Neuroscience

From recency to central tendency biases in working memory: a unifying network model

Vezha Boboeva, Alberto Pezzotta, Claudia Clopath, Athena Akrami

Sainsbury Wellcome Centre, University College London • Department of Bioengineering, Imperial College London • Gatsby Computational Neuroscience Unit, University College London • The Francis Crick Institute

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
January 5, 2024 (this version)

Reviewed preprint version 1
August 2, 2023

Sent for peer review
March 9, 2023

Posted to preprint server
January 27, 2023

Abstract

The central tendency bias, or contraction bias, is a phenomenon where the judgment of the magnitude of items held in working memory appears to be biased towards the average of past observations. It is assumed to be an optimal strategy by the brain, and commonly thought of as an expression of the brain's ability to learn the statistical structure of sensory input. On the other hand, recency biases such as serial dependence are also commonly observed, and are thought to reflect the content of working memory. Recent results from an auditory delayed comparison task in rats, suggest that both biases may be more related than previously thought: when the posterior parietal cortex (PPC) was silenced, both short-term and contraction biases were reduced. By proposing a model of the circuit that may be involved in generating the behavior, we show that a volatile working memory content susceptible to shifting to the past sensory experience – producing short-term sensory history biases – naturally leads to contraction bias. The errors, occurring at the level of individual trials, are sampled from the full distribution of the stimuli, and are not due to a gradual shift of the memory towards the sensory distribution's mean. Our results are consistent with a broad set of behavioral findings and provide predictions of performance across different stimulus distributions and timings, delay intervals, as well as neuronal dynamics in putative working memory areas. Finally, we validate our model by performing a set of human psychophysics experiments of an auditory parametric working memory task.

eLife assessment

This **important** study combines disparate results from both psychophysics and neural silencing experiments to suggest a new interpretation of how animals and humans represent and interpret recent events in our memory. A key aspect of the model put forward here is the presence of discrete jumps in neural activity within the posterior parietal region of the cortex. The model is distinct from other models, and the authors provide **convincing** evidence to support it both from existing results as well as from novel experiments.

Introduction

A fundamental question in neuroscience relates to how brains efficiently process the statistical regularities of the environment to guide behavior. Exploiting such regularities may be of great value to survival in the natural environment, but may lead to biases in laboratory tasks. Repeatedly observed across species and sensory modalities is the central tendency ("contraction") bias, where performance in perceptual tasks seemingly reflects a shift of the working memory representation towards the mean of the sensory history [1–6]. Equally common are sequential biases, either attractive or repulsive, towards the immediate sensory history [7, 5, 8–14, 6, 15].

It is commonly thought that these biases occur due to disparate mechanisms - contraction bias is widely thought to be a result of learning the statistical structure of the environment, whereas serial biases are thought to reflect the contents of working memory [16, 17]. Recent evidence, however, challenges this picture: our recent study of a parametric working memory task discovered that the rat posterior parietal cortex (PPC) plays a key role in modulating contraction bias [7]. When the region is optogenetically inactivated, contraction bias is attenuated. Intriguingly, however, this is also accompanied by the suppression of bias effects induced by the recent history of the stimuli, suggesting that the two phenomena may be interrelated. Interestingly, other behavioral components, including working memory of immediate sensory stimuli (in the current trial), remain intact. In another recent study with humans, a double dissociation was reported between three cohorts of subjects: subjects on the autistic spectrum (ASD) expressed reduced biases due to recent statistics, whereas dyslexic subjects (DYS) expressed reduced biases towards long-term statistics, relative to neurotypical subjects (NT) [16]. Finally, further complicating the picture is the observation of not only attractive serial dependency, but also repulsive biases [18]. It is as of yet unclear how such biases occur and what mechanisms underlie such history dependencies.

These findings stimulate the question of whether contraction bias and the different types of serial biases are actually related, and if so, how. Although normative models have been proposed to explain these effects [19, 18, 16], the neural mechanisms and circuits underlying them remain poorly understood. We address this question through a model of the putative circuit involved in giving rise to the behavior observed in [7]. Our model consists of two continuous (bump) attractor sub-networks, representing a working memory (WM) module and the PPC. Given the finding that PPC neurons carry more information about stimuli presented during previous trials, the PPC module integrates inputs over a longer timescale relative to the WM network, and incorporates firing rate adaptation.

We find that both contraction bias and short-term sensory history effects emerge in the WM network as a result of inputs from the PPC network. Importantly, we see that these effects do not necessarily occur due to separate mechanisms. Rather, in our model, contraction bias emerges as a statistical effect of errors in working memory, occurring due to the persisting memory of stimuli shown in the preceding trials. The integration of this persisting memory in the WM module competes with that of the stimulus in the current trial, giving rise to short-term history effects. We conclude that contraction biases in such paradigms may not necessarily reflect explicit learning of regularities or an "attraction towards the mean", on individual trials. Rather, it may be an effect emerging at the level of average performance, when in each trial, errors are made according to the recent sensory experiences whose distribution follow that of the input stimuli. Furthermore, using the same model, we also show that the biases towards long-term (short-term) statistics inferred from the performance of ASD (DYS) subjects [16] may actually reflect short-term biases extending more or less into the past with respect to NT subjects, challenging the hypothesis of a double-dissociation mechanism. Last, we show that as a result of neuronal integration of inputs and adaptation, in addition to attraction effects occurring on a short timescale, repulsion effects are observed on a longer timescale [18].

We make specific predictions on neuronal dynamics in the PPC and downstream working memory areas, as well as how contraction bias may be altered, upon manipulations of the sensory stimulus distribution, inter-trial and inter-stimulus delay intervals. We show agreements between the model and our previous results in humans and rats. Finally, we test our model predictions by performing new human auditory parametric working memory tasks. The data is in agreement with our model and not with an aternative Bayesian model.

1 Results

1.1 The PPC as a slower integrator network

Parametric working memory (PWM) tasks involve the sequential comparison of two graded stimuli that differ along a physical dimension and are separated by a delay interval of a few seconds (**Fig. 1 A** and **B**) [20, 7, 19]. A key feature emerging from these studies is contraction bias, where the averaged performance is as if the memory of the first stimulus progressively shifts towards the center of a prior distribution built from past sensory history (**Fig. 1 C**). Additionally, biases towards the most recent sensory stimuli (immediately preceding trials) have also been documented [7, 5].

In order to investigate the circuit mechanisms by which such biases may occur, we use two identical one-dimensional continuous attractor networks to model WM and PPC modules. Neurons are arranged according to their preferential firing locations in a continuous stimulus space, representing the amplitude of auditory stimuli. Excitatory recurrent connections between neurons are symmetric and a monotonically decreasing function of the distance between the preferential firing fields of neurons, allowing neurons to mutually excite one another; inhibition, instead, is uniform. Together, both allow a localized bump of activity to form and be sustained (**Fig. 1 D** and **E**) [21–29]. Both networks have free boundary conditions. Neurons in the WM network receive inputs from neurons in the PPC coding for the same stimulus amplitude (**Fig. 1 D**). Building on experimental findings [30–35], we designed the PPC network such that it integrates activity over a longer timescale compared to the WM network (*Sect. 3.1*). Moreover, neurons in the PPC are equipped with neural adaptation, that can be thought of as a threshold that dynamically follows the activation of a neuron over a longer timescale.

To simulate the parametric WM task, at the beginning of each trial, the network is provided with a stimulus s_1 for a short time via an external current I_{ext} as input to a set of neurons (see **Tab. 1**). Following s_1, after a delay interval, a second stimulus s_2 is presented (**Fig. 1 E**). The pair (s_1, s_2) is drawn from the stimulus set shown in **Fig. 1 B**, where they are all equally distant from the diagonal $s_1 = s_2$, and are therefore of equal nominal discrimination, or difficulty. The stimuli (s_1, s_2) are co-varied in each trial so that the task cannot be solved by relying on only one of the stimuli [36]. As in the study in Ref. [7] using an interleaved design, across consecutive trials, the inter-stimulus delay intervals are randomized and sampled uniformly between 2, 6 and 10 seconds. The inter-trial interval, instead, is fixed at 5 seconds.

We additionally include psychometric pairs (indicated in the box in **Fig. 1 B**) where the distance to the diagonal, hence the discrimination difficulty, is varied. The task is a binary comparison task that aims at classifying whether $s_1 < s_2$ or vice-versa. In order to solve the task, we record the activity of the WM network at two time points: slightly before and after the onset of s_2 (**Fig. 1 E**). We repeat this procedure across many different trials, and use the recorded activity to assess performance (see (*Sect. 3.2*) for details). Importantly, at the end of each trial, the activity of both networks is not re-initialized, and the state of the network at the end of the trial serves as the initial network configuration for the next trial.

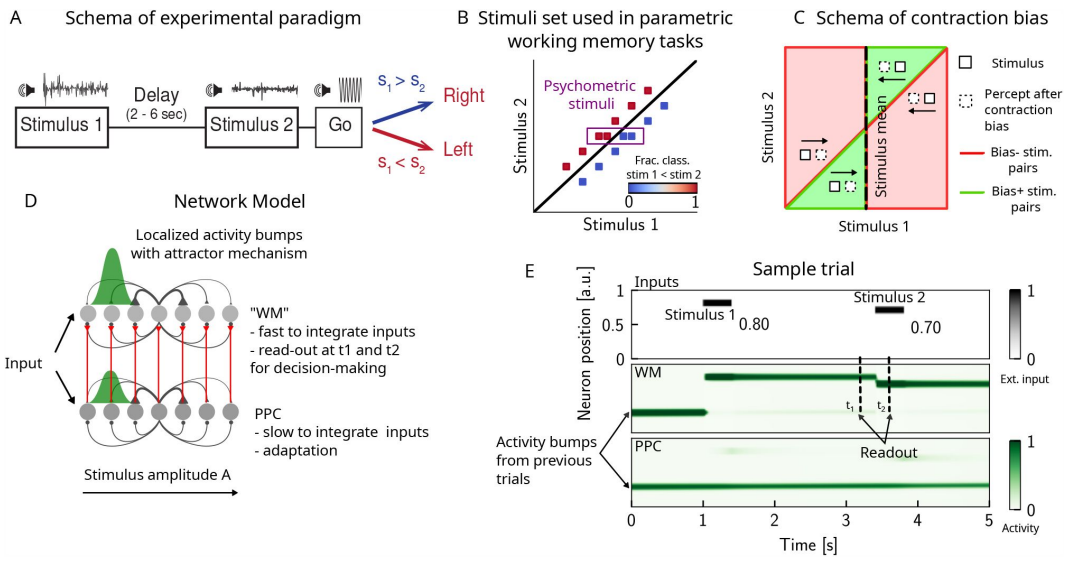

Figure 1

The PPC as a slower integrator network.

