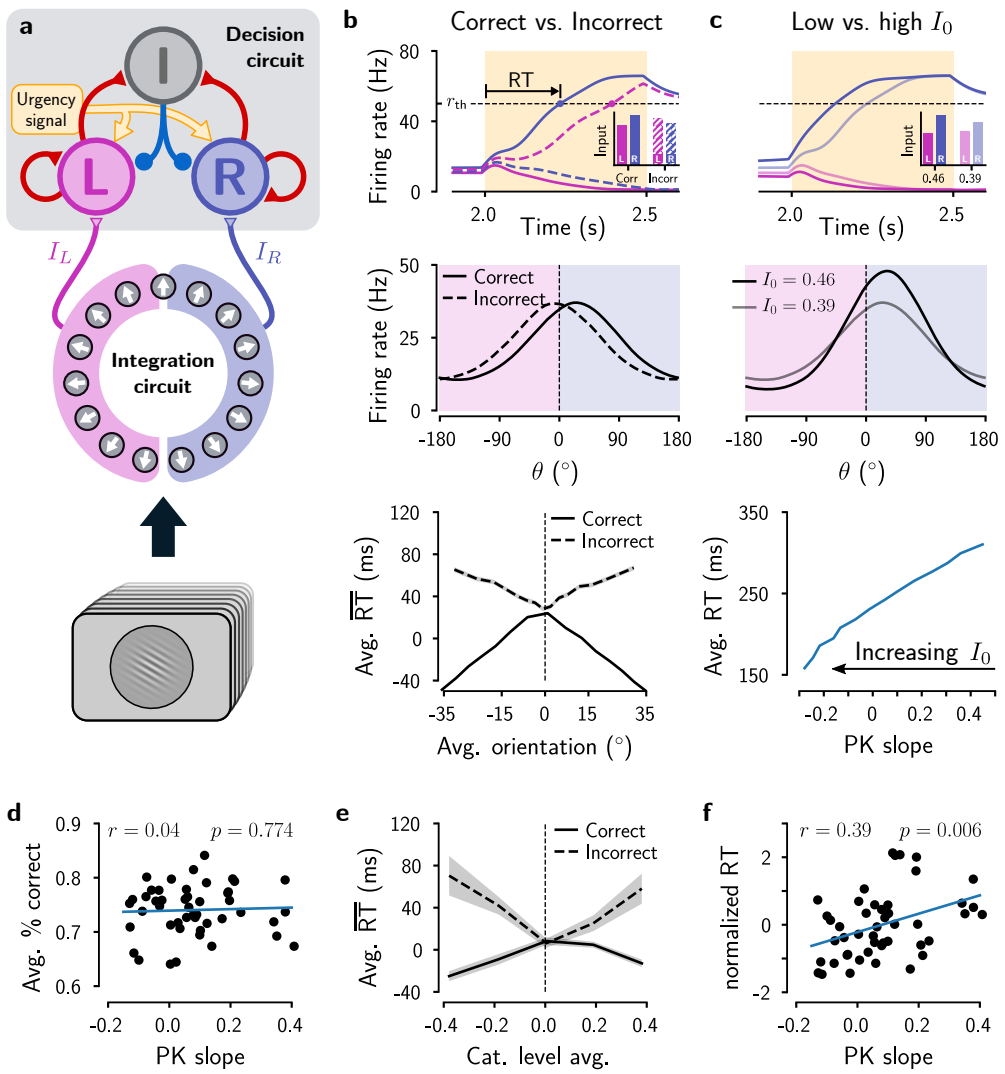



**Fig. 7 | Dependence of reaction times on evidence strength and on the shape of the psychophysical kernel. a**, Ring integration circuit coupled to a categorical readout circuit (decision circuit; Methods). Left (**L**) and right (**R**) excitatory populations in the decision circuit receive inputs, $I_L$ and $I_R$, from their corresponding side of the ring. The inhibitory population (**I**) promotes winner-take-all competition[8]. The decision process is prompted by the activation of an urgency signal at the end of the stimulus presentation that initiates the competition. **b**, Dependence of reaction times (RTs) on choice outcome and stimulus evidence in the model. Top, average activities of the **L** and **R** populations for correct (solid lines) and incorrect (dashed lines) trials for stimuli with average direction of 15°. RT is measured from the onset of the urgency signal (at $t = 2$ s; shaded area) until one of the firing rates reaches the threshold $r_{\text{th}} = 50$ Hz. The inset shows the summed bottom-up inputs, $I_L$ and $I_R$. Middle, average activity of the integration circuit at the end of the stimulus presentation. Bottom, Dependence of RTs on stimulus evidence for correct and incorrect trials. **c**, Dependence of RT on the excitatory drive $I_0$ in the model. Top and middle panels as in **b** but for two different levels of the excitatory drive and correct trials only. Bottom, Average RTs as a function of the PK slope. **d-f,** Performance and RTs in human subjects ($n = 47$). **d**, The percentage of correct trials is not correlated with the PK slope ($r = 0.04$, $P = 0.774$, $t$-test). **e**, Average RTs as a function of the category level average for correct and incorrect trials. RTs for each condition were computed relative to the trial-averaged RT across all trials, $\overline{\text{RT}} = \text{RT} - \langle \text{RT} \rangle$. Errorbars indicate the SEM across subjects. **f**, Subject's RTs (z-scored values for each data subset, Methods) vs. their PK slopes show a significant correlation ($r = 0.39$, $P = 0.006$, $t$-test).

simulations we confirmed that the phase of the activity bump tracks the running circular average of the stimulus orientations (for variable stimulus strengths we would get a weighted circular average). Concurrently, the amplitude of the bump evolves depending on the dispersion of the stimulus directions. The dynamical regime in which the model closely approximates perfect vector integration depends on the stimulus statistics as well as the stimulus strength and duration (Extended Data Fig. 3e; Extended Data Fig. 5b). It is characterized by a relatively wide and shallow potential so that the dynamics of the bump is dominated by the stimulus and not by the non-linear internal dynamics (Fig. 3). The limiting factor imposed by the internal dynamics of the network is that the bump amplitude cannot grow beyond a maximum value that depends on the model parameters and is proportional to the overall excitatory drive to the network. When the maximum bump amplitude is reached, which will eventually happen for long stimulus durations, the model has reached the neutrally stable ring attractor and can still track the average stimulus phase. However, the model is then over-weighting later stimulus frames, similar to a leaky integrator. In addition to this recency effect, the model can also show primacy temporal weighting when the internal network dynamics contributes to a rapid initial growth of the bump and initial stimulus frames have a relatively higher impact on the evolution of the bump phase. Independent of the temporal weighting regime, the average orientation estimates of the model were always unbiased; the variance of the estimates is however minimal when the model acts as perfect vector integrator and increases for non-uniform temporal weighting (Fig. 4). In sum, we have uncovered the fundamental mechanism underlying stimulus integration in the phase and the amplitude of the bump of the ring attractor model and the corresponding computation of the running circular mean. Our results suggest ring attractor dynamics as a potential neural mechanism for the estimation of average circular stimulus features and for perceptual discriminations based on the same features.

## The role of the global excitatory drive

The key parameter that controls the integration dynamics in the bump attractor model is the global excitatory drive to all neurons in the network, $I_0$ (Fig. 3). Keeping all other parameters fixed, $I_0$ determines the depth and the width of the potential (Extended Data Fig. 1) and thus the maximal bump amplitude and the intrinsic network dynamics. We showed that by varying $I_0$, the model could account for the heterogeneity in how human observers weighted sensory evidence across a stream of oriented stimulus frames (Fig. 6) as well as for their reaction times (Fig. 7). In addition to a different level of global excitatory drive, subjects are likely to show other differences as well, for example a different E/I ratio or a different gain of the neural circuits involved in evidence integration. Effectively, in the model these inter-subject differences can be captured by different values of $I_0$. In particular, it can be shown that changing the slope of the neuronal transfer function is mathematically equivalent to a change in $I_0$. The bump attractor model can also account for inter-subject differences in evidence weighting by a less parsimonious explanation that involves the joint modulation of the global excitatory drive, the internal noise and the initial state of the network. That is, constraining $I_0$ to a narrow range while modulating the initial bump amplitude or the fraction of trials that start with an initial bump would also result in different psychophysical kernels (Fig. 3c). Initial bump states generally yield more recency and larger estimation errors (Extended Data Fig. 4e,f). However, our experimental data does not support this hypothesis because the accuracy of human subjects did not depend on the slope of the PK (Fig. 7d).

The global cortical gain can also be modulated in a task-dependent manner across time in individual subjects, for example controlled by top-down signals and neuromodulators such as noradrenalin originating from the locus coeruleus[43–45]. We showed that this can explain attention-driven differences in evidence integration (Extended Data Fig. 7). Moreover, a time-varying neural gain or excitatory drive in the model can switch the network from an integration regime into a working memory regime. During stimulus estimation, the excitatory drive should be moderate so that the network would have a relatively shallow potential and would act as a perfect vector integrator. Subsequently, after stimulus presentation, an increase of $I_0$ would make the potential wider and deeper and thus increase the stability of the bump against noise and against further incoming stimuli as is needed for distractor-resistant working memory[30,46]. This transition could also be gradual, for example caused by a ramping $I_0$[47].

## Mechanisms underlying biases in perceptual estimation and categorization tasks

Starting the integration process in the model with a formed activity bump shifts temporal integration towards recency and can introduce estimation biases towards the initial bump position (Figs. 3 and 5). The existence of bump states before stimulus presentation is supported by neural population recordings[38–40]. A recent paper[5] investigated how prior expectations influence motion direction estimation in humans, and found attractive biases towards the predictive cued direction, and a correlation between reported directions and directions decoded from MEG activity that emerged before stimulus onset. These findings are consistent with the initial bump condition in our model (Fig. 5).

Moreover, it has been shown in combined discrimination and estimation tasks that stimulus estimation is influenced by a categorical decision causing post-decision biases[2,4,20,21]. One hypothesis is that these bias effects are mediated through attention signals and through changes in global gain[4,21]. In the bump attractor model, attention related signals could be modeled either as global changes on the excitatory drive to the neurons or as spatially modulated signals targeting specific regions of the network. For example, in the seminal work of[2] subjects showed a repulsive bias of direction estimates (away from a reference) while performing a fine discrimination-estimation task in which they first had to report whether a pattern of moving dots was clockwise or counterclockwise of a reference. In contrast, subjects showed an attractive bias (towards a reference) in a similar coarse discrimination task in which they had to report whether the dots moved towards or 180° away from the reference. In the model, a spatially modulated attention signal, targeting the location of the two possible choices in the ring, could potentially reproduce the experimentally observed biases. During the fine discrimination task, the attention signal would attract the bump towards the clockwise or counterclockwise directions away from the reference causing a repulsive bias, while the same mechanism would cause the attraction bias during the coarse discrimination task.

Overall, our network model provides a comprehensive computational framework for investigating the neural mechanisms underlying stimulus estimation and perceptual categorization and their interaction in future studies.

## Experimental predictions provided by our model

Our model accounts for experimentally observed PKs (Fig. 6) and we have confirmed the predicted relationship between PKs and reaction times (Fig. 7) in human subjects. Moreover, the bump attractor model yields several predictions that can be tested in human or monkey experiments to validate our theory.

First, our model predicts characteristic changes of the evidence integration dynamics when changing the stimulus properties (e.g. its strength or duration), or its statistics. PKs should shift towards recency for longer stimulus durations (Fig. 3e), which could be tested in a task with randomly interleaved trials of different durations. A similar shift should be observed for stronger, i.e., high contrast stimuli. Additionally, the model predicts changes for manipulations of the generative distribution of the stimuli. For example, for a subject with a particular $I_0$, broader distributions of the stimulus directions would shift the PK towards primacy and also increase the estimation error (Extended Data Fig. 3b,e). Alternatively, subjects could also internally adapt their $I_0$ while keeping their PK constant to minimize the error (Extended Data Fig. 3e). Overall, these predictions are constrained by the fundamental mechanisms that govern the integration dynamics in the bump attractor network which are fully captured by the amplitude equation (Eq. 1).

