Neural Network Mechanisms Underlying Confirmation Bias in Stimulus Estimation

Jose M. Esnaola-Acebes[1], **Bharath C. Talluri**[2], **Tobias Donner**[2], **Alex Roxin**[1], **Klaus Wimmer**[1]

1. Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra (Barcelona), Spain

2. Dept. of Neurophysiology & Pathophysiology, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany

Email: jmesnaola@crm.cat, kwimmer@crm.cat

Abstract

Perception is influenced by past choices. For example, an intermittent categorical choice biases the estimation of average motion direction across two stimuli (confirmation bias). To shed light on the underlying neural mechanisms, we develop a ring attractor model that integrates stimulus direction and represents a continuous estimate of the average stimulus in the phase of an activity bump. Depending on the relative strength of sensory input compared to the intrinsic network dynamics, the model can account for qualitatively distinct decision behaviors (uniform temporal weighting, and "recency" regime). We studied two potential mechanisms underlying confirmation bias and found that they predict different modulations of the estimation curve: (i) applying an urgency signal after the first stimulus leads to a shift modulation, (ii) a feature-based attention signal that boosts stimuli that are consistent with the intermittent choice leads to a gain modulation, the main feature observed in human behavior. Our work suggests bump attractor dynamics together with feature-based attention as a potential underlying mechanism of confirmation bias in stimulus estimation tasks.

Keywords: decision making; network dynamics; computational model; confirmation bias; attention; psychophysics

Introduction

Perceptual decision making often involves making categorical judgments based on estimations of continuous stimulus features. It has recently been shown that committing to a categorical choice biases a subsequent report of the stimulus estimate by selectively increasing the weighting of choice-consistent evidence (Talluri, Urai, Tsetsos, Usher, & Donner, 2018). This phenomenon is known as confirmation bias. The underlying neural mechanisms remain poorly understood. Here, we developed a computational network model that can integrate a continuous stimulus feature such as motion direction and can also account for a subsequent categorical choice. We then studied potential mechanisms underlying confirmation bias and found that including feature-based attention in the model can explain the experimentally observed bias in stimulus estimation.

Methods

Ring Model. The dynamics of the model are described in terms of the firing rate, $r(\theta,t)$, of a neural population arranged in a ring, $\theta \in [-\pi,\pi)$, obeying the following equation

$$\tau \dot{r} = -r + \Phi\left[\frac{\tau}{2\pi}\int_{-\pi}^{\pi} w\left(\theta-\theta'\right) r\left(\theta',t\right) d\theta' + I\left(\theta,t\right)\right], \quad (1)$$

where $\tau = 20$ ms is the neurons' time constant, and Φ is the current-to-rate transfer function. The synaptic input consists of a recurrent current due to the presynaptic activity at a location θ' with a weight $w(\theta-\theta')$, and an external current $I(\theta,t)$. Both functions, I and w are written in terms of their Fourier coefficients, I_k and w_k ($k = 0,1,2,\ldots$), respectively. The homogeneous external input is $I_0 = 2$, and the connectivity function w is given by $w_0 = -2$, $w_1 = 1$ and $w_2 = 0.5$. We simulated Eq. (1) following the Euler scheme with a time step $\Delta t = 0.5$ ms, and discretizing the space θ into $N = 200$ evenly distributed angular locations. Stimulus inputs $I_{\text{stim}} = I_1 \cdot G(x,\theta^*) + n(\theta,t)$ are defined as the combination of a circular Gaussian function G with average direction θ^*, and noisy fluctuations $n(\theta,t)$. The stimulus amplitude is $I_1 = 0.1$ and its width is $\sigma_{\text{stim}} = 10^\circ$. Noisy inputs are modeled as independent Ornstein-Uhlenbeck processes for each neuron ($\tau_{\text{OU}} = 1$ ms, $\sigma_{\text{OU}} = 0.4$). The urgency signal of Fig. 2B is modeled as an external input that combines two Gaussians of amplitude 0.25, mean $\pm 90^\circ$ and standard deviation of 10°. Finally, the attentional modulation of Fig. 2C is modeled by multiplying the second stimulus by a factor $[1 + A_1 \cdot G_1(x,45^\circ) + A_2 \cdot G_2(x,45^\circ)]$ for CW trials and $[1 + A_1 \cdot G_1(x,-45^\circ) + A_2 \cdot G_2(x,45^\circ)]$ for CCW trials, with $A_1 = 1.5$, $A_2 = -0.5$, $\sigma_1 = 20^\circ$, $\sigma_2 = 30^\circ$.