(A) In any given trial, a pair of stimuli (here, sounds) separated by a variable delay interval is presented to a subject. After the second stimulus, and after a go cue, the subject must decide which of the two sounds is louder by pressing a key (humans) or nose-poking in an appropriate port (rats). **(B)** The stimulus set. The stimuli are linearly separable, and stimulus pairs are equally distant from the $s_1 = s_2$ diagonal. Error-free performance corresponds to network dynamics from which it is possible to classify all the stimuli below the diagonal as $s_1 > s_2$ (shown in blue) and all stimuli above the diagonal as $s_1 < s_2$ (shown in red). An example of a correct trial can be seen in (E). In order to assay the psychometric threshold, several additional pairs of stimuli are included (purple box), where the distance to the diagonal $s_1 = s_2$ is systematically changed. The colorbar expresses the fraction classified as $s_1 < s_2$. **(C)** Schematics of contraction bias in delayed comparison tasks. Performance is a function of the difference between the two stimuli, and is impacted by contraction bias, where the base stimulus s_1 is perceived as closer to the mean stimulus. This leads to a better/worse (green/red area) performance, depending on whether this “attraction” increases (Bias+) or decreases (Bias-) the discrimination between the base stimulus s_1 and the comparison stimulus s_2. **(D)** Our model is composed of two modules, representing working memory (WM), and sensory history (PPC). Each module is a continuous one-dimensional attractor network. Both networks are identical except for the timescales over which they integrate external inputs; PPC has a significantly longer integration timescale and its neurons are additionally equipped with neuronal adaptation. The neurons in the WM network receive input from those in the PPC, through connections (red lines) between neurons coding for the same stimulus. Neurons (gray dots) are arranged according to their preferential firing locations. The excitatory recurrent connections between neurons in each network are a symmetric, decreasing function of their preferential firing locations, whereas the inhibitory connections are uniform (black lines). For simplicity, connections are shown for a single pre-synaptic neuron (where there is a bump in green). When a sufficient amount of input is given to a network, a bump of activity is formed, and sustained in the network when the external input is subsequently removed. This activity in the WM network is read out at two time points: slightly before and after the onset of the second stimulus, and is used to assess performance. **(E)** The task involves the comparison of two sequentially presented stimuli, separated by a delay interval (top panel, black lines). The WM network integrates and responds to inputs quickly (middle panel), while the PPC network integrates inputs more slowly (bottom panel). As a result, external inputs (corresponding to stimulus 1 and 2) are enough to displace the bump of activity in the WM network, but not in the PPC. Instead, inputs coming from the PPC into the WM network are not sufficient to displace the activity bump, and the trial is consequently classified as correct. In the PPC, instead, the activity bump corresponds to a stimulus shown in previous trials.

Parameter	Symbol	Default value
Number of neurons	N	2000
Neuronal gain	β	5
Range of excitatory interactions [in units of stimulus space length]	d_0	0.02
Strength of inhibitory weights	J_0	0.2
Strength of excitatory weights	J_e	1
Time scale of neuronal integration in WM net [s]	τ^W	0.01
Time scale of neuronal integration in PPC [s]	τ^P	0.5
Time scale of neuronal adaptation in PPC [s]	τ_θ^P	7.5
Amplitude of adaptation current in PPC	D^P	0.3
Amplitude of external inputs	I_{ext}	1
Strength of weights from PPC to WM net	$J^{P\to W}$	0.5
Duration of stimuli [s]		0.4
Delay interval [s]		$[2, 6, 10]$
Inter-trial interval [s]		6
Width of box stimulus [in units of stimulus space length]	δs	0.05

Table 1

Simulation parameters, when not explicitly mentioned. Used to produce **Figs. 1**, **2**, **3**, **4**, **5**, **8**, **S2**, **S3**, **S5**.

1.2 Contraction bias and short-term stimulus history effects as a result of PPC network activity

Probing the WM network performance on psychometric stimuli (**Fig. 1 B**, purple box, 10% of all trials) shows that the comparison behavior is not error-free, and that the psychometric curves (different colors) differ from the optimal step function (**Fig. 2 A**, green dashed line). The performance on pychometric trials is also better for shorter inter-stimulus delay intervals, as has been shown in previous work [37, 7]. In our model, errors are caused by the displacement of the activity bump in the WM network, due to the inputs from the PPC network. These displacements in the WM activity bump can result in different outcomes: by displacing it *away* from the second stimulus, they either do not affect the performance or improve it (**Fig. 2 B** right panel, "Bias+"), if noise is present. Conversely, the performance can suffer, if the displacement of the activity bump is *towards* the second stimulus (**Fig. 2 B** left panel, "Bias-"). Note, however, that in these two specific trials, the activity bump in PPC is strong, and it displaces the activity bump in the WM network, but this is not the only kind of dynamics present in the network (see *Sect.1.3* for a more detailed analysis of the network dynamics).

Performance on stimulus pairs that are equally distant from the $s_1 = s_2$ diagonal can be similarly impacted and the network produces a pattern of errors that is consistent with contraction bias: performance is at its minimum for stimulus pairs in which s_1 is either largest or smallest, and at its maximum for stimulus pairs in which s_2 is largest or smallest (**Fig. 2 C**, left panel) [19, 38, 7, 39, 40]. These results are consistent with the performance of humans and rats on the auditory task, as previously reported (**Fig. 2 C**, middle and right panels, data from Akrami et al 2018 [7]).

Can the same circuit also give rise to short-term sensory history biases [7, 41]? We analyzed the fraction of trials the network response was "$s_1 < s_2$" in the current trial conditioned on stimulus pairs presented in the previous trial, and we found that the network behavior is indeed modulated by the preceding trial's stimulus pairs (**Fig. 2 D**, panel 1). We quantified these history effects as well as how many trials back they extend to. We computed the bias by plotting, for each particular pair (of stimuli) presented at the current trial, the fraction of trials the network response was "$s_1 < s_2$" as a function of the pair presented in the previous trial minus the mean performance over all previous trial pairs (**Fig. 2 D**, panel 2) [7]. Independent of the current trial, the previous trial exerts an "attractive" effect, expressed by the negative slope of the line: when the previous pair of stimuli is small, s_1 in the current trial is, on average, misclassified as smaller than it actually is, giving rise to the attractive bias in the comparison performance; the converse holds true when the previous pair of stimuli happens to be large. These effects extend to two trials back, and are consistent with the performance of humans and rats on the auditory task (**Fig. 2 D**, panels 3-6, data from Akrami et al 2018 [7]).

It has been shown that inactivating the PPC, in rats performing the auditory delayed comparison task, markedly reduces the magnitude of contraction bias, without impacting non-sensory biases [7]. We assay the causal role of the PPC in generating the sensory history effects as well as contraction bias by weakening the connections from the PPC to the WM network, mimicking the inactivation of the PPC. In this case, we see that the performance for the psychometric stimuli is greatly improved (yellow curve, **Fig. 2 E**, top panel), consistent also with the inactivation of the PPC in rodents (yellow curve, **Fig. 2 E**, bottom panel, data from Akrami et al 2018 [7]). Performance is improved also for all pairs of stimuli in the stimulus set (**Fig. S3 A**). The breakdown of the network response in the current trial conditioned on the specific stimulus pair preceding it reveals that the previous trial no longer exerts a notable modulating effect on the current trial (**Fig. S3 B**). Quantifying this bias by subtracting the mean performance over all of the previous pairs reveals that the attractive bias is virtually eliminated (yellow curve, **Fig. 2 F**, left panel), consistent with findings in rats (**Fig. 2 F**, right panel, data from Akrami et al 2018 [7]).

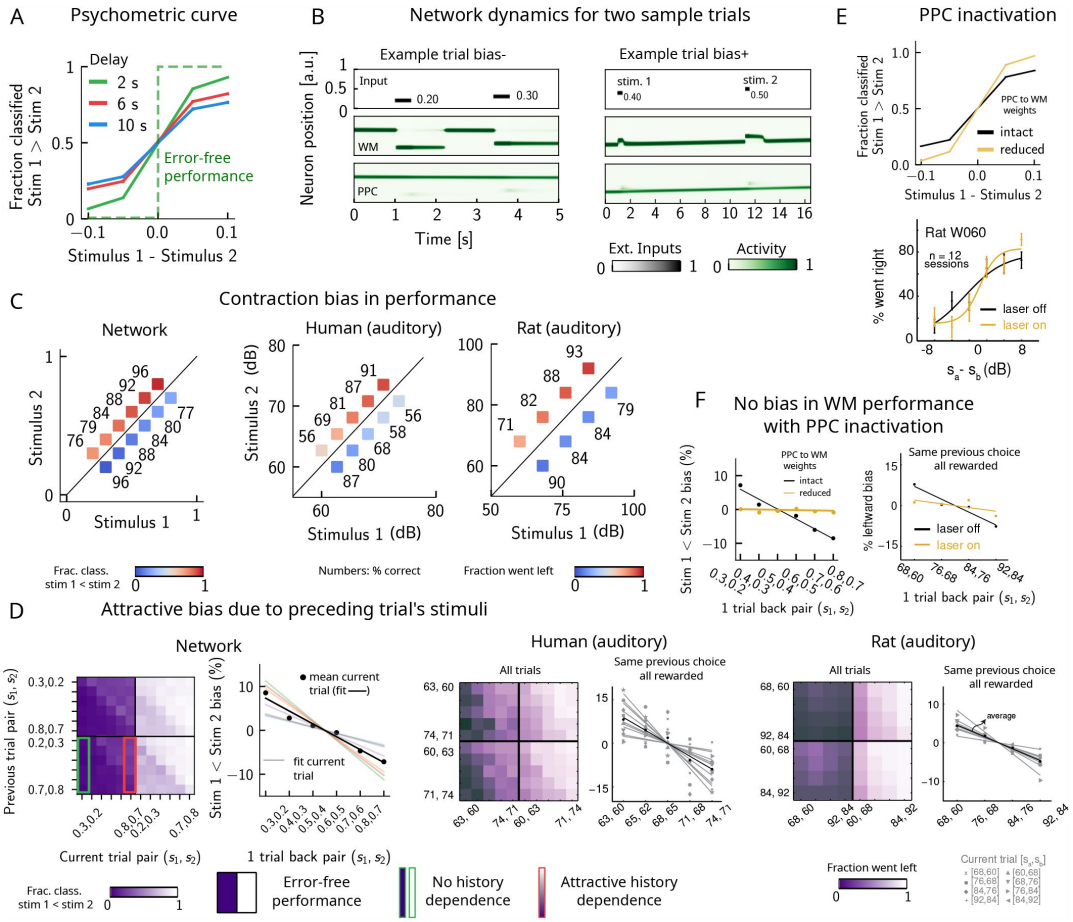

Figure 2

Contraction bias and short-term sensory history effects as a result of PPC network activity.