Second, the dynamics of evidence integration crucially depends on the global excitatory drive or gain of the model (Fig. 3) and this could be tested experimentally using pharmacological or optogenetic manipulations. A recent study has compared stimulus integration in human participants in sessions where they have been administered the NMDA receptor antagonist ketamine or a placebo[48]. Reduced excitability under ketamine led to more recency PKs (more leaky integration), as predicted by a reduction in excitatory drive in our model.

Third, the central prediction of our model is a systematic relationship between the amplitude of the

population response and the rate of change of its phase (Extended Data Fig. 1). This could be validated in simultaneous multi-unit recordings in monkeys performing an orientation averaging task. It would require decoding the current estimate of the average direction during evidence integration and measure the impact of individual stimulus frames on this estimate. The change of phase in response to the same orientation difference (between stimulus frame and current average) should decrease in a manner inversely proportional to the population firing rate. Thanks to recent advances in fine-grained decoding of sensory and decision information across multiply brain areas from MEG activity[49], it may be possible to test the same prediction in humans.

## Comparison with other models

The investigation of neural networks that can integrate information across time has a long history[50]. Neural integrators are usually categorized into either rate based[51,52], accumulating the signal in the graded firing rate of the neurons, or location based, in which the integrated information is encoded in the localized activity of the neurons, for example tracking the head direction in the phase of an activity bump[33,34,53]. The network model presented in this work belongs to the latter category and its activity is governed by bump attractor dynamics as in models of the head direction system. However, the underlying integration process is different. Head-direction models integrate self-motion information that is presented in the form of angular head velocity signals, while visual cues only serve as calibration to partially reset the network. In contrast, in our model the basic operation involves integrating a time-varying sensory stimulus to compute its average direction. Crucially, we have shown that this computation is mediated by the transient interplay between the location of the bump and also its firing rate amplitude, a feature that, to the best of our knowledge, has remained unexplored until now.

One previous study investigated whether a similar continuous recurrent network could identify stimulus features as analog quantities, and also discriminate among a set of discrete alternatives[29]. By changing the underlying dynamical regime, their network was able to perform both computations in two separate tasks. In a direction identification task, they investigated how an orientated stimulus represented in an activity bump is affected by perturbations mimicking cortical microstimulation experiments, similar to the effect of distractors in working memory models[30]. They found that two competing localized bumps can emerge, encoding both the visual stimulus and the microstimulation, and elicit either a stochastic winner-take-all or a spatial averaging process depending on the distance between the two bumps. Overall their integration mechanism is different from the temporal integration studied here, which relies on the dynamics of a single activity bump in response to a time-varying stimulus that achieves the computation of the running average orientation.

Mathematically, the bump attractor model can be viewed as a generalization of the discrete attractor model of perceptual decision making[8,9,11,54]. However, the dynamics of evidence integration in the two models are fundamentally different. Discrete attractor models have a double-well potential that usually leads to a primacy temporal weighting because once the system settles into one of the two attractors it remains there until the end of the trial. Recently we have described how uniform and recency weighting can be obtained when fluctuations in the stimulus together with the internal noise are strong enough to overcome the categorization dynamics and cause transitions between the attractor states[54]. In contrast to this, in the bump attractor model a continuous integration process – without abrupt transitions – occurs in the neutrally stable ring manifold and the stimulus always impacts the bump phase, yielding recency without the need for strong fluctuations (Fig. 3a).

## Low-dimensional ring attractor dynamics in neuronal circuits

Is bump attractor dynamics realized in a cortical circuit a plausible neural mechanism for stimulus estimation? A bump attractor realized in a physical ring layout has recently been found in the fly's heading direction system[36,55,56]. Taking advantage of simultaneous recordings of multiple neurons it has been confirmed that the high-dimensional neural population activity in the mammalian head direction system evolves along a one-dimensional ring attractor manifold[35] consistent with the dynamics of ring

models[32–34]. We hypothesize that a similar topological ring manifold may be embedded in the neural population activity of cortical areas that are involved in stimulus estimation. In order to test experimentally whether bump attractor dynamics underlies orientation averaging, it would be interesting to record neural activity from parietal and frontal areas of non-human primates that perform an estimation task and apply manifold identification approach as in Chaudhuri et al.[35]. Such data would also allow for testing whether trial-to-trial variability of neural firing rates and the estimates reported by the subject are related as predicted by the bump attractor model, as in prefrontal cortical neurons during spatial working memory[31]. In fact, our work suggests that the same neural circuits that are involved in working memory may also be able to carry out stimulus averaging in perceptual estimation tasks. Bump attractor dynamics may thus be a unifying neural mechanism underlying both working memory and evidence integration over prolonged timescales.

# Methods

## Ring model

The dynamics of the bump attractor model are described in terms of the effective firing rate, $r(\theta, t)$, of a neural population arranged in a ring, $\theta \in [-\pi, \pi)$ (Fig. 1a), obeying the integro-differential equation[27,28,57]

$$\tau \frac{\partial r}{\partial t} = -r + \Phi\left(\frac{\tau}{2\pi}\int_{-\pi}^{\pi} w\left(\theta - \theta'\right) r\left(\theta', t\right) d\theta' + I_{\text{exc}} + I_{\text{stim}}\left(\theta, t\right) + \xi\left(\theta, t\right)\right), \tag{4}$$

where $\tau$ is the neural time constant and $\Phi(\cdot)$ is the current-to-rate transfer function. The synaptic input to a neuron with preferred orientation $\theta$ consists of a recurrent current due to the presynaptic activity at a location $\theta'$ with a weight $w(\theta - \theta')$, and external currents $I_{\text{exc}} + I_{\text{stim}}(\theta, t)$ plus fluctuations $\xi(\theta, t)$. The connectivity profile $w(\theta)$, is written in terms of its Fourier coefficients, $w_k$ $(k = 0, 1, 2, \dots)$ and represents the effective excitatory/inhibitory coupling, and can therefore include both positive and negative interactions. The inset in Fig. 1a shows an example of Mexican-hat type connectivity with strong recurrent excitation and broad inhibition. External inputs are divided into a global net excitatory drive $I_{\text{exc}}$ that modulates the excitability of the network and a time-varying input $I_{\text{stim}}$ that represents the sensory stimulus. In general, the sensory stimulus can be written in terms of its Fourier coefficients $I_k$, $k = 1, 2, \dots$ with directions $\theta_k(t)$. Throughout this work we will focus on an input of the form $I_{\text{stim}}(\theta, t) = I_1 \cos\left(\theta - \theta^{\text{stim}}(t)\right)$ to model a stimulus with constant strength $I_1$ and a time-varying orientation $\theta^{\text{stim}}(t)$ but our derivations are equally valid for time-varying stimulus strengths $I_1$. A detailed description of the stimuli used in the simulations is given below (Model simulations). Finally, the fluctuation term $\xi(\theta, t)$ reflects the joint effect of the internal stochasticity of the network and temporal variations in the stimulus. For simplicity, we model these fluctuations as independent Ornstein-Uhlenbeck processes for each neuron, with amplitude $\sigma_{\text{OU}} = 0.15$ and time constant $\tau_{\text{OU}} = 1$ ms. Tab. 1 summarizes the values of the model parameters used in this work.

The effective network, Eq. 4, assumes that excitatory and inhibitory neurons follow the same dynamics, and thus can be grouped together, greatly simplifying the mathematical analysis. Nevertheless, a mathematical analysis of the more general model with separated pools of excitatory and inhibitory neurons is possible and is expected to give the same results[58–61].

## The bump attractor

In the following we describe the conditions for which the ring network shows a stable bump of activity. We consider the quadratic/square root transfer function[34,62]

$$\Phi(x) = \frac{1}{\tau} \cdot \begin{cases} 0 & , \quad x < 0 \\ x^2 & , \quad 0 \le x \le 1 \\ 2\sqrt{x - 3/4}, & \quad x > 1 \end{cases} \tag{5}$$

where $x$ is the total synaptic input arriving to the neurons. Other continuous non-linear transfer functions will lead to the same qualitative behavior provided they are a monotonically increasing function of the input.

In the absence of sensory stimuli, $I_{\text{stim}} = 0$, the ring network can show different spatial activity profiles depending on the form of the connectivity profile and the global excitatory drive $I_{\text{exc}}$. For $I_{\text{exc}} \leq 0$ the network has a unique equilibrium solution with zero activity, $r(\theta) = 0$ Hz. Increasing the excitatory drive, $I_{\text{exc}} > 0$, will increase the activity of this equilibrium solution uniformly along the ring $r(\theta) = r_\infty = \Phi(\tau w_0 r_\infty + I_{\text{exc}})/\tau > 0$. By studying the linear stability of this solution we can find a critical value of the global excitatory drive, $I_{\text{exc}} = I_{\text{crit}}$, for which the spatially homogeneous state destabilizes into a spatially inhomogeneous state via a pattern forming instability known as Turing bifurcation. This new stable state will have a single bump of activity as shown in Fig. 1b for a connectivity profile that satisfies $w_1 > w_k$ for all $k \neq 1$. For our choice of the transfer function (Eq. 5) and the connectivity profile (see Tab. 1) this critical value is given by

$$w_1^{\text{crit}} = \frac{1}{\tau \Phi'_\infty} \Rightarrow I_{\text{crit}} = \frac{1}{4w_0}\left[1 - \left(\frac{w_0 - w_1}{w_1}\right)^2\right] = 1, \tag{6}$$

where $\Phi'_\infty = \Phi'(\tau r_\infty w_0 + I_{\text{crit}})$ is the slope of the transfer function evaluated at the bifurcation point.

**Table 1 |** List of the parameters of the ring network and the stimulus.

| **Ring network (Eq. 4)** | *Symbol* | Value |
|---|---|---|
| Time constant of the network | $\tau$ | 20 ms |
| Fourier modes of the connectivity | $w_k$, $k = 0, 1, 2$ | $-2, 1, 0.5$ |
| Global net excitatory drive | $I_{\text{exc}}$ | $\in [1.01, 1.2]$ |
| Noise term | $\xi(\theta, t)$ | Eq. 4 |
| Noise amplitude | $\sigma_{\text{OU}}$ | 0.15 |
| Noise time constant | $\tau_{\text{OU}}$ | 1 ms |
| Transfer function | $\Phi(\cdot)$ | Eq. 5 |
| Critical excitatory drive | $I_{\text{crit}}$ | 1 |
| Stationary FR at $I_{\text{exc}} = I_{\text{crit}}$ | $r_\infty$ | 12.5 Hz |
| **Stimulus (Eqs. 27 and 28)** | *Symbol* | Value |
| Stimulus amplitude | $I_1$ | $5 \cdot 10^{-3}$ |
| Duration of the stimulus | $T$ | 2 s |
| Number of stimulus frames | $N$ | 8 |