Dynamics of the bump attractor. For stationary homogeneous inputs, Eq. (1) has either an homogenous or a localized solution depending on the spatial profile of the connectivity function w (Ben-Yishai, Bar-Or, & Sompolinsky, 1995). Near the bifurcation, the dynamics of the bump reduce to the normal form of a supercritical Turing bifurcation, and can be expressed in terms of the amplitude, R, and phase, Ψ, of the bump as

$$\tau \dot{R} = I_0 R + I_1 \cos\left(\Psi - \theta^*\right) - cR^3 + \xi_1(t), \quad (2a)$$

$$\tau \dot{\Psi} = -\frac{I_1}{R}\sin\left(\Psi - \theta^*\right) + \frac{\xi_2(t)}{R}, \quad (2b)$$

where c is a constant that depends on the particular form of the connectivity profile, θ^* is the average direction of the stimulus, and $\xi_1(t)$ and $\xi_2(t)$ are independent Ornstein-Uhlenbeck processes.

Results

Stimulus integration with bump attractor dynamics. We studied the integration of noisy stimuli with a continuous fea-

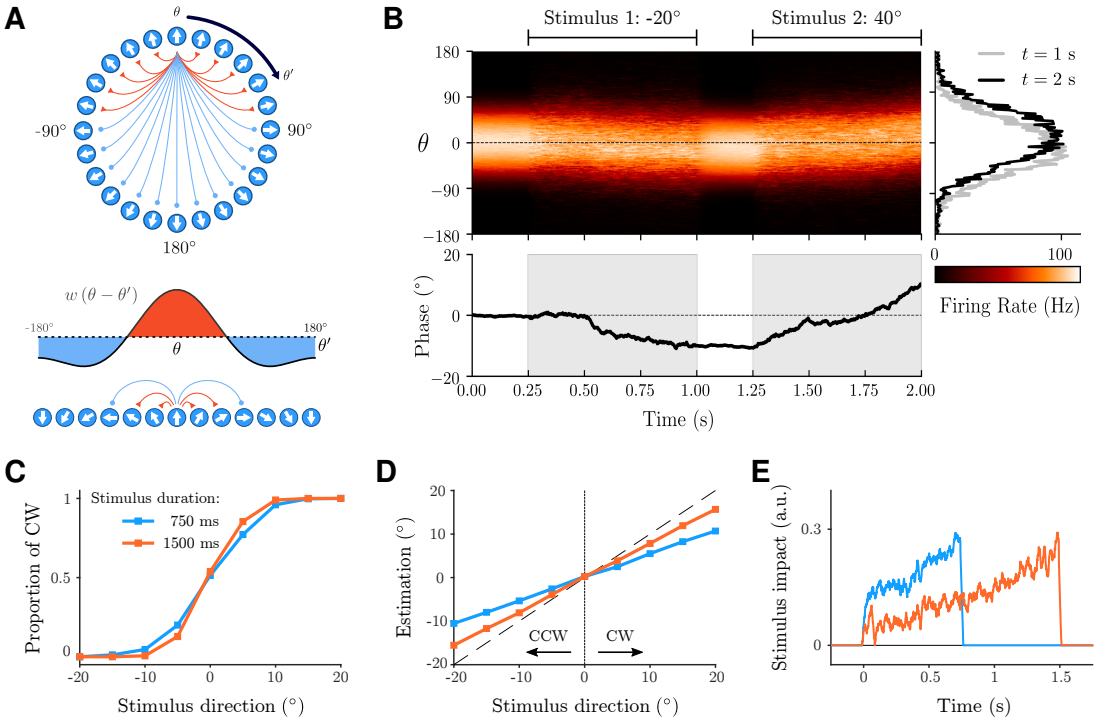

Figure 1: Stimulus estimation and categorization in the bump attractor network. **(A)** Schematic of the ring network with Mexican hat like connectivity. **(B)** Network activity in a single trial starting with a bump of activity at 0° followed by two noisy stimuli (average direction -20° and $+40^\circ$, respectively). Top left: color-coded firing rate. Top-right: Activity bump after the first and second stimulus. Bottom: temporal evolution of the bump phase. **(C)** Proportion of CW choices as a function of stimulus direction, for two different stimulus durations. A CW choice corresponds to a positive bump phase at the end of the trial. **(D)** Continuous stimulus estimation as a function of stimulus direction, obtained from the same simulations as in (C). The estimation corresponds to the phase of the bump at the end of the trial. **(E)** Time-course of evidence integration (psychophysical kernels) obtained for average stimulus directions of 0°.