(A) Performance of network model for psychometric stimuli (colored lines) is not error-free (green dashed lines). A shorter inter-stimulus delay interval yields a better performance. **(B)** Errors occur due to the displacement of the bump representing the first stimulus s_1 in the WM network. Depending on the direction of this displacement with respect to s_2, this can give rise to trials in which the comparison task becomes harder (easier), leading to negative (positive) biases (top and bottom panels). Top sub-panel: stimuli presented to both networks in time. Middle/ bottom sub-panels show activity of WM and PPC networks (in green). **(C)** Left: performance is affected by contraction bias – a gradual accumulation of errors for stimuli below (above) the diagonal upon increasing (decreasing) s_1. Colorbar indicates fraction of trials classified as $s_1 < s_2$. Middle and Right: for comparison, data from the auditory version of the task performed in humans and rats. Data from Ref. [7]. **(D)** Panel 1: For each combination of current (x-axis) and previous trial's stimulus pair (y-axis), fraction of trials classified as $s_1 < s_2$ (colorbar). Performance is affected by preceding trial's stimulus pair (modulation along the y-axis). For readability, only some tick-labels are shown. Panel 2: bias, quantifying the (attractive) effect of previous stimulus pairs. Colored lines correspond to linear fits of this bias for each pair of stimuli in the current trial. Black dots correspond to average over all current stimuli, and black line is a linear fit. These history effects are attractive: the smaller the previous stimulus, the higher the probability of classifying the first stimulus of the current trial s_1 as small, and vice-versa. Panel 3: human auditory trials. Percentage of trials in which humans chose left for each combination of current and previous stimuli; vertical modulation indicates attractive effect of preceding trial. Panel 4: Percentage of trials in which humans chose left minus the average value of left choices, as a function of the stimuli of the previous trial, for fixed previous trial response choice and reward. Panel 5 and 6: same as panels 3 and 4 but with rat auditory trials. Data from Ref. [7]. **(E)** Top: performance of network, when the weights from the PPC to the WM network is weakened, is improved for psychometric stimuli (yellow curve), relative to the intact network (black curve). Bottom: psychometric curves for rats (only shown for one rat) are closer to error-free during PPC inactivation (yellow) than during control trials (black). **(F)** Left: the attractive bias due to the effect of the previous trial is present with the default weights (black line), but is eliminated with reduced weights (yellow line). Right: while there is bias induced by previous stimuli in the control experiment (black), this bias is reduced under PPC inactivation (yellow). Experimental figures reproduced with permission from Ref.[7].

Together, our results suggest a possible circuit through which both contraction bias and short-term history effects in a parametric working memory task may arise. The main features of our model are two continuous attractor networks, both integrating the same external inputs, but operating over different timescales. Crucially, the slower one, a model of the PPC, includes neuronal adaptation, and provides input to the faster one, intended as a WM circuit. In the next section, we show how the slow integration and firing rate adaptation in the PPC network give rise to the observed effects of sensory history.

1.3 Multiple timescales at the core of short-term sensory history effects

The activity bumps in the PPC and WM networks undergo different dynamics, due to the different timescales with which they integrate inputs, the presence of adaptation in the PPC, and the presence of global inhibition. The WM network integrates inputs over a shorter timescale, and therefore the activity bump follows the external input with high fidelity (**Fig. 3 A** (purple bumps) and B (purple line)). The PPC network, instead, has a longer integration timescale, and therefore fails to sufficiently integrate the input to induce a displacement of the bump to the location of a new stimulus, at each single trial. This is mainly due to the competition between the inputs from the recurrent connections sustaining the bump, and the external stimuli that are integrated elsewhere: if the former is stronger, the bump is not displaced. If, however, these inputs are weaker, they will not displace it, but may still exert a weakening effect via the global inhibition in the connectivity. The external input, as well as the presence of adaptation (**Fig. S1 B** and **C**) induce a small continuous drift of the activity bump that is already present from the previous trials (lower right panel of **Fig. 2 B**, **Fig. 3 A** (pink bumps) and B (pink line)). The build-up of adaptation in the PPC network, combined with the global inhibition from other neurons receiving external inputs, can extinguish the bump in that location (see also **Fig. S1** for more details). Following this, the PPC network can make a transition to an incoming stimulus position (that may be either s_1 or s_2), and a new bump is formed. The resulting dynamics in the PPC are a mixture of slow drift over a few trials, followed by occasional jumps (**Fig. 3 A**).

As a result of such dynamics, relative to the WM network, the activity bump in the PPC represents the stimuli corresponding to the current trial in a smaller fraction of the trials, and represents stimuli presented in the previous trial in a larger fraction of the trials (**Fig. 3 C**). This yields short-term sensory history effects in our model (**Fig. 2 D**, and **E**), as input from the PPC lead to the displacement of the WM bump to other locations (**Fig. 3 D**). Given that neurons in the WM network integrate this input, once it has built up sufficiently, it can surpass the self-sustaining inputs from the recurrent connections in the WM network. The WM bump, then, can move to a new location, given by the position of the bump in the PPC (**Fig. 3 D**). As the input from the PPC builds up gradually, the probability of bump displacement in WM increases over time. This in return leads to an increased probability of contraction bias (**Fig. 3 E**), and short-term history (one-trial back) biases (**Fig. 3 F**), as the inter-stimulus delay interval increases.

Additionally, a non-adapted input from the PPC has a larger likelihood of displacing the WM bump. This is highest immediately following the formation of a new bump in the PPC, or in other words, following a “bump jump” (**Fig. 3 F**). As a result, one can reason that those trials immediately following a jump in the PPC are the ones that should yield the maximal bias towards stimuli presented in the previous trial. We therefore separated trials according to whether or not a jump has occurred in the PPC in the preceding trial (we define a jump to have occurred if the bump location across two consecutive trials in the PPC is displaced by an amount larger than the typical width of the bump (*Sect. 3.1*)). In line with this reasoning, only the set that included trials with jumps in the preceding trial yields a one-trial back bias (**Fig. 3 G**).

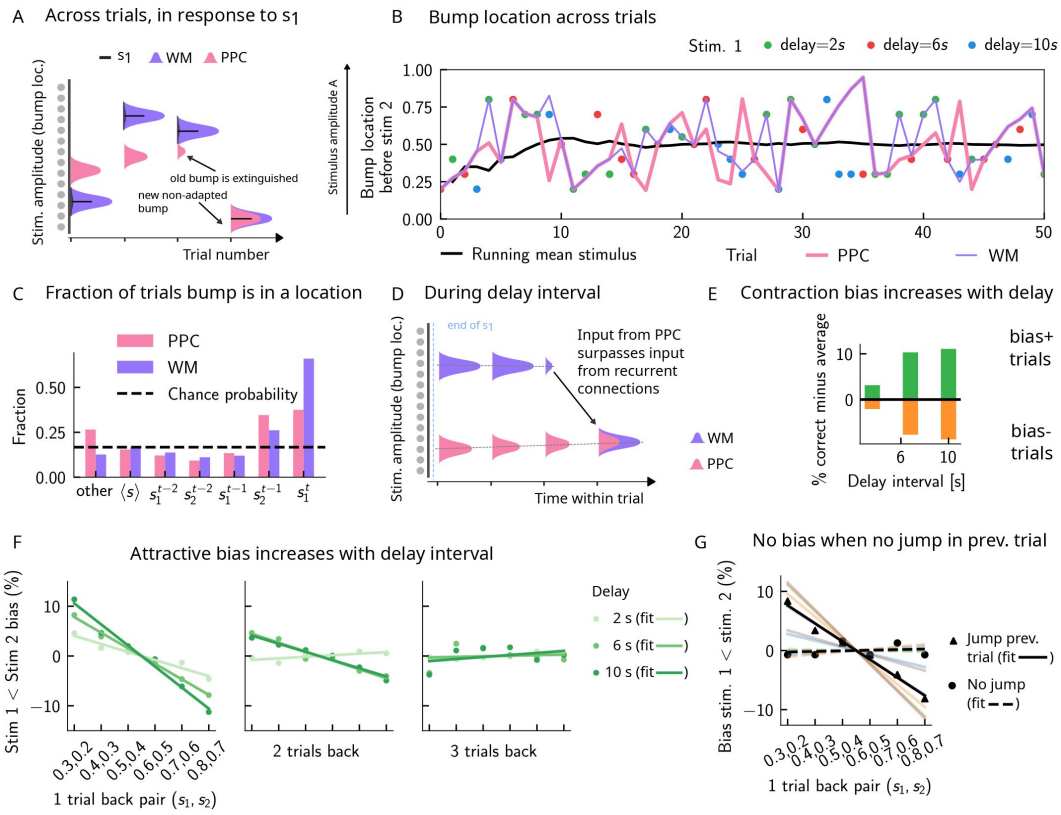

Figure 3

Multiple timescales at the core of short-term sensory history effects.

(A) Schematics of activity bump dynamics in the WM vs PPC network. Whereas the WM responds quickly to external inputs, the bump in the PPC drifts slowly and adapts, until it is extinguished and a new bump forms. **(B)** The location of the activity bump in both the PPC (pink line) and the WM (purple line) networks, immediately before the onset of the second stimulus s_2 of each trial. This location corresponds to the amplitude of the stimulus being encoded. The bump in the WM network closely represents the stimulus s_1 (shown in colored dots, each color corresponding to a different delay interval). The PPC network, instead, being slower to integrate inputs, displays a continuous drift of the activity bump across a few trials, before it jumps to a new stimulus location, due to the combined effect of inhibition from incoming inputs and adaptation that extinguishes previous activity. **(C)** Fraction of trials in which the bump location corresponds to the base stimulus that has been presented (s_1^t) in the current trial, as well as the two preceding trials $(s_2^{t-1}$ to $s_1^{t-2})$. In the WM network, in the majority of trials, the bump coincides with the first stimulus of the current trial s_1^t . In a smaller fraction of the trials, it corresponds to the previous stimulus s_2^{t-1}, due to the input from the PPC. In the PPC network instead, a smaller fraction of trials consist of the activity bump coinciding with the current stimulus s_1^t. Relative to the WM network, the bump is more likely to coincide with the previous trial's comparison stimulus (s_2^{t-1}) . **(D)** During the inter-stimulus delay interval, in the absence of external sensory inputs, the activity bump in the WM network is mainly sustained endogenously by the recurrent inputs. It may, however, be destabilized by the continual integration of inputs from the PPC. **(E)** As a result, with an increasing delay interval, given that more errors are made, contraction bias increases. Green (orange) bars correspond to the performance in Bias+ (Bias-) regions, relative to the mean performance over all pairs (**Fig. 1 C**). **(F)** Left and middle: longer delay intervals allow for a longer integration times which in turn lead to a larger frequency of WM disruptions due to previous trials, leading to a larger previous-trial attractive biases (2*s* vs. 6*s* vs. 10*s*). Right: Weak repulsive effects for larger delays become apparent. Colored dots correspond to the bias computed for different values of the inter-stimulus delay interval, while colored lines correspond to their linear fits. **(G)** When neuronal adaptation is at its lowest in the PPC i.e. following a bump jump, the WM bump is maximally susceptible to inputs from the PPC. The attractive bias (towards previous stimuli) is present in trials in which the PPC network underwent a jump in the previous trial (black triangles, with black line a linear fit). Such biases are absent in trials where no jumps occur in the PPC in the previous trial (black dots, with dashed line a linear fit). Colored lines correspond to bias for specific pairs of stimuli in the current trial, regular lines for the jump condition, and dashed for the no jump condition.