## Amplitude equation

The bump solution described in the previous section is analytically intractable. However, we can study its dynamics near the onset of the bump, i.e. in the vicinity of the bifurcation point $I_{\text{exc}} = I_{\text{crit}}$, by performing a multiple-scale analysis based on a perturbation method[63]. We chose $I_{\text{exc}}$ as our bifurcation parameter and define the perturbation

$$I_{\text{exc}}(\theta) = I_{\text{crit}} + \varepsilon^2 I_0 + \varepsilon^3 2 I_1 \cos\left(\theta - \theta^{\text{stim}}\right), \tag{7}$$

and the corresponding expansion of the firing rate

$$r = r_\infty + \sum_{n=1}^{\infty} \varepsilon^n r_n. \tag{8}$$

The multiple scale method assumes that the dynamics of the network can be separated into two independent time scales, such that we can define the slow time $T = \varepsilon^2 t$ in which the dynamics of the bump evolve. Plugging all these definitions into the equation of the ring network, Eq. 4, we obtain the equation that describes the evolution of the amplitude of the bump, $A(t)$, close to the bifurcation point:

$$\tau \frac{\partial A}{\partial T} = \tilde{I}_1 e^{-i\theta^{\text{stim}}} + \tilde{I}_0 A - c\,|A|^2 A + \zeta(t). \tag{9}$$

where $\tilde{I}_1 = \Phi'_\infty I_1$ is the scaled stimulus strength, $\tilde{I}_0$ corresponds to the scaled distance to the bifurcation point, which is related to the global excitatory drive:

$$\tilde{I}_0 = \frac{\Phi''_\infty}{\Phi'_\infty\left(1 - \tau w_0 \Phi'_\infty\right)} \left(I_0 - 2\sigma^2\right), \tag{10}$$

and the cubic factor $c$ is

$$c = -\left[\frac{\left(\Phi''_\infty\right)^2}{\left(\Phi'_\infty\right)^3}\left(\frac{\tau w_0}{1 - \tau w_0 \Phi'_\infty} + \frac{\tau w_2}{2\left(1 - \tau w_2 \Phi'_\infty\right)}\right) + \frac{\tau^2 \Phi'''_\infty}{2\left(\Phi'_\infty\right)^3}\right]. \tag{11}$$

The last term in eq. 9 corresponds to the first Fourier component of the noise process $\xi(\theta, t)$ in Eq. 4, that can be written as

$$\zeta(t) = \frac{\Phi'_\infty}{\sqrt{2n}}\left(\xi_a(t) + i\xi_b(t)\right) = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\; e^{-i\cdot \text{atan2}(\xi_b, \xi_a)}, \tag{12}$$

where, $n$ is the discrete number of Fourier modes of the noise process $\xi(\theta, t)$, and $\xi_a(t)$ and $\xi_b(t)$ are independent Ornstein-Uhlenbeck processes with the same amplitude, $\sigma_{\text{OU}}$, and time constant, $\tau_{\text{OU}}$ as $\xi(\theta, t)$. Additionally, the non-linear filtering of the noise process, $\xi(\theta, t)$, through the transfer function, introduces spatial correlations[64,65] that effectively shift the bifurcation point to $I_{\text{crit},\sigma} = I_{\text{crit}} + 2\sigma_{\text{OU}}^2$, which explains the correction term in Eq. 10. Thus, changing the noise level not only increases the variability of responses (Fig. 4c), but also affects the integration of the stimulus by modulating the amplitude of the bump in a similar way as the global excitatory drive, $I_0$.

Finally, by considering the polar form of $A(t) = Re^{-i\psi}$, Eq. 9 results in the system described by Eqs. 1a and 1b with the noise terms $\xi_1$ and $\xi_2$ being

$$\xi_1 = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\cos\left(\psi - \text{atan2}\left(\xi_b, \xi_a\right)\right),\; \xi_2 = \frac{\Phi'_\infty}{\sqrt{2n}}\sqrt{\xi_a^2 + \xi_b^2}\sin\left(\psi - \text{atan2}\left(\xi_b, \xi_a\right)\right). \tag{13}$$

The values of the parameters used for the simulation of the amplitude equation are listed in Tab. 2.

The amplitude equation (Eq. 9) derived here corresponds to the normal form of a Turing bifurcation where the first spatial mode of the homogeneous state becomes unstable. Similar derivations can be found in the literature (see for example, refs.[57,58] and ref.[27] section 8.1.2), which also study the general case of separated excitatory-inhibitory populations. This low-dimensional description of the network accurately describes the bump dynamics close to the bifurcation point where the bump emerges, and it is qualitatively valid for a wide parameter regime (Fig. 3b), allowing us to study the dynamics of the bump analytically.

**Table 2 |** List of the parameters of the amplitude equation, and parameters of the numerical integration, used to simulate both Eqs. 1 and 4.

| **Amplitude equation (Eq. 1)** | *Symbol* | Value |
|---|---|---|
| Additional excitatory drive, $I_{\text{exc}} - I_{\text{crit}}$ | $I_0$ | $\in [0.01, 0.2]$ |
| Slope of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$ | $\Phi'_\infty$ | 50 |
| 2nd derivative of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$ | $\Phi''_\infty$ | 100 |
| 3rd derivative of $\Phi(\cdot)$ at $I_{\text{exc}} = I_{\text{crit}}$ | $\Phi'''_\infty$ | 0 |
| 1st order prefactor | $\tilde{I}_0$ | $\in [-7/300, 31/300]$ |
| Cubic prefactor | $c$ | $1/3750$ |
| Scaled stimulus amplitude | $\tilde{I}_1$ | 0.125 |
| Noise terms | $\xi_i(t)$, i=1,2 | Eq. 13 |
| **Numerical integration** | *Symbol* | Value |
| Integration time step | $\Delta t$ | 0.2 ms |
| Spatial discretization | $n$ | 200 |
| Number of trials | $K$ | $10^5$ |

## Dynamics of the phase for a stationary bump amplitude

An analytical expression for the approximate evolution of the phase, $\psi(t)$, can be obtain by assuming an almost static bump amplitude ($\dot{R} \sim 0$), and then solving the ordinary differential equation Eq. 1b,

$$\psi(t) = \theta^{\text{stim}} - 2\arctan\left[\tan\left(\frac{\theta^{\text{stim}} - \psi_0}{2}\right) e^{-\frac{\tilde{I}_1}{\tau R_\infty} t}\right], \quad \psi(0) = \psi_0, \quad \psi_\infty = \theta^{\text{stim}}, \tag{14}$$

where $\psi_0$ is the initial phase of the bump, and the stationary amplitude, $R_\infty$, is obtained by solving the third order polynomial that results from equating Eq. 1a to zero. The solution to this polynomial equation depends on the value of $\tilde{I}_1$, which if set to 0, gives $R_\infty = \sqrt{\frac{\tilde{I}_0}{c}}$. Extended Data Fig. 2a shows the evolution of the phase of the bump for an initial condition in which the bump is completely formed and centered at 0° for a stimulus $\theta^{\text{stim}} = 90°$. The approximation (shaded lines), Eq. 14, is plotted along with the corresponding simulation of the amplitude equation showing good agreement. As expected, the approximation works better for larger values of $\tilde{I}_0$, that make the amplitude of the bump less sensitive to the stimulus.

A first order approximation of Eq. 14, as a function of $\theta^{\text{stim}}$ (Extended Data Fig. 2c), is given by

$$\Delta\psi\left(\theta^{\text{stim}}\right) \approx \left(1 - e^{-\frac{\tilde{I}_1}{\tau R_\infty} t}\right) \Delta\theta^{\text{stim}}. \tag{15}$$

The error incurred by this approximation is small (approx. 10% for $\Delta\theta^{\text{stim}} = 90°$, $t = 250$ ms and $\tilde{I}_0 = 0.02$), confirming that stimulus integration in the bump attractor model is relatively linear across a wide range of $\Delta\theta^{\text{stim}}$.

## Potential of the amplitude equation

Here we show that the system $\dot{\mathbf{x}} = \mathbf{f}(R, \psi)$ describing the evolution of the bump (Eq. 1), is indeed a gradient system, $\dot{\mathbf{x}} = -\nabla\Theta(R, \psi)$, with a compatible potential function $\Theta(R, \psi)$ given by Eq. 2:

$$\begin{aligned}
\nabla\Theta &= \left(\frac{\partial}{\partial R}, \frac{1}{R}\frac{\partial}{\partial \psi}\right)\Theta(R, \psi), \\
&= -\left(\tilde{I}_1 \cos\left(\psi - \theta^{\text{stim}}(t)\right) + \tilde{I}_0 R - cR^3,\ -\tilde{I}_1 \sin\left(\psi - \theta^{\text{stim}}(t)\right)\right) \\
&= -\left(\frac{dR}{dt}, R\frac{d\psi}{dt}\right) = -\mathbf{f}(R, \psi)
\end{aligned} \tag{16}$$

where we have used the polar form of the gradient operator and the differential of $\mathbf{x} = (R, \psi)$ is $(dR, Rd\psi)$. Similarly, we can compute its time-evolution

$$\begin{aligned}
\frac{d\Theta}{dt} &= \frac{\partial\Theta}{\partial R}\frac{dR}{dt} + \frac{\partial\Theta}{\partial\psi}\frac{d\psi}{dt} \\
&= -\left(\frac{\partial\Theta}{\partial R}\right)^2 - \left(\frac{1}{R}\frac{\partial\Theta}{\partial\psi}\right)^2 \\
&= -(\nabla\Theta)^2 \leq 0,
\end{aligned} \tag{17}$$

that shows that for all initial conditions $\mathbf{x}_0 = (R_0, \psi_0)$, the potential will always decrease towards a stationary stable state $(R_\infty, \psi_\infty)$.

## Dynamical system for the cumulative circular average (perfect vector integrator)

In order to understand the integration process carried out by the ring attractor model, we derived a dynamical system that perfectly integrates a circular variable to compute the circular average. In general, the optimal strategy to keep track of the average, $\bar{z}(t)$, of a time-varying stimulus, $z(t)$, is to compute the cumulative running average. For discrete-time stimuli the cumulative running average can be computed by the following iterative equation:

$$\bar{z}_t = \bar{z}_{t-1} + \frac{1}{t}\left(z_t - \bar{z}_{t-1}\right),\ t = 1, 2, \ldots \tag{18}$$

where $t$ represents discrete time points. If we assume that the updates $\bar{z}_t - \bar{z}_{t-1}$ are small enough, we can write the above equation as a continuous process in time:

$$\frac{d\bar{z}}{dt} = \frac{1}{t}\left(z(t) - \bar{z}\right),\ t \geq 1; \quad \bar{z}(1) = z(1). \tag{19}$$

In order to avoid dealing with a non-autonomous system we can rewrite this equation by absorbing the time dependence into a new variable $\alpha(t) = t\bar{z}(t)$ which leads to the integral form of the perfect integrator:

$$\frac{d\alpha}{dt} = z(t). \tag{20}$$

According to this equation, $\alpha(t)$ is just the sum of all $z(t)$ (integral over time), and the average is obtained by dividing the summation by $t$:

$$\bar{z}(t) = \frac{1}{t}\int_{t_0}^{t} z(t')dt'. \tag{21}$$

However, even if the result of the above integral is equivalent to that obtained by means of the iterative method, Eq. 19, the computational approach is fundamentally different. Here we would be required to memorize the entire time-varying stimulus up to the instant $t$, and divide its summation by the duration of the stimulus. In contrast, using the iterative method, Eq. 19, only requires keeping track of the average

quantity at the last time point and update that value by weighting successive evidence in a time-dependent manner.