ture (e.g. a random dot stimulus with different net motion directions), in a neural network model based on a ring attractor network (Fig. 1). Due to strong recurrent connectivity, a bump of activity emerges in this model at a position determined by the input, and this bump state persists when the input is removed (Fig. 1B). Similar ring models have previously been studied in the context of orientation selectivity in primary visual cortex (Ben-Yishai et al., 1995) and in the context of spatial working memory (Compte, Brunel, Goldman-Rakic, & Wang, 2000). However, here we take a different perspective and interpret the position (phase) of the activity bump as the estimate of the integrated stimulus direction.

We found that the transient population response to changing stimulus input effectively integrates the stimulus as is needed in typical decision making and estimation tasks that require accumulation of evidence over time (Fig. 1B). In the example trial (Fig. 1B), the activity bump starts at 0° (the initial position may correspond to a reference in a given task or it may be random if no prior knowledge is available). The first noisy stimulus with -20° average direction results in a transient population response: the bump initiates a slow movement towards the -20° position, with only minimal change in its amplitude ("virtual rotation", Ben-Yishai et al. (1995)). The second stimulus, with 40° average direction, causes a slow bump movement towards 40°. With the chosen parameters, the network computes approximately the average of the two stimuli and the resulting bump position is close to 10°.

To investigate whether the stimulus-dependent drift of the bump is actually yielding a continuous estimation of accumulated sensory input, we simulated many trials with noisy input stimuli of different average motion directions (Fig. 1C-E). We found that the probability of a CW choice depends on the evidence strength (Fig. 1C), similar to the observed performance ins psychophysical experiments (Talluri et al., 2018). This psychometric curve becomes steeper for longer stimulus duration (blue and orange lines in Fig. 1C), indicating that the model is able to integrate sensory evidence over long timescales. The precision of the estimation of the stimulus average also improves with stimulus duration (Fig. 1D).

In order to directly test how the stimulus is integrated over time, we computed the model's psychophysical kernel, obtained from the difference between the average stimuli yield-