Removing neuronal adaptation entirely from the PPC network further corroborates this result. In this case, the network dynamics show a very different behavior: the activity bump in the PPC undergoes a smooth drift (**Fig. S2 A**), and the bump distribution is much more peaked around the mean (**Fig. S2 B**), relative to when adaptation is present (**Fig. 4 A**). In this regime, there are no jumps in the PPC (**Fig. S2 A**), and the activity bump corresponds to the stimuli presented in the previous trial in a fewer fraction of the trials (**Fig. S2 C**), relative to when adaptation is present (**Fig. 3 B**). As a result, no short-term history effects can be observed (**Fig. S2 C** and **D**), even though a strong contraction bias persists (**Fig. S2 E**).

As in the study in Ref.[7], we can further study the impact of the PPC on the dynamics of the WM network by weakening the weights from the PPC to the WM network, mimicking the inactivation of PPC (**Fig. 2 E** and **F**, **Fig. S3 A** and **B**). Under this manipulation, the trajectory of the activity bump in the WM network immediately before the onset of the second stimulus s_2 closely follows the external input, consistent with an enhanced WM function (**Fig. S3 C** and **D**).

The drift-jump dynamics in our model of the PPC give rise to short-term (notably one and two-trial back) sensory history effects in the performance of the WM network. In addition, we observe an equally salient contraction bias (bias towards the sensory mean) in the WM network's performance, increasing with the delay period (**Fig. 3 E**). However, we find that the activity bump in both the WM and the PPC network corresponds to the mean over all stimuli in only a small fraction of trials, expected by chance (**Fig. 3 B**, see *Sect. 4.1* for how it is calculated). Rather, the bump is located more often at the current trial stimulus (s_1^t), and to a lesser extent, at the location of stimuli presented at the previous trial (s_2^{t-1}). As a result, contraction bias in our model cannot be attributed to the representation of the running sensory average in the PPC. In the next section, we show how contraction bias arises as an averaged effect, when single trial errors occur due to short-term sensory history biases.

1.4 Errors are drawn from the marginal distribution of stimuli, giving rise to contraction bias

In order to illustrate the statistical origin of contraction bias in our network model, we consider a mathematical scheme of its performance (**Fig. 4 B**). In this simple formulation, we assume that the first stimulus to be kept in working memory, s_1^t, is volatile. As a result, in a fraction ϵ of the trials, it is susceptible to replacement with another stimulus $\hat{s}$ (by the input from the PPC, that has a given distribution p_m (**Fig. 4 A**)). However, this replacement does not always lead to an error, as evidenced by Bias- and Bias+ trials (i.e. those trials in which the performance is affected, negatively and positively, respectively (**Fig. 2 B**)). For each stimulus pair, the probability to make an error, p_e, is integral of p_m over values lying on the wrong side of the $s_1 = s_2$ diagonal (**Fig. 4 C**). For instance, for stimulus pairs below the diagonal (**Fig. 4 C**, blue squares) the trial outcome is erroneous only if $\hat{s}$ is displaced above the diagonal (red part of the distribution). As one can see, the area above the diagonal increases as s_1 increases, giving rise to a gradual increase in error rates (**Fig. 4 C**). This mathematical model can capture the performance of the attractor network model, as can be seen through the fit of the network performance, when using the bump distribution in the PPC as p_m, and ϵ as a free parameter (see *Eq. 9* in *Sect. 4.2*, **Fig. 4 D, E**).

Can this simple statistical model also capture the behavior of rats and humans (**Fig. 2 C**)? We carried out the same analysis for rats and humans, by replacing the bump location distribution of PPC with that of the marginal distribution of the stimuli provided in the task, based on the observation that the former is well-approximated by the latter (**Fig. 4 A**). In this case, we see that the model roughly captures the empirical data (**Fig. 4 F** and **G**), with the addition of another parameter δ that accounts for the lapse rate. Interestingly, such "lapse" also occurs in the network model (as seen by the small amount of errors for pairs of stimuli where s_2 is smallest and

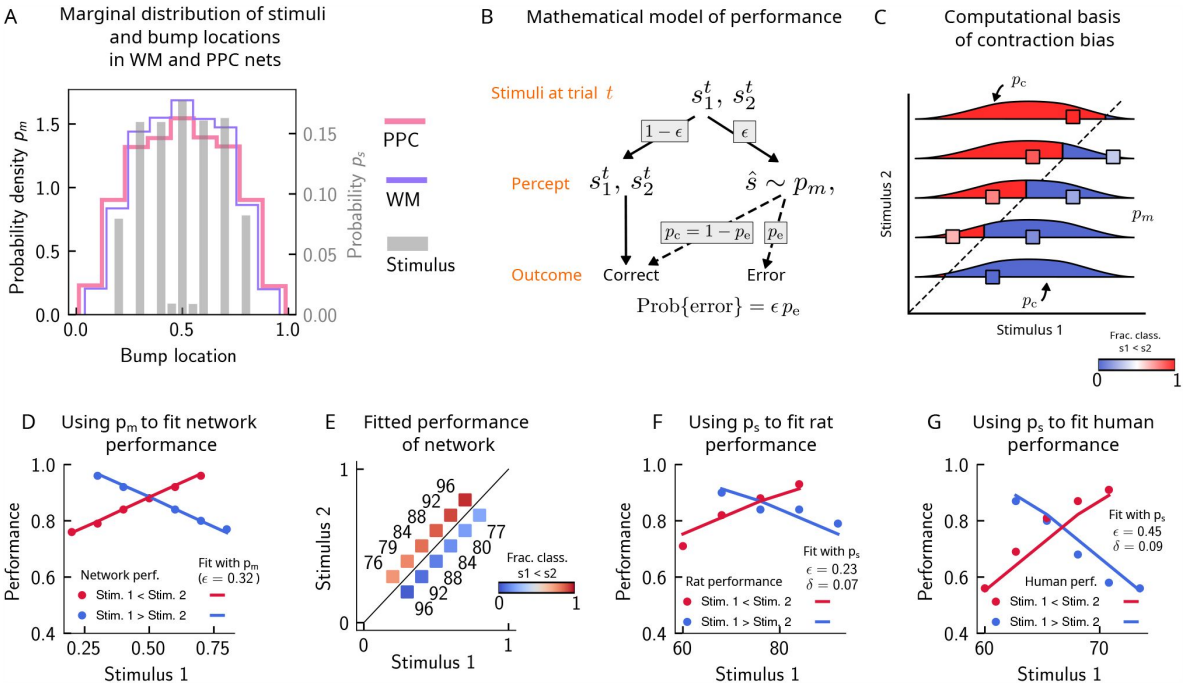

Figure 4

Errors are drawn from the marginal distribution of stimuli, giving rise to contraction bias.

(A) The bump locations in both the WM network (in pink) and the PPC network (in purple) have identical distributions to that of the input stimulus (marginal over s_1 or s_2, shown in gray). **(B)** A simple mathematical model illustrates how contraction bias emerges as a result of a volatile working memory for s_1. A given trial consists of two stimuli s_1^t and s_2^t. We assume that the encoding of the second stimulus s_2^t is error-free, contrary to the first stimulus that is prone to change, with probability ϵ. Furthermore, when s_1 does change, it is replaced by another stimulus, $\hat{s}$ (imposed by the input from the PPC in our network model). Therefore, $\hat{s}$ is drawn from the marginal distribution of bump locations in the PPC, which is similar to the marginal stimulus distribution (see panel B), p_m (see also *Sect. 4.2*). Depending on the new location of $\hat{s}$, the comparison to s_2 can either lead to an erroneous choice (Bias-, with probability p_e) or a correct one (Bias+, with probability $p_c = 1 - p_e$). **(C)** The distribution of bump locations in PPC (from which replacements $\hat{s}$ are sampled) is overlaid on the stimulus set, and repeated for each value of s_2. For pairs below the diagonal, where $s_1 > s_2$ (blue squares), the trial outcome will be an error if the displaced WM bump $\hat{s}$ ends up above the diagonal (red section of the p_m distribution). The probability to make an error, p_e, equals the integral of p_m over values above the diagonal (red part), which increases as s_1 increases. Vice versa, for pairs above the diagonal ($s_1 < s_2$, red squares), p_e equals the integral of p_m over values below the diagonal, which increases as s_1 decreases. **(D)** The performance of the attractor network as a function of the first stimulus s_1, in red dots for pairs of stimuli where $s_1 > s_2$, and in blue dots for pairs of stimuli where $s_1 < s_2$. The solid lines are fits of the performance of the network using *Eq. 9*, with ϵ as a free parameter. **(E)** Numbers correspond to the performance, same as in (D), while colors expresses the fraction classified as $s_1 < s_2$ (colorbar), to illustrate the contraction bias. **(F)** Performance of rats performing the auditory delayed-comparison task in Ref.[7]. Dots correspond to the empirical data, while the lines are fits with the statistical model, using the distribution of stimuli. The additional parameter δ captures the lapse rate. **(G)** Same as (F), but with humans performing the task. Data in (F) and (G) reproduced with permission from Ref.[7].

largest (**Fig. 4 E**). This occurs because of the drift present in the PPC network, that eventually, for long enough delay intervals, causes the bump to arrive at the boundaries of the attractor, which would result in an error.

This simple analysis implies that contraction bias in the WM network in our model is not the result of the representation of the mean stimulus in the PPC, but is an effect that emerges as a result of the PPC network's sampling dynamics, mostly from recently presented stimuli. Indeed, a "contraction to the mean" hypothesis only provides a general account of which pairs of stimuli should benefit from a better performance and which should suffer, but does not explain the gradual accumulation of errors upon increasing (decreasing) s_1, for pairs below (above) the $s_1 = s_2$ diagonal [38, 39, 7]. Notably, it cannot explain why the performance in trials with pairs of stimuli where s_2 is most distant from the mean stand to benefit the most from it. All together, our model suggests that contraction bias may be a simple consequence of errors occurring at single trials, driven by inputs from the PPC that follow a distribution similar to that of the external input (**Fig. 4 B**).

1.5 Contraction bias in continuous recall

Can contraction bias also be observed in the activity of the WM network prior to binary decision-making? Many studies have evidenced contraction bias also in delayed estimation (or production) paradigms, where subjects must retain the value of a continuous parameter in WM and reproduce it after a delay [15, 42]. Given that we observe contraction bias in the behavior of the network, we reasoned that this should also be evident prior to binary decision-making. Similar to delayed estimation tasks, we therefore analyzed the position of the bump $\hat{s}$, at the end of the delay interval, for each value of s_1. Consistent with our reasoning, we observe contraction bias of the value of $\hat{s}$, as evidenced by the systematic departure of the curve corresponding to the bump location from that of the nominal value of the stimulus (**Fig. 5 A**). We also find that this contraction bias becomes greater as the delay interval increases (**Fig. 5 A**, right). We next analyzed the effect of the previous trial on the current trial by computing the displacement of the bump during the WM delay, as a function of the distance between the current trial's stimulus and the previous trial's stimulus $s_1(t) - s_2(t-1)$ (**Fig. 5 B**). We found that when this distance is larger, the displacement of the bump during WM is on average also larger (**Fig. 5 B**). This displacement is also attractive. Breaking down these effects by delay, we find that longer delays lead to greater attraction (**Fig. 5 B**, right).