Eq. 20 holds for any time-varying process $z(t)$, in particular, here we are interested in $z(t) \in \mathbb{C}$. In this case we can interpret $z(t)$ as stimulus vectors defined by the stimulus strength and the stimulus direction and $\bar{z}(t)$ as the vector average up to time $t$. The resulting complex quantities can be written in polar form as $\alpha(t) = R(t)e^{i\psi(t)}$ and $z(t) = I(t)e^{i\theta(t)}$ which gives the two equations

$$\frac{dR}{dt} = I(t)\cos\left(\psi - \theta(t)\right), \tag{22a}$$

$$\frac{d\psi}{dt} = -\frac{I(t)}{R}\sin\left(\psi - \theta(t)\right). \tag{22b}$$

We refer to these equations as the *perfect vector integrator (PVI)* (Extended Data Fig. 3a,c). As described previously in this work, our main goal is to compute the circular average of the stimulus orientations, i.e. the argument of $\bar{z}(t)$. From the definition of $\alpha(t) = t\bar{z}(t)$, we can already see that $\arg(\bar{z}) = \arg(\alpha) = \psi$. Therefore Eq. 22b exactly describes the evolution of the circular average of the orientations of the stimulus frames, i.e. the PVI computes the cumulative circular average. Moreover, note that Eq. 22b will compute the circular cumulative running average for any given normalization function $f(t)$ such that $\alpha(t) = f(t)\bar{z}(t)$.

The variable $R$, needed in the computation of the circular average, has several interpretations. First, the average magnitude of $\bar{z}(t)$, can be computed as $|\bar{z}(t)| = \frac{1}{t}R(t)$. In other words, $R(t)$ is the magnitude of the integrated vectors (i.e., the length of the resulting vector). Second, from this geometric interpretation it is easy to see that $R(t)$ measures the dispersion of the directions $\theta$, with $R(t)$ growing linearly with $t$ if $\theta(t)$ is a constant. If $\theta(t)$ is sampled from some underlying distribution, the growth will be slower and depends on the width of the distribution (Extended Data Fig. 3b,d). The circular variance of the samples can be computed from the ratio of $R(t)$ and the maximum $R$ obtained for $\theta(t) = \text{const.}$ as $\text{var} = 1 - \frac{R(t)}{\int I(t)dt}$. Third, if we assume that $\theta(t)$ is sampled from a von Mises distribution with mean $\mu$ and concentration $\kappa$, the running circular mean $\psi(t)$ will also be distributed according to a von Mises distribution, with concentration proportional to $R(t)\kappa$. In a Bayesian framework, $\psi(t)$ and $R(t)$ thus track the mean and the concentration of the posterior distribution[66].

### Potential landscape for the perfect vector integrator

We follow the same approach used in the main text to graphically picture the dynamics of the perfect vector integrator (Eqs. 22a and 22b) by defining its potential landscape,

$$\Theta^{\text{PVI}}(R, \psi) = -RI\cos\left(\psi - \theta^{\text{stim}}\right). \tag{23}$$

Proceeding in the same way as with the amplitude equation (Eq. 16) it is easy to see that the gradient of this potential gives the system Eq. 22.

## Categorical readout: decision circuit

In order to obtain a categorical decision from the continuous estimate of the ring network (Fig. 7), we use a standard two-population discrete attractor network[8,9,11] that receives excitatory inputs from neurons located at each side of the ring (Fig. 7a). The decision circuit is composed of two recurrently coupled excitatory populations selective to left and right inputs, respectively, and an inhibitory population that is coupled to both excitatory populations. The activity of the network is described in terms of the firing rate

of its populations, $r_L$, $r_R$ and $r_I$, and obeys[9]:

$$\tau_E \frac{dr_L}{dt} = -r_L + \Phi\left(sr_L - cr_I + I_L + I_E(t)\right), \tag{24a}$$

$$\tau_E \frac{dr_R}{dt} = -r_R + \Phi\left(sr_R - cr_I + I_R + I_E(t)\right), \tag{24b}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \Phi\left(g\left(r_L + r_R\right) + I_I\right), \tag{24c}$$

where $\tau_{E(I)}$ represents the membrane time constant of the excitatory (inhibitory) neurons and $\Phi(x) = \frac{1.5}{1+e^{-2.5(x-1)}}$ is the current-to-rate transfer function[9]. The inhibitory population receives excitatory inputs from both excitatory populations with synaptic strength $g$, and a constant positive current $I_I$. The excitatory populations are recurrently coupled with synaptic strength $s$ and receive inhibition from the inhibitory population with synaptic strength $c$. The values of all model parameters are summarized in Tab. 3. Additionally the L and R populations receive excitatory inputs from the integration circuit, $I_L$ and $I_R$, that will depend on the size and location of the bump, and are computed as the convolution of the activity of the bump and a spatially constant kernel targeting its corresponding side of the ring:

$$I_i(t) = \frac{\gamma}{2\pi} \int_{-\pi}^{\pi} r(\theta, t)\, H_i(\theta)\, d\theta, \quad H_i(\theta) = H(\pm\theta + \pi) \cdot H(\mp\theta); \quad i = L, R \tag{25}$$

where $\gamma = 0.0015$, and $H_i(\theta)$ is the left or right Heaviside step function.

Finally, $I_E$ represents a time varying external input that determines the dynamical regime of the decision network. During the integration period, $I_E = I_R$ is set such that the activities of both populations remain at the low symmetrical state with bistability between the symmetric and the winner-take-all states. To force the decision, $I_E$ is increased by activating a transient urgency signal, $I_U$ (with duration $\Delta T_U$), that breaks the stability of the low-activity symmetric state and triggers the winner-take-all competition between the excitatory populations:

$$I_U(t) = I_R + H(t - T_{\text{stim}})\, H(T_{\text{stim}} + \Delta T_U - t), \tag{26}$$

where $H(t)$ is the Heaviside step function and $T_{\text{stim}}$ is the stimulus duration. The decision is made whenever the activity of one of the excitatory populations crosses the 50 Hz threshold, $r_{th}$ (Fig. 7b,c).

**Table 3 |** List of the parameters of the decision circuit and the urgency signal.

| **Decision circuit (Eq. 24)** | *Symbol* | Value |
|---|---|---|
| Time constant of the excitatory/inhibitory neurons | $\tau_E, \tau_I$ | 20 ms |
| Strength of the recurrent coupling | $s$ | 1.9 |
| Strength of the inhibitory cross coupling | $c$ | 1.8 |
| Strength of the excitatory cross coupling | $g$ | 1 |
| External input to inhibitory neurons | $I_I$ | 0.2 |
| **External input,** $I_E(t)$ | *Symbol* | Value |
| Resting input | $I_R$ | 0.33 |
| Resting input + Urgency signal | $I_R + I_U$ | 0.46 |
| Duration | $\Delta T_U$ | 500 ms |

## Modeling focused and divided attention

To model evidence integration in the focused vs. divided attention conditions (Extended Data Fig. 7) we built an extended network model consisting of two identical ring networks, each integrating one of the

two spatially separated orientation streams (Extended Data Fig. 7d). The dynamics of both circuits are described by Eq. 4, using the parameters shown in Tab. 1. Focused and divided attention conditions are modeled by including a modulatory attention signal, in the form of an external input, that distributes a fixed amount of resources between the two circuits. For simplicity, the attention signal is included in the global excitatory drive, $I_{\text{exc}} = I_{\text{crit}} + I_0$. In the focused condition ($I_0^{\text{target}} = 0.035$; $I_0^{\text{distractor}} \sim 0$) the resources are mainly directed towards the circuit encoding the target stream, while in the divided attention condition ($I_0^{\text{neutral}} = 0.01$) they are equally divided between both neutral circuits. Note that the amplitude of the bump goes as $R \propto \sqrt{I_0}$, and therefore the overall activity in both conditions remains similar. The attention signal is the only difference between focused and divided attention trials. In particular, the stimulus strength is the same for cued and neutral stimulus streams. This is justified by the empirical finding that occipital and parietal EEG signals recorded from the participants represented neutral and target stimuli (in divided and focused attention, respectively) with a similar strength[41]. EEG signals also revealed that the information about the distractor stimulus is (partly) filtered out at early processing stages. Thus, in the model, the distractor stream in the focused attention condition has a 50% reduced stimulus strength and it is unattended ($I_0^{\text{distractor}} \sim 0$). As a result, the distractor stream is not integrated, as has been observed in humans.

## Model simulations

### Stimuli

The stimulus in each trial consists of a sequence of $N$ i.i.d. oriented frames $\theta_j^{\text{stim}}, j = 1, \ldots, N$ that are sampled from some probability distribution. We define the stimulus orientation as a piecewise constant function of time

$$\theta^{\text{stim}}(t) = \sum_{j=1}^{N} \theta_j^{\text{stim}} H\left(t - \frac{j-1}{N} T_{\text{stim}}\right) H\left(\frac{j}{N} T_{\text{stim}} - t\right), \tag{27}$$

where $H(t)$ is the Heaviside step function and $T_{\text{stim}}$ is the stimulus duration. The input current to each neuron on the ring was modeled as

$$I_{\text{stim}}(\theta, t) = I_1(t) \cos\left(\theta - \theta^{\text{stim}}(t)\right), \tag{28}$$

where $I_1$ is the stimulus strength and $\theta$ is the position of the neuron on the ring (Eq. 4). In most of the simulations we run $K = 10^5$ trials each consisting of $N = 8$ stimulus frames with a constant strength $I_1$ and stimulus orientations sampled from a uniform distribution in the range $\theta_j^{\text{stim}} \in [-90°, 90°)$, but randomly selecting combinations $\left\{\theta_1^{\text{stim}}, \ldots, \theta_N^{\text{stim}}\right\}$ such that their circular mean, $\bar{\theta}^{\text{stim}}$, is uniformly distributed in the range $\bar{\theta}^{\text{stim}} \in [-80°, 80°]$.

### Initial conditions

To simulate the **No initial bump** condition (Figs. 1c-e, 2b, 3b-f, 4 and 5a) the activity of the network at the beginning of each trial is set at a spatially homogeneous state ($I_{\text{exc}} = I_{\text{crit}}$) with an average activity around $r_\infty = 12.5$ Hz. The bump state is formed by increasing the global excitatory drive to the network, $I_{\text{exc}}$, such that the bump starts to emerge at the beginning of the stimulus presentation.

On the other hand, for the **Initial bump** condition (Figs. 3a,c, 5b and 4b) the activity of the network at the beginning of each trial is set at a stationary bump. We compute this state taking into account both the strength of the stimulus, $I_1$, and the effect of the noise (see the derivation of the Amplitude equation), by waiting long enough such that the system reaches the steady state. For the *biased* condition (Figs. 3a,c, and 5b), the bump is centered at 0°, while for the *unbiased* condition (Extended Data Fig. 4b), the initial phase of the bump is shifted to match the orientation of the first stimulus frame. This shift is taken to be equal to the position of the bump in the *No initial bump* condition at the end of the first stimulus frame, which still captures the impact of fluctuations at the beginning of the trial.

### Numerical integration

We simulated Eqs. 4 and 9 following the Euler-Maruyama scheme with a time step $\Delta t = 0.5$ ms, and discretizing the angular space $\theta$ into $n = 200$ evenly distributed locations, $\theta_k = k \cdot 360°/n - 180°$. All simulations and analysis were done in Python 3.7.3, using standard libraries including among others: `NumPy`, `ScyPy` and `Pandas`.

## Statistical measures and fits

### Estimation bias and estimation error

To characterize the accuracy with which an estimate $\bar{\theta}_{\text{est}}$ describes the true stimulus average $\bar{\theta}$, we define the following quantities.

*Bias of the estimator*. Usually, the estimation bias is defined as $b_{\text{est}}\left(\bar{\theta}\right) = \langle\bar{\theta}_{\text{est}}\rangle - \bar{\theta}$, i.e., the difference of the trial-average of the estimates $\bar{\theta}_{\text{est}}$ and the true value $\bar{\theta}$. For our simulations (Fig. 5c), the bias defined in this way would be zero for $\bar{\theta} = 0°$ and increase approximately linearly with $\bar{\theta}$. Thus, we preferred to define the estimation bias as one minus the slope of the estimation curve obtained from a regression fit ($k_1$ in Eq. 32). An unbiased estimator would have a slope of 1 and thus zero bias ($1 - k_1 = 0$), while slopes smaller than 1 correspond to increasingly large values of the estimation bias.