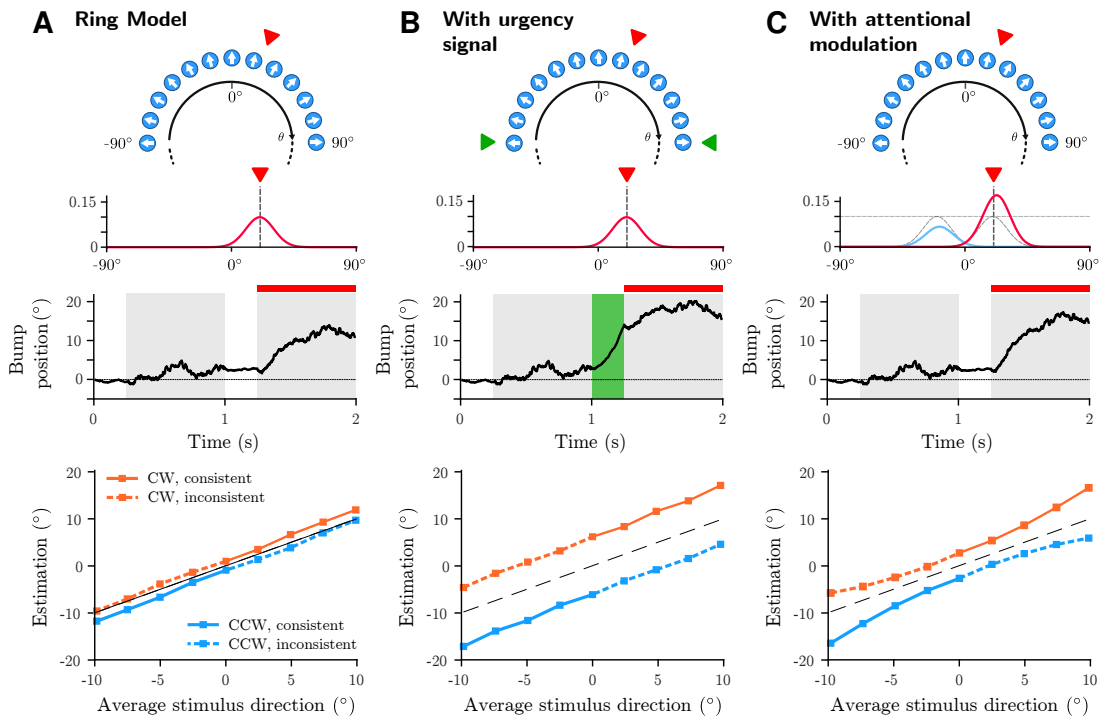

Figure 2: Confirmation bias in continuous estimation after a categorical choice. **(A)** Ring model without additional mechanisms. Top: Model architecture and illustration of a stimulus with 20° average direction . Middle: Single trial with a 0° stimulus leading to a positive bump position corresponding to a CW choice, followed by a 20° stimulus. Bottom: Continuous estimation after the second stimulus as a function of the average of the two stimuli, separately for CW and CCW choices. **(B)** Ring model with urgency signal applied in the decision phase after the first stimulus. **(C)** Ring model with choice-dependent attentional modulation. The exact same stimulus fluctuations and internal noise has been used in the three example trials in order to make the bump trajectory comparable across the three models.

ing each choice (Fig. 1E). Stimulus fluctuations impacted the choice throughout the trial, with increasing weight towards the end of the stimulus (recency effect; e.g. Cheadle et al. (2014)). This integration dynamics can be understood with the help of a reduced two-dimensional equation for the amplitude and the phase of the bump (Eqs. (2a) and (2b) in Methods). The model can account for qualitatively distinct decision behaviors, depending on the relative strength of sensory stimuli, I_1, compared to the amplitude of the bump, R. When sensory inputs dominate over the intrinsic network dynamics, later parts of the stimulus have an higher impact on the final phase and the categorical choice than earlier parts (recency regime; Fig. 1E). On the other hand, when the internal dynamics are stronger, the temporal weighting of stimulus information is uniform (not shown). In sum, the ring model can accumulate evidence at prolonged time scales and it can reproduce the experimentally observed dependence of discrimination and estimation accuracy on stimulus direction (Talluri et al., 2018).

Neural network mechanisms underlying confirmation bias. In order to reveal the potential neural mechanisms underlying confirmation bias, we then used the model to mimic a recent psychophysical experiment that required both a categorical choice and a continuous estimation (Talluri et al., 2018). In the experiment, subjects viewed two successive random dot motion stimuli, and they had to make a categorical choice (CW vs CCW) after the first stimulus and a continuous estimation of the average direction across both stimuli after the second stimulus. Subjects successfully integrated evidence across both stimuli but their estimations were biased reflecting selectively enhanced sensitivity for the second stimulus if it was consistent with the intermittent choice.

We tested whether two plausible mechanisms in the model can account for the observed confirmation bias: an urgency signal applied after the first stimulus and an attentional gain modulation during the second stimulus (Fig. 2). For the ease of exposition we restricted our simulations to trials where the first stimulus has 0° average direction, and the bias effect can simply be visualized by comparing the estimation as a function of the average stimulus direction for CW vs CCW choices (Fig. 2, bottom panel). Since the first stimulus is uninformative, the intermittent choice is determined only by stimulus fluctuations and internal noise (Fig. 2A, middle), and approximately half of the trials lead to a CW and a CCW choice, respectively. This

effect of the noise explains why in the model without any additional mechanism the estimation is slightly different for CW vs. CCW choices (Fig. 2A, bottom). This difference is however much smaller than the psychophysically observed biases.