These results point to attractive effects of the previous trial, leading in turn to contraction bias in our model. To better understand the dynamics leading to them, we next looked at the distribution of bump displacements conditioned on a specific value of the second stimulus of the previous trial $s_2(t-1)$ (**Fig. 5 C**). These distributions are characterized by a mode around 0, corresponding to a majority of trials in which the bump is not displaced, and another mode around $s_1(t) - s_2(t-1)$, corresponding to the displacement in the direction of the preceding trial's stimulus described in *Sect. 1.3* and **Fig. 2 B**. However, note that the variance of this second mode can be large, reflecting displacements to locations other than $s_2(t-1)$, due to the complex dynamics in both networks that we have described in detail in *Sect. 1.3*.

1.6 Model predictions

1.6.1 The stimulus distribution impacts the pattern of contraction bias through its cumulative

In our model, the pattern of errors is determined by the cumulative distribution of stimuli from the correct decision boundary $s_1 = s_2$ to the left (right) for pairs of stimuli below (above) the diagonal (**Fig. 4 C** and **Fig. S4 A**). This implies that using a stimulus set in which this distribution is deformed makes different predictions for the gradient of performance across

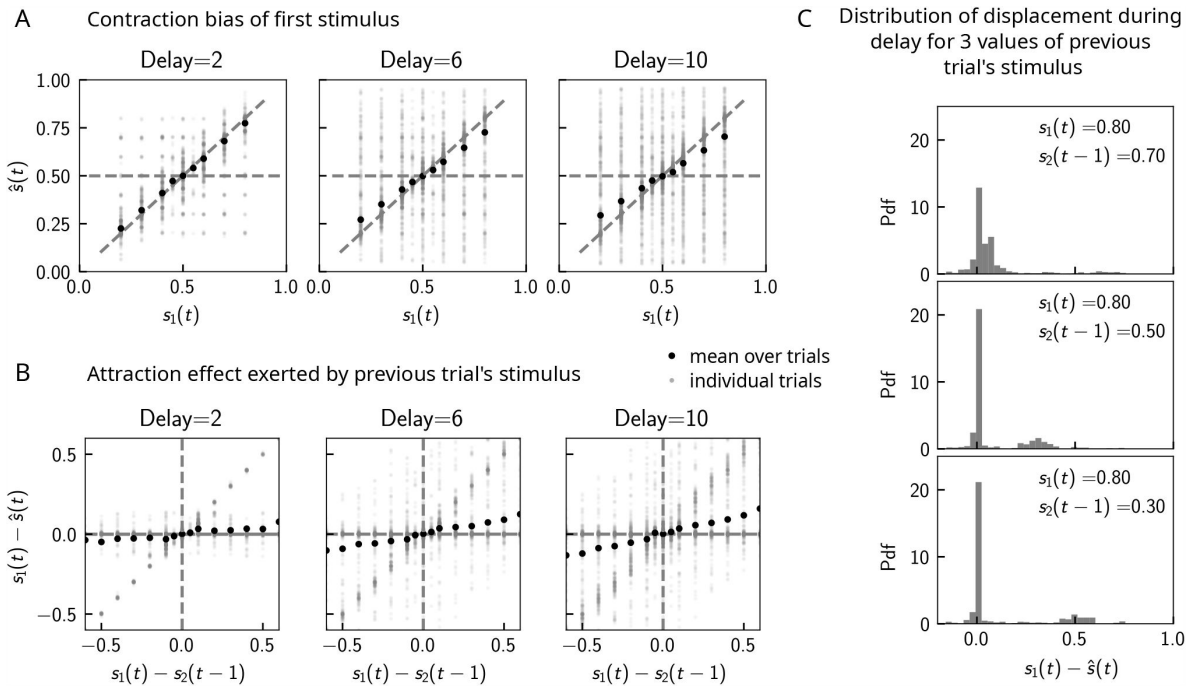

Figure 5

Contraction bias in continuous recall.

(A) We observe contraction bias of the bump of activity after the delay period ŝ: the average $\hat{s}$ over trials (black dots) deviates from the identity line (diagonal dashed line) toward the mean of the marginal stimulus distribution (0.5). This effect is stronger as the delay interval is longer (left to right panel). **(B)** This contraction bias is actually largely due to the effect of the previous trial: the larger the difference between the current trial and the previous trial's stimulus $s_2(t-1)$, the larger is this attractive effect on average. Accordingly with panel (A), this effect is stronger for longer delay intervals (left to right panel). **(C)** The distribution of the bump displacement during delay period is characterized by two modes: a main one centered around 0, corresponding to correct trials where the WM bump is not displaced during the delay interval, and another one centered around $s_1(t) - s_2(t-1)$, where the bump is displaced during WM (delay interval is randomly selected between 2, 4 and 10 seconds. We show here this distribution for three values of $s_2(t-1)$.

different stimulus pairs. A distribution that is symmetric (**Fig. S4 A**) yields an equal performance for pairs below and above the $s_1 = s_2$ diagonal (blue and red lines) when s_1 is at the mean (as well as the median, given the symmetry of the distribution). A distribution that is skewed, instead, yields an equal performance when s_1 is at the median for both pairs below and above the diagonal. For a negatively skewed distribution (**Fig. S4 B**) or positively skewed distribution (**Fig. S4 C**) the performance curves for pairs of stimuli below and above the diagonal show different concavity. For a distribution that is bimodal, the performance as a function of s_1 resembles a saddle, with equal performance for intermediate values of s_1 (**Fig. S4 D**). These results indicate that although the performance is quantitatively shaped by the form of the stimulus distribution, it persists as a monotonic function of s_1 under a wide variety of manipulations of the distributions. This is a result of the property of the cumulative function, and may underlie the ubiquity of contraction bias under different experimental conditions.

We compare the predictions from our simple statistical model to the Bayesian model in [41], outlined in *Sec. 4.3*. We compute the predicted performance of an ideal Bayesian decision maker, using a value of the uncertainty in the representation of the first stimulus ($\sigma = 0.12$) that yields the best fit with the performance of the statistical model (where the free parameter is $\epsilon = 0.5$, **Fig. S4 A, B, C**, and **D**, second panels). Our model makes different predictions across all types of distributions from that of the Bayesian model. Across all of the distributions (used as priors, in the Bayesian model), the main difference is that of a monotonic dependence of performance as a function of s_1 for our model (**Fig. S4 A, B, C**, and **D**, second panels). The biggest difference can be seen with a prior in which pairs of stimuli with extreme values are much more probable than middle-range values. Indeed, in the case of a bimodal prior, for pairs of stimuli where our model would predict a worse-than-average performance (**Fig. S4 D**, third panel), the Bayesian model predicts a very good performance (**Fig. S4 D**, fourth panel).

Do human subjects perform as predicted by our model (**Fig. 6 A**)? We tested 34 human subjects on the auditory modality of the task. The experimental protocol was identical to the one used in Ref.[7]. Briefly, participants were presented with two sounds separated by a delay interval that varied across trials (randomly selected from 2, 4 and 6 seconds). After the second sound, participants were required to decide which sound was louder by pressing the appropriate key. We tested two groups of participants on two stimulus distributions: a negatively skewed and a bimodal distribution (**Fig. 6 A**, see *Sect. 3.3* for more details). Participants performed the task with a mean accuracy of approximately 75%, across stimulus distribution groups and across pairs of stimuli (**Fig. 6 B**). The experimental data was compatible with the predictions of our model. First, for the negatively skewed stimulus distribution condition, we observe a shift of the point of equal performance to the right, relative to a symmetric distribution (**Fig. 6 C**, left panel). For the bimodal condition, such a shift is not observed, as predicted by our model (**Fig. 6 C**, right panel). Second, the monotonic behavior of the performance, as a function of s_1 also holds, across both distributions (**Fig. 6 C**). Our model provides a simple explanation: the percent correct on any given pair is given by the probability that, given a shift in the working memory representation, this representation still does not affect the outcome of the trial (**Fig. 4 C**). This probability, is given by cumulative of the probability distribution of working memory representations, for which we assume the marginal distribution of the stimuli to be a good approximation (**Fig. 4 A**). As a result, performance is a monotonic function of s_1, independent of the shape of the distribution, while the same does not always hold true for the Bayesian model (**Fig. 6 C**).

We further fit the performance of each participant, using both our statistical model and the Bayesian model, by minimizing the mean squared error loss (MSE) between the empirical curve and the model, with ϵ and σ as free parameters (**Fig. 6 C**), respectively (for the Bayesian model, we used the marginal distribution of the stimuli p_m as the prior). Across participants in both distributions, our statistical model yielded a better fit of the performance, relative to the Bayesian

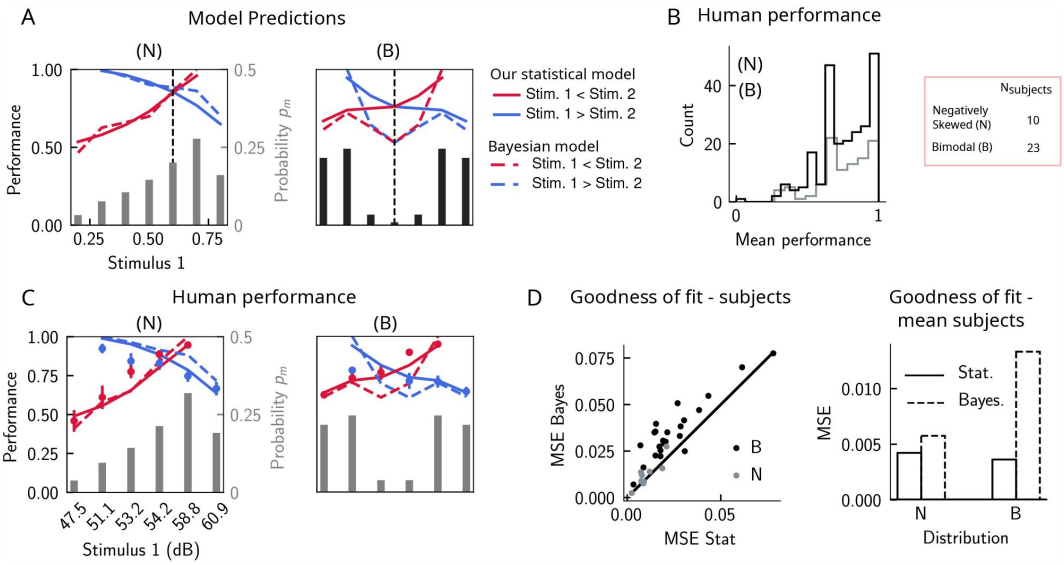

Figure 6

The stimulus distribution impacts the pattern of contraction bias through its cumulative.