*Variance of the estimator*. The variance of the estimator is defined as $\sigma^2_{\text{est}}\left(\bar{\theta}\right) = \langle\left(\bar{\theta}_{\text{est}} - \langle\bar{\theta}_{\text{est}}\rangle\right)^2\rangle$, i.e. the variation of the estimates about their mean value. In Fig. 5e, we plot the square root of this quantity, the standard deviation of the estimates.

*Estimation error*. To assess the total deviations of estimated average orientations $\bar{\theta}_{\text{est}}$ from the true stimulus average $\bar{\theta}$, due to estimation bias and variance, we compute the root-mean-square error

$$\text{RMSE}\left(\bar{\theta}\right) = \sqrt{\langle\left(\bar{\theta}_{\text{est}} - \bar{\theta}\right)^2\rangle}, \tag{29}$$

i.e., the square root of the trial-average squared estimation error. For Figs. 4c and 5e, we computed the mean of the RMSE across all average orientations $\bar{\theta}$.

*Weighted circular average*. To check whether other factors, besides the intrinsic nonlinearities of the model, are contributing to the estimation error, we compute the weighted circular average by simulating the PVI (Eq. 22), with the same exact stimulus $I^{\text{stim}}(t)$ used to simulate the ring network (Eqs. 27 and 28) but with a time-varying amplitude obeying

$$I_1\left(t\right) = I_1 \sum_{j=1}^{N} \beta_j H\left(t - \frac{j-1}{N} T_{\text{stim}}\right) H\left(\frac{j}{N} T_{\text{stim}} - t\right), \tag{30}$$

where $\beta_j$ correspond to the weights of the PK obtained for the ring network, and $H(t)$ and $T_{\text{stim}}$ are the Heaviside step function and the stimulus duration, respectively. This measure corresponds to the purple line in Fig. 4c.

### Regression models for circular data

We use regression models to assess the impact of each stimulus frame on the final estimate of the stimulus average provided by the bump attractor model. Standard linear regression is not well suited for this purpose because we are estimating the average of a circular quantity. We thus derived a regression model for circular data. We define the weighted circular mean as the argument of a weighted sum of complex numbers:

$$\bar{\alpha} = \phi\left(\boldsymbol{\theta}; \boldsymbol{\beta}\right) = \arg \sum_{j=1}^{N} \beta_j \exp\left(i \cdot \theta_j\right), \tag{31}$$

where $N$ in the number of stimulus frames, $\theta_j$ the orientation of the $j$th frame and $\beta_j$ the corresponding weight. In order to account for systematic deviations of estimated average orientations, we include a

bias term $k_0$ and a scale factor $k_1$ (determining the slope of the estimation curve) and model the average stimulus estimation in trial $i$ as:

$$\mu_i = k_0 + k_1 \phi\left(\boldsymbol{\theta_i}; \boldsymbol{\beta}\right). \tag{32}$$

Assuming that the data follows a von Mises distribution with concentration parameter $\kappa$, we can derive the log-likelihood of observing the data $\boldsymbol{\psi}$ (i.e., the reported average angles) given the weights $\boldsymbol{\beta}$ as

$$LL\left(\boldsymbol{\beta}, k_0, k_1, \kappa; \boldsymbol{\psi}\right) = \sum_{i=1}^{K} \kappa \cos\left(\psi_i - \mu_i\right) - K \log\left(2\pi I_0\left(\kappa\right)\right), \tag{33}$$

where $K$ is the number of trials and $I_0(\kappa)$ is the modified Bessel function of order 0. To uniquely define the weights $\boldsymbol{\beta}$ we include the constraint that the sum of $\sum \beta_j = 1$. The parameters $(\boldsymbol{\beta}, k_0, k_1, \kappa)$ that maximize the log-likelihood were estimated using the `minimize` function from `scipy.optimize` module in Python 3.7.3.

### Psychophysical kernel

For the model simulations, we obtain the PK as the stimulus weights of a circular regression model. To quantify the shape of a PK, we define the PK slope[54]. It is the slope of a linear regression of the PK, normalized between $-1$ (decaying PK, primacy) and $+1$ (increasing PK, recency). Formally, we first fit the PK with a linear function of time,

$$\mathrm{PK}(t) = \beta_0 + k\beta_1 t, \tag{34}$$

where $\beta_1$ is the PK slope and $k = \frac{1}{2\mathrm{Var}(t)}$ is a factor that normalizes the PK slope to the interval $(-1, 1)$.

## Psychophysical data and data analysis

### Psychophysical experiment and datasets

We used published data from three studies carrying out psychophysical experiments in which human subjects had to categorize the average direction of a stream of Gabor patterns[6,7,41] (Fig. 6a). To deal with slight differences in the experimental procedures (see below), we organized the data in 5 datasets: dataset 1 containing data from 15 subjects of Wyart et al.[6], dataset 2 with data from 10 subjects of experiment 1 from Cheadle et al.[7], dataset 3 with data from 19 subjects of experiment 2 from Cheadle et al.[7], dataset 4 with data from 17 subjects of the focused attention condition from Wyart et al.[41], and dataset 5 with data from the same 17 subjects in the divided attention condition Wyart et al.[41]. The stimuli in all experiments consisted of a stream of eight oriented high-contrast (50%) Gabor patches with orientations $\theta_j^{\mathrm{stim}},\ j = 1, 2, \ldots, 8$ uniformly distributed in the range between $-90°$ and $90°$. The duration of each sample was 250 ms in datasets 1-3 and 333 ms in dataset 4-5. In each trial participants reported whether, on average, the orientation of the eight samples fell closer to the cardinal or diagonal axes. The category level of each trial was computed by means of a folded decision-mapping rule consisting of a w-shaped profile that mapped the orientation of each sample to a decision variable ranging from $-1$ (diagonal) to 1 (cardinal), and then averaged across samples. Participants responded by pressing, with their left or right index finger, either two keys of a standard keyboard (dataset 1) or a one of two response buttons (DirectIN high-speed button box, Empirisoft; datasets 3-5). In the experiments of dataset 2 they responded at a continuous scale by clicking the mouse to a position on the screen corresponding to the integrated decision value. Auditory feedback was given at the end of each trial depending on the agreement between the response and the sign of the average decision value. The average number of trials performed by each participant was 546, 387, 433, 172 and 182 in datasets 1-5, respectively.

### Analysis of psychophysical kernels for behavioral data

In this analysis (Fig. 6b) we included data from 61 subjects (datasets 1-4). Dataset 5 was not included because of the different experimental condition (divided attention that leads to more recency[41]). PKs of each participant are obtained as the weights of logistic regression (Extended Data Fig. 6). To characterize the shape of individual PKs, we used logistic regression with the constraint that the PK is either uniform (constant) or a linear function of time (Extended Data Fig. 6). We compare the model fits using AIC and classify each PK as uniform (best model has a constant PK), primacy (linear model with negative slope) or recency (linear model with positive slope).

### Analysis of reaction times

In this analysis (Fig. 7) we included data from 47 subjects from datasets 1, 3 and 4. Dataset 2 was not included because subjects had to move and click the mouse which is not comparable to pressing a key or a button as in the other experiments. In addition, 4 subjects from dataset 3 were excluded because they had reaction times shorter than 240 ms in some trials and those reaction times had not been recorded. To compute the dependence of reaction times on the PK slope (Fig. 7f), we normalized the mean reaction times of each subject across the different datasets by z-scoring each dataset individually. This removes the overall differences in reaction times between experiments with a standard keyboard (dataset 1) and response buttons (dataset 3 and 4).

### Analysis of evidence integration under focused and divided attention

For this analysis (Extended Data Fig. 7), we included data from the 17 subjects of dataset 4 and 5.

## Code availability

The code to generate the figures of the paper will be uploaded in `github`.

## References


1. Britten, K., Shadlen, M., Newsome, W. & Movshon, J. The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).

2. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).

3. Nichols, M. J. & Newsome, W. T. Middle temporal visual area microstimulation influences veridical judgments of motion direction. *J. Neurosci.* **22**, 9530–9540 (2002).

4. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation bias through selective overweighting of choice-consistent evidence. *Curr. Biol.* **28**, 3128–3135.e8 (2018).

5. Aitken, F., Turner, G. & Kok, P. Prior expectations of motion direction modulate early sensory processing. *J. Neurosci.* **40**, 6389–6397 (2020).

6. Wyart, V., de Gardelle, V., Scholl, J. & Summerfield, C. Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron* **76**, 847–858 (2012).

7. Cheadle, S. et al. Adaptive gain control during human perceptual choice. *Neuron* **81**, 1429–1441 (2014).

8. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).

9. Roxin, A. & Ledberg, A. Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Comput. Biol.* **4**, e1000046 (2008).

10. Wimmer, K. et al. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat. Commun.* **6**, 1–13 (2015).

11. Wong, K.-F. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).

12. Ratcliff, R. & McKoon, G. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).

13. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).

14. Pardo-Vazquez, J. L. et al. The mechanistic foundation of Weber's law. *Nat. Neurosci.* **22**, 1493–1502 (2019).

15. Kiani, R., Hanks, T. & Shadlen, M. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *J. Neurosci.* **28**, 3017–3029 (2008).

16. Mazurek, M. E., Roitman, J. D., Ditterich, J. & Shadlen, M. N. A role for neural integrators in perceptual decision making. *Cereb. Cortex* **13**, 1257–1269 (2003).

17. Zhang, J., Bogacz, R. & Holmes, P. A comparison of bounded diffusion models for choice in time controlled tasks. *J. Math. Psychol.* **53**, 231–241 (2009).

18. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).

19. Ratcliff, R. Decision making on spatially continuous scales. *Psychol. Rev.* **125**, 888–935 (2018).

20. Luu, L. & Stocker, A. A. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife* **7** (2018).

21. Talluri, B. C. et al. Choices change the temporal weighting of decision evidence. *J. Neurophysiol.* **125**, 1468–1481 (2021).

22. Jazayeri, M. & Movshon, J. A. Optimal representation of sensory information by neural populations. *Nat. Neurosci.* **9**, 690–696 (2006).

23. Latham, P. E., Deneve, S. & Pouget, A. Optimal computation with attractor networks. *J. Physiology-paris* **97**, 683–694 (2003).

24. Deneve, S., Latham, P. & Pouget, A. Efficient computation and cue integration with noisy population codes. *Nat. Neurosci.* **4**, 826–831 (2001).

25. Pouget, A., Dayan, P. & Zemel, R. S. Inference andcomputation withpopulationcodes. *Annu. Rev. Neurosci.* **26**, 381–410 (2003).

26. Amari, S.-i. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **27**, 77–87 (1977).

27. Ermentrout, B. Neural networks as spatio-temporal pattern-forming systems. *Rep. Prog. Phys.* **61**, 353–430 (1998).

28. Ben-Yishai, R., Bar-Or, R. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc. Natl Acad. Sci.* **92**, 3844–3848 (1995).

29. Liu, F. & Wang, X.-J. A common cortical circuit mechanism for perceptual categorical discrimination and veridical judgment. *PLoS Comput. Biol.* **4**, e1000253 (2008).

30. Compte, A. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).

31. 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).

32. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A model of the neural basis of the rat's sense of direction. *Adv Neural Inf Process Syst* **7**, 173–180 (1995).

33. Zhang, K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *J. Neurosci.* **16**, 2112–2126 (1996).

34. Boucheny, C., Brunel, N. & Arleo, A. A continuous attractor network model without recurrent excitation: Maintenance and integration in the head direction cell system. *J. Comp. Physiol.* **18**, 205–227 (2005).

35. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat. Neurosci.* **22**, 1512–1520 (2019).

36. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in theDrosophila-central brain. *Science* **356**, 849–853 (2017).

37. Burak, Y. & Fiete, I. R. Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput. Biol.* **5**, e1000291 (2009).

38. Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A. & Arieli, A. Spontaneously emerging cortical representations of visual attributes. *Nature* **425**, 954–956 (2003).

39. 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**, bhw106 (2016).

40. Barbosa, J. et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).

41. Wyart, V., Myers, N. & Summerfield, C. Neural mechanisms of human perceptual choice under focused and divided attention. *J. Neurosci.* **35**, 3485–3498 (2015).

42. Georgopoulos, A., Schwartz, A. & Kettner, R. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).

43. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).

44. Eckhoff, P., Wong-Lin, K. & Holmes, P. Optimality and robustness of a biophysical decision-making model under norepinephrine modulation. *J. Neurosci.* **29**, 4301–4311 (2009).

45. Niyogi, R. K. & Wong-Lin, K. Dynamic excitatory and inhibitory gain modulation can produce flexible, robust and optimal decision-making. *PLoS Comput. Biol.* **9**, e1003099 (2013).

46. Murray, J. D., Jaramillo, J. & Wang, X.-J. Working memory and decision-making in a frontoparietal circuit model. *J. Neurosci.* **37**, 12167–12186 (2017).

47. Finkelstein, A. et al. Attractor dynamics gate cortical information flow during decision-making. *Nat. Neurosci.* **24**, 843–850 (2021).

48. Salvador, A. et al. Premature commitment to uncertain beliefs during human NMDA receptor hypofunction. *bioRxiv* 2020.06.17.156539 (2020).

49. Murphy, P. R., Wilming, N., Hernandez-Bocanegra, D. C., Prat-Ortega, G. & Donner, T. H. Adaptive circuit dynamics across human cortex during evidence accumulation in changing environments. *Nat. Neurosci.* **24**, 987–997 (2021).

50. Goldman, M., Compte, A. & Wang, X.-J. Neural integrator models. In *Encyclopedia of Neuroscience*, 165–178 (Elsevier, 2009).

51. Cannon, S. C., Robinson, D. A. & Shamma, S. A proposed neural network for the integrator of the oculomotor system. *Biol. Cybernetics* **49**, 127–136 (1983).

52. Machens, C. Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* **307**, 1121–1124 (2005).

53. Song, P. Angular path integration by moving "Hill of activity": A spiking neuron model without recurrent excitation of the head-direction system. *J. Neurosci.* **25**, 1002–1014 (2005).

54. Prat-Ortega, G., Wimmer, K., Roxin, A. & de la Rocha, J. Flexible categorization in perceptual decision making. *Nat. Commun.* **12**, 1283 (2021).

55. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).

56. Green, J. et al. A neural circuit architecture for angular integration in drosophila. *Nature* **546**, 101–106 (2017).

57. Hansel, D. & Sompolinsky, H. Modeling Feature Selectivity in Local Cortical Circuits. *Methods in Neuronal Modeling: From Ions to Networks* 499–567 (1998).

58. Tass, P. Cortical pattern formation during visual hallucinations. *J. Biol. Phys.* **21**, 177–210 (1995).

59. Bressloff, P. C. Spatiotemporal dynamics of continuum neural fields. *J. Phys. A: Math. Theor.* **45**, 033001 (2011).

60. Esnaola-Acebes, J. M., Roxin, A., Avitabile, D. & Montbrió, E. Synchrony-induced modes of oscillation of a neural field model. *Phys. Rev. E* **96** (2017).

61. Bressloff, P. C. Stochastic neural field model of stimulus-dependent variability in cortical neurons. *PLoS Comput. Biol.* **15**, e1006755 (2019).

62. Brunel, N. Dynamics and plasticity of stimulus-selective persistent activity in cortical network models. *Cereb. Cortex* **13**, 1151–1161 (2003).

63. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence*, vol. 19 of *Springer Series in Synergetics* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1984).

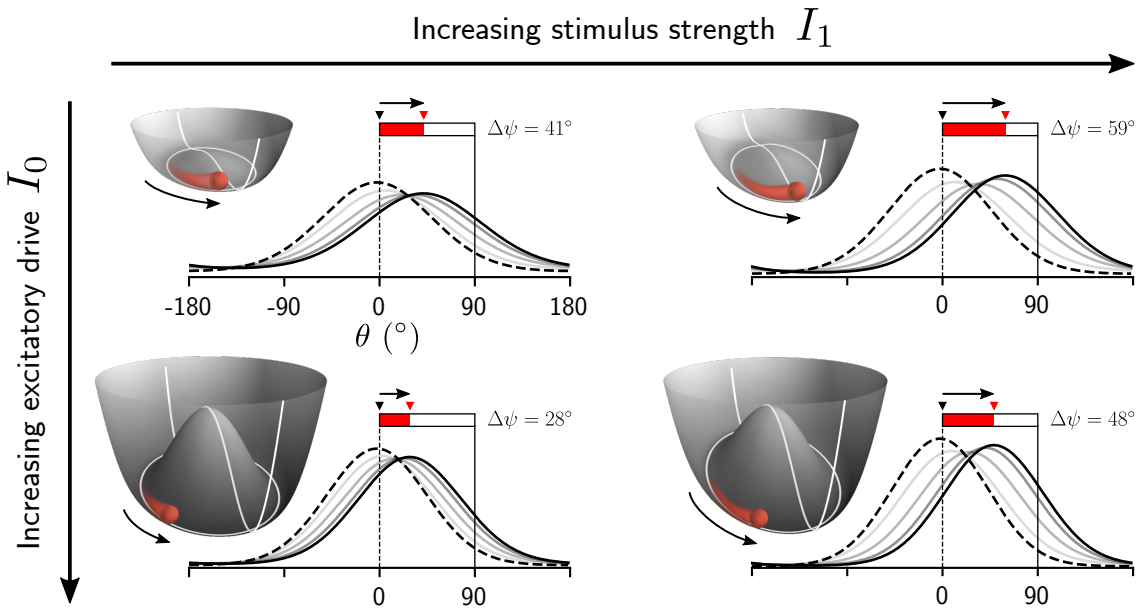
64. Hutt, A., Longtin, A. & Schimansky-Geier, L. Additive global noise delays Turing bifurcations. *Phys. Rev. Lett.* **98**, 230601 (2007).

65. Hutt, A., Longtin, A. & Schimansky-Geier, L. Additive noise-induced Turing transitions in spatial systems with application to neural fields and the Swift–Hohenberg equation. *Physica D* **237**, 755–773 (2008).

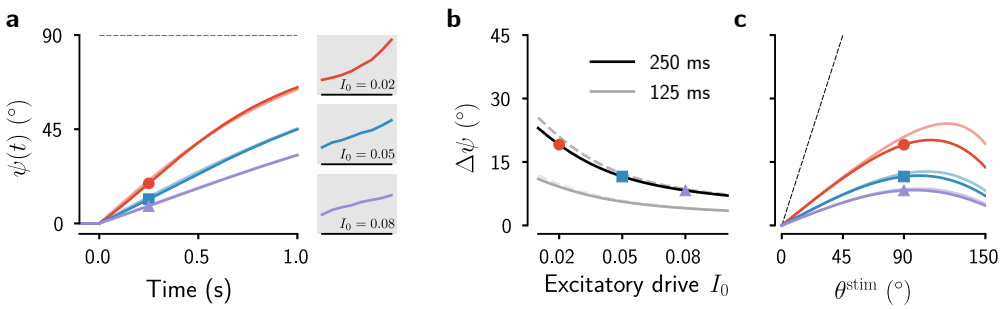
66. Kutschireiter, A., Rast, L. & Drugowitsch, J. Angular path integration by projection filtering with increment observations. *arXiv* **2102.09650** (2021).

# Acknowledgements

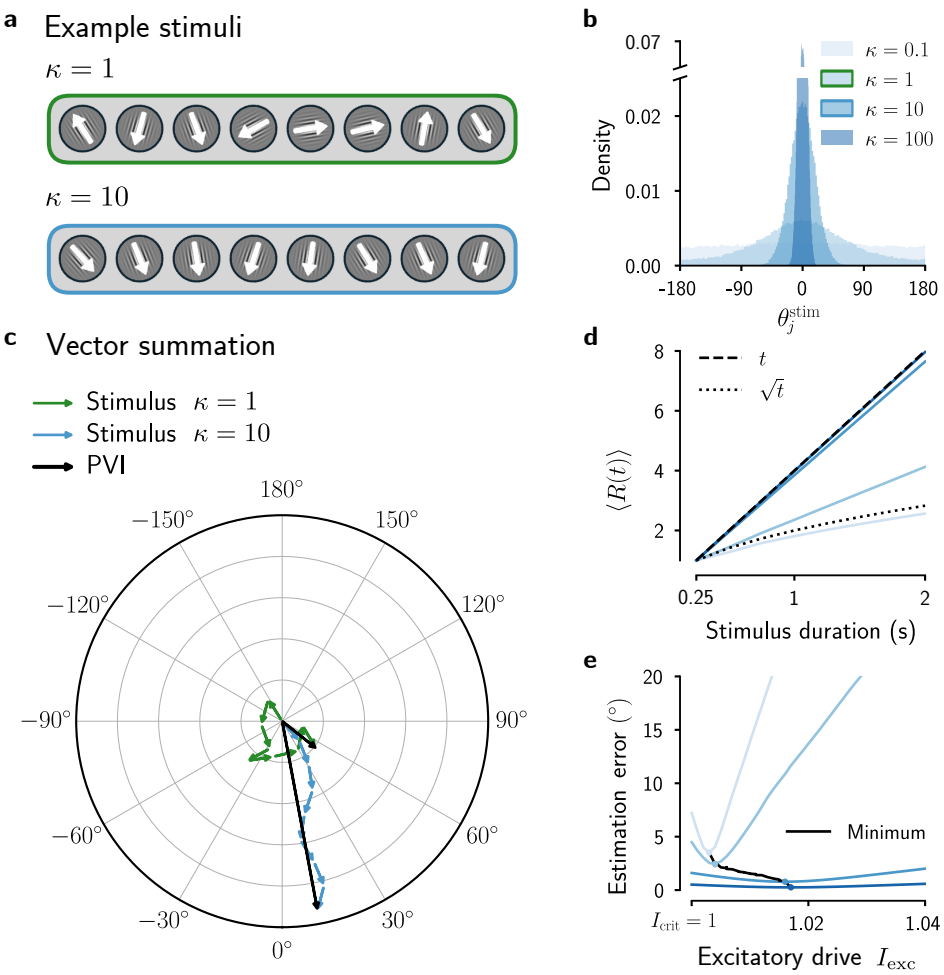
We thank Valentin Wyarth and Christopher Summerfield for sharing the experimental data and for fruitful discussions, Tobias H. Donner, Bharath Chandra Talluri and Federico Devalle for excellent discussions, and Albert Compte, Joao Barbosa and Genis Prat-Ortega for comments on the manuscript. The research leading to these results has received funding from the Spanish Ministry of Science and Innovation together with the European Regional Development Fund (RYC-2015-17236 and BFU2017-86026-R to K.W, MTM2015-71509-C2-1-R and RTI2018-097570-B-I00 to A.R.) and from the Generalitat de Catalunya (grant AGAUR 2017 SGR 1565 to A.R. and K.W.).