A possible way of introducing a choice-dependent bias is to include an urgency signal in the model (Fig. 2B). This urgency signal consists of an unspecific input to neurons around -90° and 90° in the ring (see Methods), causing a movement of the bump towards either -90° or $+90^\circ$, depending on which one is closer to the current bump position. Thus, the bump moves further away from the decision boundary, making it easier to read out the categorical choice. Moreover, the magnitude of the displacement is almost independent of the bump phase and its direction is always consistent with the intermittent choice. This leads to a choice-dependent shift modulation of the final stimulus estimation (Fig. 2B, bottom), an effect that contributed to the confirmation bias in the psychophysical experiment.

The second way of introducing a choice-dependent bias was through including an attentional modulation of the second stimulus (Fig. 2C). The attention effect is modeled as a feature-dependent gain modulation of visual neurons (Martinez-Trujillo & Treue, 2004) that provide the input to the ring model. Importantly, the "attentional spotlight" is chosen to be consistent with the intermittent choice (Fig. 2C, top). This gain modulation leads to a boost of the second stimulus if it is consistent with the choice, and to a decrease if it is inconsistent. The boost is larger for stimuli that are further away from the 0° reference, yielding a choice-dependent gain-modulation of the final estimation (Fig. 2C, bottom). This gain-modulation resembles the stimulus-dependent confirmation bias that was the main signature of confirmation bias in human subjects.

Discussion

We showed that bump attractor dynamics emerging in a neural network model with ring-like connectivity structure provides a mechanistic model for stimulus integration of continuous features such as motion direction. A categorical choice (e.g. CW vs. CCW relative to a reference direction) can then be obtained by "reading out" the bump position. Using this model, we studied the potential neural mechanisms underlying the recently observed choice-dependent bias that an intermittent choice has on the estimation of average motion direction across two stimuli (Talluri et al., 2018). We found that a shift modulation of the estimation curve, observed in a minority of subjects, can be explained by an urgency signal applied after the first stimulus. Gain modulation of the estimation curve, as observed in most of the subjects, can be explained by a feature-based attention signal that boosts stimuli that are consistent with the intermittent choice. In future work we plan put to test the hypothesis that feature-based attention underlies the confirmation bias in human MEG experiments.

Our novel model has several unique features: Modulation of the excitability of the network results in either recency or uniform temporal integration, a temporal weighting that has for example been observed in a category-level averaging task (Cheadle et al., 2014; Wyart, de Gardelle, Scholl, & Summerfield, 2012). This uniform or recency temporal integration is in contrast to classical attractor models of decision making that typically give more weight to the early part of the stimulus (primacy effect; Wang (2002); Wimmer et al. (2015)). Furthermore, the architecture of our model is similar to the classical model of parametric working memory (Compte et al., 2000) but operating in a regime where it slowly integrates inputs. Both models show persistent stimulus-related activity in delay periods after stimulus presentation. Bump attractor dynamics might therefore be a unifying neural mechanism underlying both working memory and decision making, and it may provide a substrate for evidence integration at prolonged timescales (Waskom & Kiani, 2018).

Acknowledgments

Funded by the Spanish Ministry of Science and the European Regional Development Fund (RYC-2015-17236, BFU2017-86026-R, MTM2015-71509-C2-1-R) and the Generalitat de Catalunya (AGAUR 2017 SGR 1565).

References

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *PNAS*, *92*(9), 3844-3848.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429-1441.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb Cortex*, *10*(9), 910-923.

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr Biol*, *14*(9), 744-751.

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Curr Biol*, *28*(19), 3128-3135.e8.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955-968.

Waskom, M. L., & Kiani, R. (2018). Decision making through integration of sensory evidence at prolonged timescales. *Curr Biol*, *28*(23), 3850-3856.e9.

Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A., & de la Rocha, J. (2015). Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat Commun*, *6*, 6177.

Wyart, V., de Gardelle, V., Scholl, J., & Summerfield, C. (2012). Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron*, *76*(4), 847-858.