(A) Left panel: prediction of performance (left y-axis) of our statistical model (solid lines) and the Bayesian model (dashed lines) for a negatively skewed stimulus distribution (gray bars, to be read with the right y-axis). Blue (red): performance as a function of s_1 for pairs of stimuli where $s_1 > s_2$ ($s_1 < s_2$). Vertical dashed line: median of distribution. Right: same as left, but for a bimodal distribution. **(B)** The distribution of performance across different stimuli pairs and subjects for the negatively skewed (gray) and the bimodal distribution (black). On average, across both distributions, participants performed with an accuracy of 75%. **(C)** Left: mean performance of human subjects on the negatively skewed distribution (dots, error-bars correspond to the standard deviation across different participants). Solid (dashed) lines correspond to fits of the mean performance of subjects with the statistical (Bayesian) model, $\epsilon = 0.55$ ($\sigma = 0.38$). Red (blue): performance as a function of s_1 for pairs of stimuli where $s_1 < s_2$ ($s_1 > s_2$), to be read with the left y-axis. The marginal stimulus distribution is shown in gray bars, to be read with the right y-axis. Right: same as left panel, but for the bimodal distribution. Here $\epsilon = 0.54$ ($\sigma = 0.73$). **(D)** Left: goodness of fit, as expressed by the mean-squared-error (MSE) between the empirical curve and the fitted curve (statistical model in the x-axis and the Bayesian model in the y-axis), computed individually for each participant and each distribution. Right: goodness of fit, computed for the average performance over participants in each distribution.

model (**Fig. 6 D**, left panel). We further fit the mean performance across all participants within a given distribution group, and similarly found that the statistical model yields a better fit, using the MSE as a goodness of fit metric (**Fig. 6 D**, right panel).

Finally, in order to better understand the parameters that affect the occurrence of errors in human participants, we computed the performance and fraction classified as $s_1 < s_2$ separately for different delay intervals. We found that the larger the delay interval, the lower the average performance (**Fig. 7 A**), accompanied by a larger contraction bias for larger intervals (**Fig. 7 B**). We further analyzed the fraction of trials in which subjects responded $s_1 < s_2$, conditioned on the specific pair of stimuli presented in the current and the previous trial (**Fig. 7 C**) for all distributions (one negatively skewed, and two bimodal distributions, of which only one is shown in **Fig. 6 C**). Compatible with the previous results [7], we found attractive history effects that increased with the delay interval (**Fig. 7 D** and **E**).

1.6.2 A prolonged inter-trial interval improves average performance and reduces attractive bias

If errors are due to the persistence of activity resulting from previous trials, what then, is the effect of the inter-trial interval (ITI)? In our model, a shorter ITI (relative to the default value of 6*s* used in **Figs. 2** and **3**) results in a worse performance and vice versa (**Fig. 8 A, B, C**). This change in performance is reflected in reduced biases toward the previous trial (**Fig. 8 D** and **E**). A prolonged ITI allows for a drifting bump to vanish due to the effect of adaptation: as a result, the performance improves with increasing ITI and conversely, worsens with a shorter ITI.

Do human subjects express less bias with longer ITIs, as predicted by our model? In our simulations, we set the ITI to either 2.2, 6 or 11 seconds, whereas in the experiment, since it is self-paced, the ITI can vary considerably. In order to emulate the simulation setting as closely as possible, we divided trials into two groups: “short” ITIs (shorter than 3 seconds), and “long” ITIs (longer than 3 seconds). This choice was motivated by the shape of the distribution of ITIs, which is bimodal, with a peak around 1 second, and another after 3 seconds (**Fig. 8 F**). Given the shape of the ITI distribution, we did not divide the ITIs into smaller intervals as this would result in too little data in some intervals. In line with our model, we found a better average performance with increasing ITI accompanied by decreasing contraction bias (**Fig. 8 G**). In order to quantify one-trial back effects, we used data pertaining to all of the distributions we tested – the negatively skewed, and also two bimodal distributions (of which only one was shown in this manuscript, in **Fig. 6 C**). This allowed us to obtain clear one-trial-back attractive biases, decreasing with increasing ITI (**Fig. 8 H**), in line with our model predictions (**Fig. 8 B** and **D**).

1.6.3 Working memory is attracted towards short-term and repelled from long-term sensory history

Although contraction bias is robustly found in different contexts, surprisingly similar tasks, such as perceptual estimation tasks, sometimes highlight opposite effects, i.e. repulsive effects [43–45]. Interestingly, recent studies have found both effects in the same experiment: in a study of visual orientation estimation [18], it has been found that attraction and repulsion have different timescales; while perceptual decisions about orientation are attracted towards recently perceived stimuli (timescale of a few seconds), they are repelled from stimuli that are shown further back in time (timescale of a few minutes). Moreover, in the same study, they find that the long-term repulsive bias is spatially specific, in line with sensory adaptation [46–48] and in contrast to short-term attractive serial dependence [18]. Given that adaptation is a main feature of our model of the PPC, we sought to determine whether such repulsive effects can emerge from the model. We extended the calculation of the bias to up to ten trials back, and quantified the slope of the bias as a function of the previous trial stimulus pair. We observe robust repulsive effects appear after the third trial back in history, and up to six trials back (**Fig. 8 I**). In our model, both short-term attractive effects and longer-term repulsive effects can be attributed to the multiple

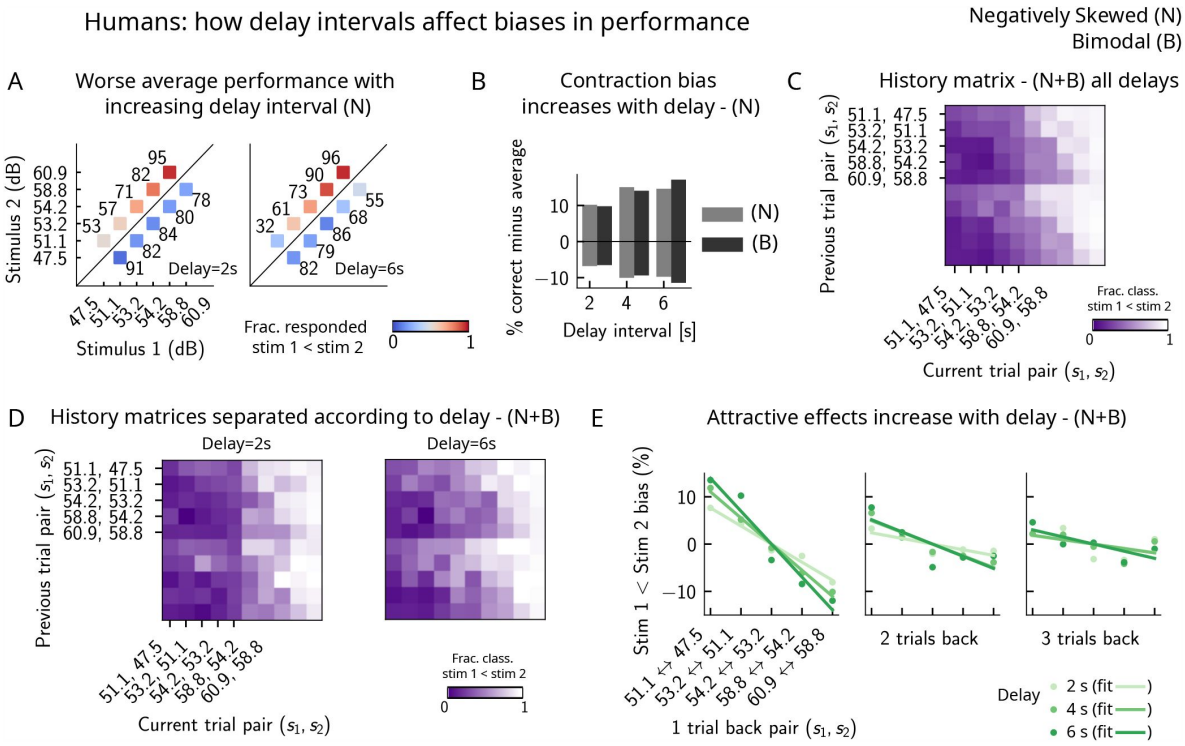

Figure 7

Attractive effects of the previous trials lead to contraction bias in human subjects, both increasing with delay interval.

(A) The performance (in percentage correct, shown in numbers above each stimulus pair) of human subjects is better with lower delay intervals (left, 2 seconds), than with higher delay intervals (right, 6 seconds). Colorbar expresses the fraction of trials in which participants responded that $s_1 < s_2$. Results are for the negatively skewed stimulus distribution, noted (N). **(B)** Concurrently, contraction bias on bias+ and bias-trials (quantification explained in text) also increases with an increased delay interval, for both stimulus distributions (negatively skewed in gray and bimodal in black). **(C)** History matrix, expressing the fraction of trials in which subjects responded $s_1 < s_2$ (in color) for every pair of current (x-axis) and previous (y-axis) stimuli, for negatively skewed and bimodal stimulus distributions (N+B). The one-trial back history effects can be seen through the vertical modulation of the color. Colorbar codes for the fraction of trials in which subjects responded $s_1 < s_2$. **(D)** History matrices (as in (C)), computed for all distributions, and separated according to delay intervals (left: 2 seconds and right: 6 seconds). **(E)** Bias, quantifying the (attractive) effect of previous stimulus pairs, for 1– 3 trials back in history. The attractive bias, computed for all distributions, increases with the delay interval separating the two stimuli (light to dark green: increasing delay).