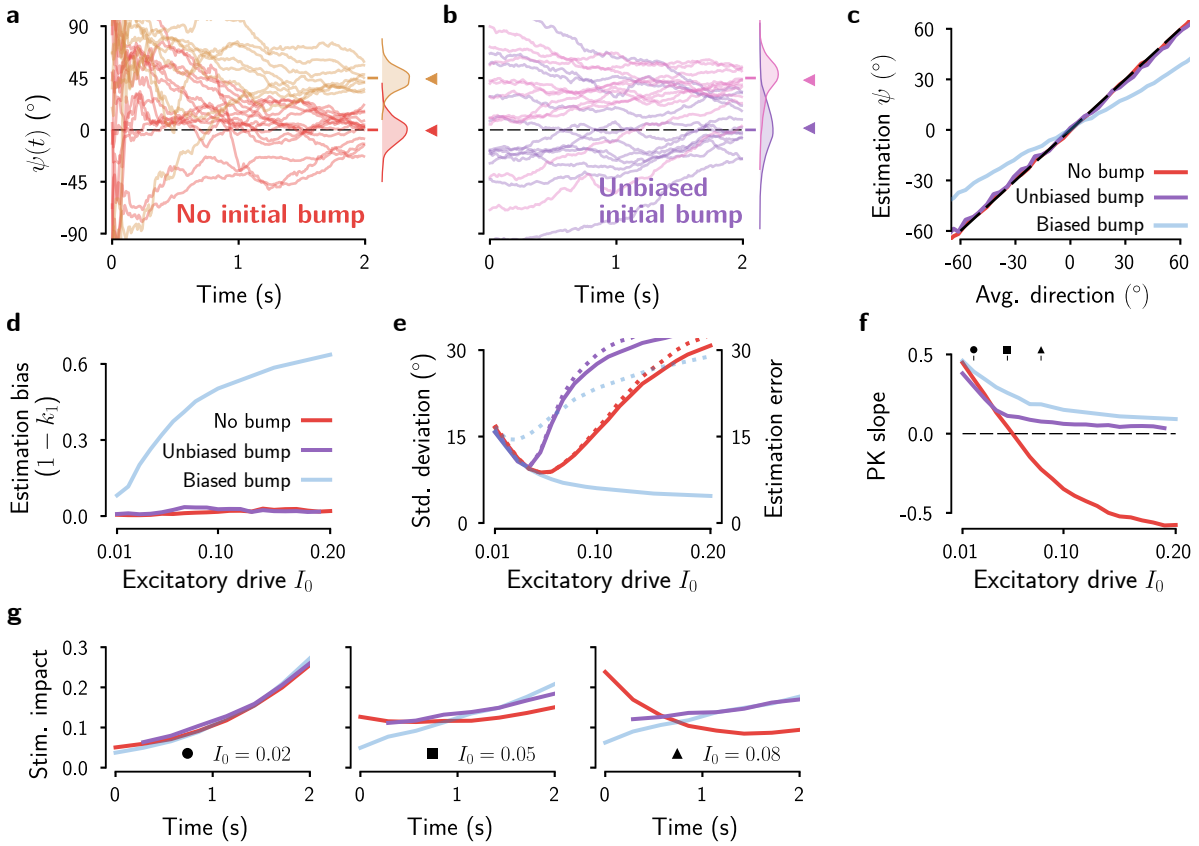

**Extended Data Fig. 1 | The global excitatory drive to the network and the stimulus strength determine the speed of movement of the activity bump.** Displacement of a bump centered at 0° in response to a stimulus with average direction of 90° and duration of 250 ms (partially filled bar above the activity profiles), for different values of the excitatory drive and input strength. Mathematically, the angular speed of the bump is described by $\dot{\psi} \propto I_1/\sqrt{I_0}$. This relationship can be understood by representing the network activity as the movement of a particle in the potential (Eq. 2) corresponding to the amplitude equation, Eq. 1. On the one hand, increasing the excitatory drive to the network, $I_0$, deepens the potential and increases the radius of the stable manifold (white circle), which in turn elongates the path that the particle has to travel in order to achieve a given angular displacement. Therefore, for the same stimulus strength, increasing the excitatory drive results in smaller displacements of the bump. On the other hand, increasing the stimulus input $I_1$ has the opposite effect, i.e., the stimulus leads to a larger angular displacement for a given $I_0$.



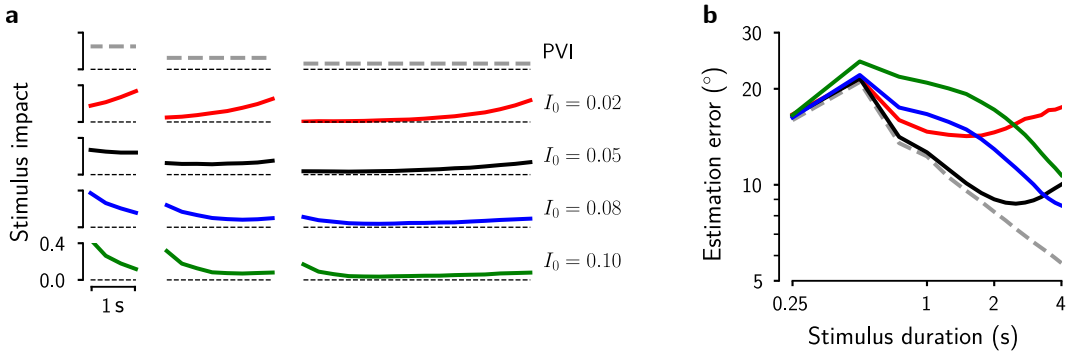

**Extended Data Fig. 2 | Impact of a single stimulus frame on the translation of the fully formed bump. a**, Evolution of the phase of the bump for a single stimulus frame of strength $I_1 = 0.005$ and duration $T_{\mathrm{stim}} = 1$ s centered at $\theta^{\mathrm{stim}} = 90°$, for three different values of the excitatory drive $I_0$. The solid lines correspond to simulations of the amplitude equation, Eq. 1, while the superimposed translucent lines have been analytically obtained using Eq. 14 which assumes a stationary bump amplitude. At the beginning of the trial the bump is fully formed and centered at 0°. Insets: PKs for each value of the excitatory drive (same as Fig. 3a). **b**, Change of phase as a function of the global excitatory drive to the network for a stimulus as in (a) but with a duration of $T_{\mathrm{stim}} = 250$ ms (solid black) and $T_{\mathrm{stim}} = 125$ ms (solid gray). The dashed lines correspond to the analytical approximation Eq. 14. **c**, Change of the phase of the bump as a function of the stimulus direction for $T_{\mathrm{stim}} = 250$ ms. The dotted line indicates a bump displacement equal to the stimulus direction.



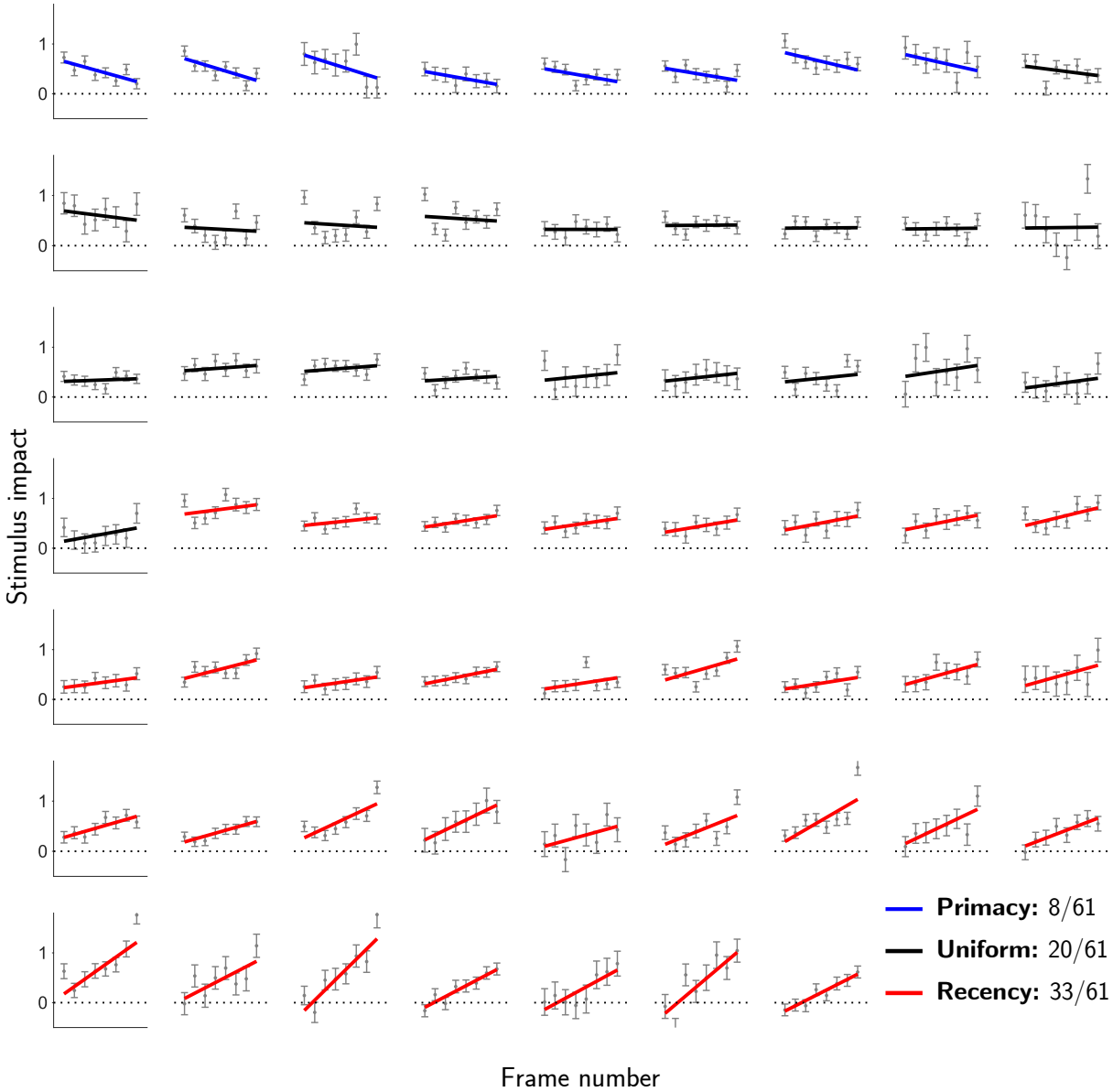

**Extended Data Fig. 3 | Perfect Vector Integrator (PVI) and the effect of the stimulus distribution on the estimation error of the bump attractor network. a-b**, Example stimuli (**a**) with directions of individual frames, $\theta_j^{\text{stim}}$ (indicated by white arrows), sampled from von Mises distributions with zero mean and different concentration parameter $\kappa$ (**b**). **c**, Illustration of stimulus integration with vector summation for the two example stimuli from **a**. Each stimulus frame corresponds to a vector of length $I_1$ (stimulus strength) and direction $\theta_j^{\text{stim}}$ (green and blue arrows; see Methods, Eqs. 27 and 28). The direction of the vector sum (black arrows) gives the circular average of the stimulus directions $\theta_j^{\text{stim}}$. This is exactly the operation carried out by the PVI (Eq. 22), with the length of the vector corresponding to the modulus $R$, and the direction of the vector corresponding to the phase $\psi$. **d**, Evolution of the trial-averaged modulus $\langle R(t) \rangle$ of the PVI (Eq. 22a) for increasing stimulus duration. Stimuli had a fixed strength ($I_1 = 1$), consisted of frames with 250 ms duration, and directions $\theta_j^{\text{stim}}$ sampled from the distributions shown in **b**. When the concentration parameter $\kappa$ is large, i.e., the dispersion of the stimulus directions is small, $\langle R(t) \rangle$ grows approximately linearly in time (as in the example shown in **c**, blue arrows). On the other hand, when the dispersion is large (e.g. $\kappa = 0.1$), $\langle R(t) \rangle$ grows much slower (as in the example shown in **c**, green arrows). A growth with $\sqrt{t}$ is shown for comparison. **e**, The root-mean-square estimation error (Methods) of the bump attractor model as a function of the global excitatory drive $I_{\text{exc}}$ for the stimulus distributions from **b**. Simulations were performed using the amplitude equation (Eq. 1) with the noise term set to zero ($\sigma_{\text{OU}} = 0$). The optimal value of the excitatory drive $I_{\text{exc}}$ to achieve the minimum error depends on the stimulus distribution (black line): For small values of $\kappa$ (large dispersion), the model performs best close to the bifurcation point $I_{\text{crit}}$. As $\kappa$ increases (small dispersion), the optimal value of $I_{\text{exc}}$ increases. A larger value of $I_{\text{exc}}$ is beneficial in this case because it allows for larger bump amplitudes needed for the integration of stimulus frames from a narrow distribution (as in **d**). This holds even though larger $I_{\text{exc}}$ increasingly lead to a juice-squeezer-shaped potential that deviates from the flat potential of the PVI (Fig. 3a). Note that $I_0 = I_{\text{exc}} - I_{\text{crit}}$ (Methods).