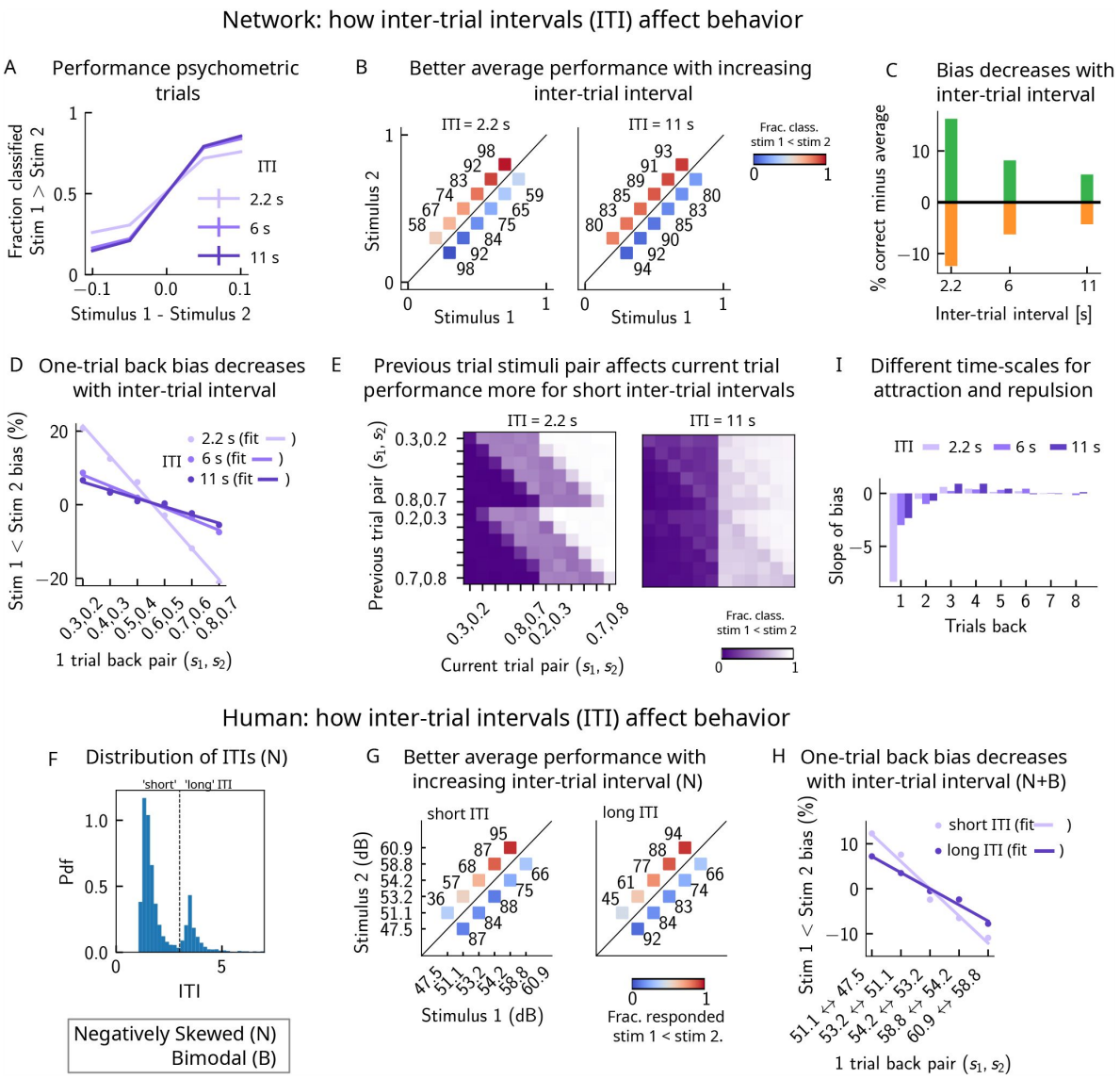

Figure 8

A prolonged inter-trial interval (ITI) improves average performance and reduces attractive biases. Working memory is attracted towards short-term and repelled from long-term sensory history.

(A) Performance of the network model for the psychometric stimuli improves with an increasing inter-trial interval. Errorbars (not visible) correspond to the s.e.m. over different simulations. **(B)** The network performance (numbers next to stimulus pairs) is on average better for longer ITIs (right panel, ITI=11*s*), compared to shorter ones (left panel, ITI=2.2*s*). Colorbar indicates the fraction of trials classified as $s_1 < s_2$. **(C)** Quantifying contraction bias separately for bias+ trials (green) and bias- trials (orange) yields a decreasing bias as the inter-trial interval increases. **(D)** The bias, quantifying the (attractive) effect of the previous trial, decreases with ITI. Darker shades of purple correspond to increasing values of the ITI, with dots corresponding to simulation values and lines to linear fits. **(E)** Performance is modulated by the previous stimulus pairs (modulation along the y-axis), more for a short ITI (left, ITI=2.2*s*) than for a longer ITI (right, ITI=11*s*). The colorbar corresponds to the fraction classified $s_1 < s_2$. **(F)** The distribution of ITIs in the human experiment is bimodal. We define as having a “short” ITI, those trials where the preceding ITI is shorter than 3 seconds and conversely for “long” ITI. **(G)** The human performance for the negatively skewed stimulus distribution is on average worse for shorter ITIs (left panel), compared to longer ones (right panel). Colorbar indicates the fraction of trials subjects responded $s_1 < s_2$. **(H)** The bias, quantifying the (attractive) effect of the previous trial, increases with ITI in human subjects. Darker shades of purple correspond to increasing values of the ITI, with dots corresponding to empirical values and lines to linear fits. **(I)** Although the stimuli shown up to two trials back yield attractive effects, those further back in history yield repulsive effects, notably when the ITI is larger. Such repulsive effects extend to up to 6 trials back.

timescales over which the networks operate. The short-term attractive effects occur due to the long time it takes for the adaptive threshold to build up in the PPC, and the short timescale with which the WM network integrates input from the PPC. The longer-term repulsive effects occur when the activity bump in the PPC persists in one location and causes adaptation to slowly build up, effectively increasing the activation threshold. The raised threshold takes equally long to return to baseline, preventing activity bumps to form in that location and thereby creating repulsion toward all the other locations in the network. Crucially, however, the amplitude of such effects depend on the inter-trial interval; in particular, for shorter inter-trial intervals, the repulsive effects are less observable.

1.7 The timescale of adaptation in the PPC network can control perceptual biases similar to those observed in dyslexia and autism

In a recent study [16], a similar PWM task with auditory stimuli was studied in human neurotypical (NT), autistic spectrum (ASD) and dyslexic (DYS) subjects. Based on an analysis using a generalized linear model (GLM), a double dissociation between different subject groups was suggested: ASD subjects exhibit a stronger bias towards long-term statistics – compared to NT subjects –, while for DYS subjects, a higher bias is present towards short-term statistics.

We investigated our model to see if it is able to show similar phenomenology, and if so, what are the relevant parameters controlling the timescale of the biases in behavior. We identified the adaptation timescale in the PPC as the parameter that affects the extent of the short-term bias, consistent with previous literature [49], [50]. Calculating the mean bias towards the previous trial stimulus pair (**Fig. 9 A**), we find that a shorter-than-NT adaptation timescale yields a larger bias towards the previous trial stimulus. Indeed, a shorter timescale for neuronal adaptation implies a faster process for the extinction of the bump in PPC – and the formation of a new bump that remains stable for a few trials – producing "jumpier" dynamics that lead to a larger number of one-trial back errors. In contrast, increasing this timescale with respect to NT gives rise to a stable bump for a longer time, ultimately yielding a smaller short-term bias. This can be seen in the detailed breakdown of the network's behavior on the current trial, when conditioned on the stimuli presented at the previous trial (**Fig. 9 B**, see also *Sect. 1.3* for a more detailed explanation of the dynamics). We performed a GLM analysis as in Ref. [16] to the network behavior, with stimuli from four trials back and the mean stimulus as regressors (see *Sect. 4.4*). This analysis shows that a reduction in the PPC adaptation timescale with respect to NT, produces behavioral changes qualitatively compatible with data from DYS subjects; on the contrary, an increase of this timescale yields results consistent with ASD data (**Fig. 9 C**).

This GLM analysis suggests that dissociable short- and long-term biases may be present in the network behavior. Having access to the full dynamics of the network, we sought to determine how it translates into such dissociable short- and long-term biases. Given that all the behavior arises from the location of the bump on the attractor, we quantified the fraction of trials in which the bump in the WM network, before the onset of the second stimulus, was present in the vicinity of any of the previous trial's stimuli (**Fig. S6 B**, right panel, and C), as well as the vicinity of the mean over the sensory history (**Fig. S6 B**, left panel, and C). While the bump location correlated well with the GLM weights corresponding to the previous trial's stimuli regressor (comparing the right panels of **Fig. S6 A** and **B**), surprisingly, it did not correlate with the GLM weights corresponding to the mean stimulus regressor (comparing the left panels of **Fig. S6 A** and **B**). In fact, we found that the bump was in a location given by the stimuli of the past two trials, as well as the mean over the stimulus history, in a smaller fraction of trials, as the adaptation timescale parameter was made larger (**Fig. S6 C**).

Given that the weights, after four trials in the past, were still non-zero, we extended the GLM regression by including a larger number of past stimuli as regressors. We found that doing this greatly reduced the weight of the mean stimulus regressor (**Fig. 9 C, D** and **E**, see *Sect. 4.4*

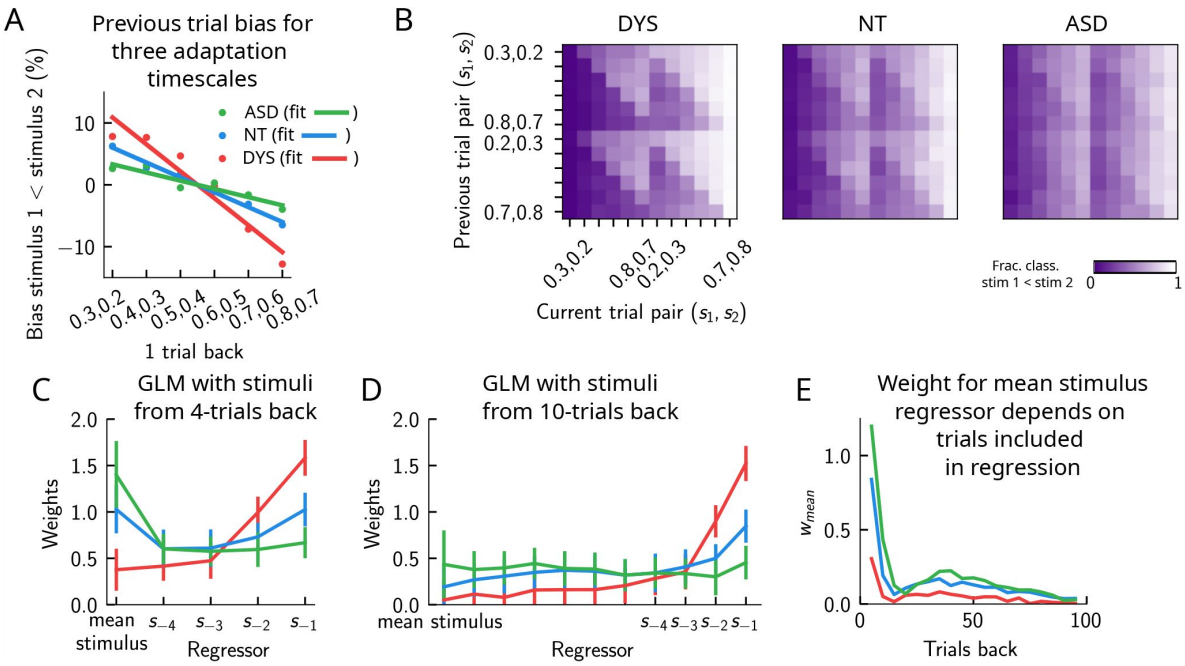

Figure 9

Apparent trade-off between short- and long-term biases, controlled by the timescale of neural adaptation.