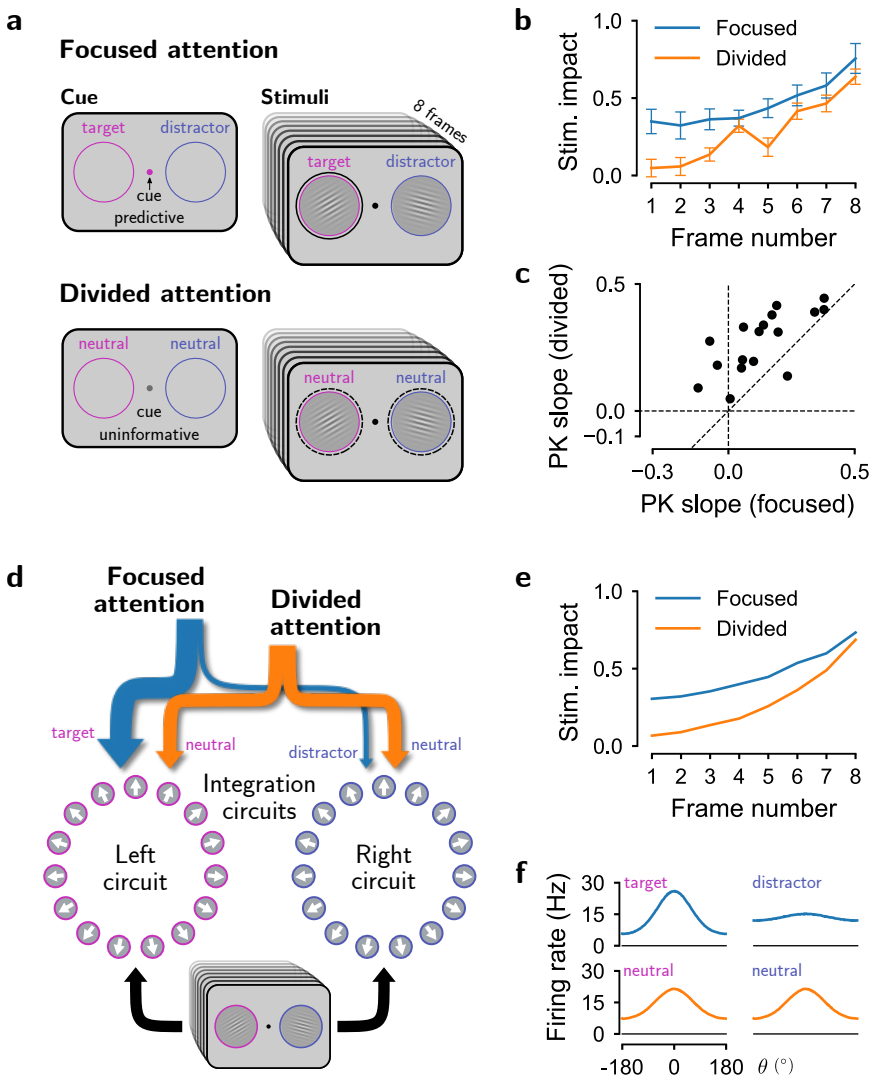

**Extended Data Fig. 4 | Bias and variance of the stimulus estimates with an initially formed bump and unbiased initial condition. a - e**, as in Fig. 5 but instead of initializing the bump at 0° in the fully formed condition, its initial position is now determined by the direction of the first stimulus frame (**b**; Methods). **a**, Bump phase in example trials when starting with homogeneous network activity, re-plotted from Fig. 5a. **c**, Continuous estimations as a function of the mean stimulus direction. **d**, Systematic bias of the estimations as a function of the global excitatory drive $I_0$. **e**, Standard deviation (solid lines, left axis) and root-mean-square estimation error (dotted lines, right axis) of the final bump position as a function of the excitatory drive $I_0$. **f**, Slope of the psychophysical kernel versus the global excitatory drive $I_0$. **g**, PKs (Methods) for the three conditions: no initial bump condition (red), initially biased bump condition (faint blue), and initially unbiased bump condition (purple). To compute the PKs of the initially unbiased condition the contribution of the first stimulus frame has been discarded. In sum, the figure confirms that the fixed initial bump condition, but not a fully formed bump per se, leads to an estimation bias. Compared to the *no initial bump* condition, the *unbiased initial bump* condition leads to a higher variance of the estimates and recency or uniform (but not primacy) PKs.




**Extended Data Fig. 5 | Dynamics of evidence integration in the bump attractor model for increasing stimulus duration. a**, PKs of the PVI and the bump attractor model for increasing stimulus durations ($T_{\text{stim}} = 1$, 2 and 4 seconds) for different values of the excitatory drive $I_0$. Stimuli were composed of frames with a fixed duration of 250 ms. The slopes corresponding to these PKs are shown in Fig. 3e. **b**, Root-mean-square estimation error (Methods) vs. stimulus duration for different levels of the excitatory drive (as in **a**). Frame duration was 250 ms. For comparison, we plot the estimation error of the PVI (Eq. 22; dashed line) to which we added noise terms obtained using Eq. 12. The estimation error in the PVI is entirely caused by this noise because the integration process itself is optimal. The particular form of the error evolution depends on the distribution of the stimulus orientations, e.g. for a pair of oppositely directed vectors the resulting vector is null and its direction is completely determined by the noise term. In general, the average modulus of the summation of two vectors will be widely distributed, and thus the fluctuation term determines the average orientation for a significant part of the distribution. This explains the initial increase of the estimation error, with a peak at 0.5 s (two stimulus frames). In contrast, the increased estimation errors—with respect the PVI—of the bump attractor model (colored lines) are due to the intrinsic dynamics of the network, and therefore depend on the integration regime.




**Extended Data Fig. 6 | Individual psychophysical kernels from human subjects.** Psychophysical kernels of individual subjects[6,7,41]. Colored lines are regression weights obtained from logistic regression with the constraint that the PK is a linear function of time (these fits are used to compute the PK slope). Plots are sorted and color coded by kernel shape. Gray errorbars show the weights obtained from logistic regression fits without applying constraints.




**Extended Data Fig. 7 | Evidence integration under focused and divided attention.** As a further test of the explicative power of the ring attractor model, we investigated whether it could shed light on the neural mechanisms underlying a change in temporal evidence weighting when attention is divided between spatially separated information sources. **a**, Category-level averaging task with focused and divided attention conditions[41]. The task is similar as in Fig. 6a, but with two streams of visual gratings presented simultaneously. In the focused attention condition, the relevant stream is indicated by a cue in the beginning of the trial whereas in the divided attention condition both streams need to be integrated and one of them will be probed randomly. **b**, PKs for the relevant stream in the two attention conditions (data re-analyzed from[41]). Errorbars indicate the SEM across $n = 17$ participants. **c**, PK slope in the divided attention condition is consistently higher than in the focused attention condition (except for 1 / 17 subjects). This shows that humans can integrate information from two streams of visual gratings simultaneously, though with an increased integration leak, that is, with PKs that show more recency effect compared to the focused attention condition in which only one stream needs to be integrated. **d**, Model architecture, with two ring circuits that integrate two spatially distinct sensory streams independently in parallel. The strength of the stimulus input is the same across focused and divided attention conditions (Methods). The integration dynamics of the circuits is controlled by sustained attention modulation signals (blue and orange arrows, respectively) that effectively add to the global excitatory drive ($I_0$) of the corresponding circuit. **e**, Model PKs for focused and divided attention conditions. We parameterized the model such that in the absence of the attention signal it operates in the recency regime. In the focused attention condition, the circuit integrating the target stimulus receives a large attention signal, increasing the total excitatory drive and pushing the network towards more uniform evidence integration (cf. Fig. 3c).

In the divided attention condition, both circuits receive an attention signal of equal strength, albeit smaller than under focused attention. Therefore, inputs are now integrated with a more recency PK. **f**, Average population firing rate at the end of the trial for the target and the distractor stream in the focused attention condition (top) and for the neutral streams in the divided attention condition (bottom). This illustrates that the strong attention signal towards the circuit integrating the target stream in the focused attention condition led to a representation of the average stimulus orientation in a larger activity bump compared to the bumps produced in the divided attention condition. This is consistent with the experimental finding that the subject's choice can be better decoded from preparatory EEG signals that reflect choice build-up in the focused attention condition (Fig. 6 in ref.[41]).

**Supplementary Video 1 | Single-trial activity of the ring network operating close to the perfect integration regime.** The **top left** panel shows the instantaneous activity of the network, with the red triangle marking the phase (center) of the bump that encodes the estimated average orientation of the stimuli. The true running average, computed by means of the perfect vector integration (PVI), is indicated by the blue triangle. At the beginning of the trial the amplitude of the bump is almost zero, and it grows and moves as the stimulus is presented. The **bottom left** panel shows the time-course of the position (phase) of the bump (red) and the PVI (blue) together with the stimulus stream on top. The light shaded region corresponds to the stimulus interval and the darker region marks the current stimulus frame. The black arrow marks the orientation of the current stimulus frame. As the simulation progresses the graph is updated to show eight additional stimulus frames. The **right panel** shows the potential (Eq. 2) corresponding to the network operating at approximately the uniform integration regime, with a global excitatory drive $I_{\mathrm{exc}} = 1.0055$. The polar graph at the bottom corresponds to the axial projection of the potential with the red trace showing the state of the network in terms of the phase and the radius of the bump as described by the amplitude equation Eq. 1. The temporal evolution of the PVI is shown for comparison (blue trace). During the first two seconds (corresponding to eight stimulus frames) the network perfectly integrates the stimulus which is reflected on the equivalence between the phase of the bump and the perfect vector integrator. Once the bump amplitude cannot grow further due to the shape of the potential, the network starts to overweight new stimuli and thus shows a recency behavior.

**Supplementary Video 2 | Recency in the ring network.** As in Supplementary Video 1 but with a lower global excitatory drive $I_{\mathrm{exc}} = 1.002$, that sets the network at the recency regime (see Fig. 3c,e). The maximal bump amplitude is small, limited by the form of the potential. As a consequence, the network overweighs the momentary stimulus information causing a constant leakage of the previously integrated sensory information, which results in a strong recency effect.

**Supplementary Video 3 | Primacy in the ring network.** As in Supplementary Video 1 but with a higher global excitatory drive $I_{\mathrm{exc}} = 1.01$, that puts the network in the primacy regime (see Fig. 3c,e). During the first two seconds (eight orientation frames) the activity bump grows rapidly which causes an overweighing of early stimulus frames. However, once the amplitude of the bump saturates (Fig. 3d), the network will start to partly erase the accumulated information (leaky integration) exactly as in the recency regime (Supplementary Video 2).