(A) the bias exerted on the current trial by the previous trial (see main text for how it is computed), for three values of the adaptation timescale that mimic similar behavior to the three cohorts of subjects. **(B)** As in (**Fig. 2 D**), for three different values of adaptation timescale. The colorbar corresponds to the fraction classified $s_1 < s_2$. **(C)** GLM weights corresponding to the three values of the adaptation parameter marked in (**Fig. S6 A**), including up to 4 trials back. In a GLM variant incorporating a small number of past trials as regressors, the model yields a high weight for the running mean stimulus regressor. Errorbars correspond to the standard deviation across different simulations. **(D)** Same as in C, but including regressors corresponding to the past 10 trials as well as the running mean stimulus. With a larger number of regressors extending into the past, the model yields a small weight for the running mean stimulus regressor. Errorbars correspond to the standard deviation across different simulations. **(E)** The weight of the running mean stimulus regressor as a function of extending the number of past trial regressors decays upon increasing the number of previous-trial stimulus regressors.

and **E**, see *Sect. 4.4* and *4.4* for more details). Therefore, we propose an alternative interpretation of the GLM results given in Ref. [16]. In our model, the increased (reduced) weight for long-term mean in the ASD (DYS) subjects can be explained as an effect of a larger (smaller) window in time of short-term biases, without invoking a double dissociation mechanism (**Fig. 9 D** and **E**). In *Sect. 4.4*, we provide a mathematical argument for this, which is empirically shown by including a large number of individual stimuli from previous trials in the regression analysis.

2 Discussion

2.1 Contraction bias in the delayed comparison task: simply a statistical effect or more?

Contraction bias is an effect emerging in working memory tasks, where in the averaged behavior of a subject, the magnitude of the item held in memory appears to be larger than it actually is when it is "small" and, vice-versa, it appears to be smaller when it is "large" [51, 3, 4, 52, 19, 53, 54]. Recently, Akrami et al [7] have found that contraction bias as well as short-term history-dependent effects occur in an auditory delayed comparison task in rats and humans: the comparison performance in a given trial, depends on the stimuli shown in preceding trials (up to three trials back) [7], similar to previous findings in human 2AFC paradigms [5]. These findings raise the question: does contraction bias occur independently of short-term history effects, or does it emerge as a result of the latter?

Akrami et al [7] have also found the PPC to be a critical node for the generation of such effects, as its optogenetic inactivation (specifically during the delay interval) greatly attenuated both effects. WM was found to remain intact, suggesting that its content was perhaps read-out in another region. Electrophysiological recordings as well as optogenetic inactivation results in the same study suggest that while sensory history information is provided by the PPC, its integration with the WM content must happen somewhere downstream to the PPC. Different brain areas can fit the profile. For instance there are known projections from the PPC to mPFC in rats [55], where neural correlates of parametric working memory have been found [40]. Building on these findings, we suggest a minimal two-module model aimed at better understanding the interaction between contraction bias and short-term history effects. These two modules capture properties of the PPC (in providing sensory history signals) and a downstream network holding working memory content. Our WM and PPC networks, despite having different timescales, are both shown to encode information about the marginal distribution of the stimuli (**Fig. 4 A**). Although they have similar activity distributions to that of the external stimuli, they have different memory properties, due to the different timescales with which they process incoming stimuli. The putative WM network, from which information to solve the task is read-out, receives additional input from the PPC network. The PPC is modelled as integrating inputs slower relative to the WM network, and is also endowed with firing rate adaptation, the dynamics of which yield short-term history biases and consequently, contraction bias.

It must be noted, however, that short-term history effects (due to firing rate adaptation) do not necessarily need to be invoked in order to recover contraction bias: as long as errors are made following random samples from a distribution in the same range as that of the stimuli, contraction bias should be observed [56]. Indeed, when we manipulated the parameters of the PPC network in such a way that short-term history effects were eliminated (by removing the firing-rate adaptation), contraction bias persisted. As a result, our model suggests that contraction bias may not simply be given by a regression towards the mean of the stimuli during the inter-stimulus interval [57, 58], but brought about by a richer dynamics occurring at the level of individual trials [2], more in line with the idea of random sampling [59].

The model makes predictions as to how the pattern of errors may change when the distribution of stimuli is manipulated, either at the level of the presented stimuli or through the network dynamics. When we tested these predictions experimentally, by manipulating the skewness of the stimulus distribution, such that the median and the mean were dissociated (**Fig. 6 A**) the results from our human psychophysics experiments were in agreement with the model predictions. In further support of this, in a recent tactile categorization study [45], where rats were trained to categorize tactile stimuli according to a boundary set by the experimenter, the authors have shown that rats set their decision boundary according to the statistical structure of the stimulus set to which they are exposed. More studies are needed to fully verify the extent to which the statistical structure of the stimuli affect the performance. Finally, we note that in our model, the stimulus distribution is not explicitly learned (but see [60]): instead, the PPC dynamics follows the input, and its marginal distribution of activity is similar to that of the external input. This is in agreement with Ref. [45], where the authors used different stimulus ranges across different sessions and noted that rats initiated each session without any residual influence of the previous session's range/boundary on the current session, ruling out long-term learning of the input structure.

Importantly, our results are not limited to the delayed "comparison" paradigm, where binary decision making occurs. We show that by analyzing the location of the WM bump at the end of the delay interval, similar to the continuous recall tasks, we can retrieve the averaged effects of contraction bias, similar to previous reports [42]. Such continuous readout of the memory reveals a rich dynamics of errors at the level of individual trials, similar to the delayed comparison case, but to our knowledge this has not been studied in previous experimental studies. Papadimitriou et al [15] have characterised residual error distribution, in an orientation recall task, when limiting previous trials to orientations in the range of +35 to +85 degrees relative to the current trial. This distribution is unimodal, leading the authors to conclude that the current trial shows a small but systematic bias toward the location of the memorandum of the previous trial. It remains to be tested whether the error distribution remains unimodal, if conditioned on other values of the current and previous orientations, similar to our analysis in **Fig. 5 C**.

2.2 Attractor mechanism riding on multiple timescales

Our model assumes that the stimulus is held in working memory through the persistent activity of neurons, building on the discovery of persistent selective activity in a number of cortical areas, including the prefrontal cortex (PFC), during the delay interval [61–67]. To explain this finding, we have used the attractor framework, in which recurrently connected neurons mutually excite one another to form reverberation of activity within populations of neurons coding for a given stimulus [68–70]. However, subsequent work has shown that persistent activity related to the stimulus is not always present during the delay period and that the activity of neurons displays far more heterogeneity than previously thought [71]. It has been proposed that short-term synaptic facilitation may dynamically operate to bring a WM network across a phase transition from a silent to a persistently active state [72, 73]. Such mechanisms may further contribute to short-term biases [74], an alternative possibility that we have not specifically considered in this model.

An important model feature that is crucial in giving rise to all of its behavioral effects is its operation over multiple timescales (**Fig. S2 F**). Such timescales have been found to govern the processing of information in different areas of the cortex [30–32], and may reflect the heterogeneity of connections across different cortical areas [75].

2.3 Relation to other models

In many early studies, groups of neurons whose activity correlates monotonically with the stimulus feature, known as "plus" and "minus" neurons, have been found in the PFC [65, 76]. Such neurons have been used as the starting point in the construction of many models [77,

78, 71, 79]. It is important, however, to note that depending on the area, the fraction of such neurons can be small [40], and that the majority of neurons exhibit firing profiles that vary largely during the delay period [80]. Such heterogeneity of the PFC neurons' temporal firing profiles have prompted the successful construction of models that have not included the basic assumption of plus and minus neurons, but these have largely focused on the plausibility of the dynamics of neurons observed, with little connection to behavior [71].

A separate line of research has addressed behavior, by focusing on normative models to account for contraction bias [19, 5, 59, 81]. The abstract mathematical model that we present (**Fig. 4**), can be compatible with a Bayesian framework [19] in the limit of a very broad likelihood for the first stimulus and a very narrow one for the second stimulus, and where the prior for the first stimulus is replaced by the distribution of $\hat{s}$, following the model in **Fig. 4 B** (see *Sect. 4.2* for details). However, it is important to note that our model is conceptually different, i.e. subjects do not have access to the full prior distribution, but only to *samples* of the prior. We show that having full knowledge of the underlying sensory distribution is not needed to present contraction bias effects. Instead, a point estimate of past events that is updated trial to trial suffices to show similar results. This suggests a possible mechanism for the brain to approximate Bayesian inference and it remains open whether similar mechanisms (based on interaction of networks with different integration timescales) can approximate other Bayesian computations. It is also important to note the differences between the predictions from the two models. As shown in **Figs. 6 A** and **S4**, depending on the specific sensory distributions, the two models can have qualitatively different testable predictions. Data from our human psychophysical experiments, utilizing auditory Parametric Working Memory, show better agreement with our model predictions as compared to the Bayesian model.

Moreover, an ideal Bayesian observer model alone cannot capture the temporal pattern of short-term attraction and long-term repulsion observed in some tasks, and the model has had to be supplemented with efficient encoding and Bayesian decoding of information in order to capture both effects [18]. In our model, both effects emerge naturally as a result of neuronal adaptation, but their amplitudes crucially depend on the time parameters of the task, perhaps explaining the sometimes contradictory effects reported across different tasks.

Finally, while such attractive and repulsive effects in performance may be suboptimal in the context of a task designed in a laboratory setting, this may not be the case in more natural environments. For example, it has been suggested that integrating information over time serves to preserve perceptual continuity in the presence of noisy and discontinuous inputs [6]. This continuity of perception may be necessary to solve more complex tasks or make decisions, particularly in a non-stationary environment, or in a noisy environment.

3 Methods

3.1 The model

Our model is composed of two populations of N neurons, representing the PPC network and the putative WM network. We consider that each population is organized as a continuous line attractor, with recurrent connectivity described by an interaction matrix J_{ij}, whose entries represent the strength of the interaction between neuron i and j. The activation function of the neurons is a logistic function, i.e. the output r_i of neuron i, given the input h_i, is

$$r_i = \frac{1}{1 + e^{-\beta h_i}} \tag{1}$$

where β is the neuronal gain. The variables r_i take continuous values between 0 and 1, and represent the firing rates of the neurons. The input h_i to a neuron is given by

$$\tau \frac{dh_i}{dt} + h_i = \sum_{j(\neq i)} J_{ij} r_j + I_i^{\text{ext}} \tag{2}$$

where τ is the timescale for the integration of inputs. In the first term on the right hand side, $J_{ij}r_j$ represents the input to neuron i from neuron j, and I_i^{ext} corresponds to the external inputs. The recurrent connections are given by

$$J_{ij} = \frac{1}{d_0}(K_{ij} - J_0)\,, \tag{3}$$

with

$$K_{ij} = J_e\, e^{-\frac{|x_i - x_j|}{d_0}}\,. \tag{4}$$

The interaction kernel, K, is assumed to be the result of a time-averaged Hebbian plasticity rule: neurons with nearby firing fields will fire concurrently and strengthen their connections, while firing fields far apart will produce weak interactions [82]. Neuron i is associated with the firing field $x_i = i/N$. The form of K expresses a connectivity between neurons i and j that is exponentially decreasing with the distance between their respective firing fields, proportional to $|i - j|$; the exponential rate of decrease is set by the constant d_0, i.e. the typical range of interaction. The amplitude of the kernel is also rescaled by d_0, in such a way that $\sum_{i,j} K_{ij}$ is constant. The strength of the excitatory weights is set by J_e; the normalization of K, together with the sigmoid activation function saturating to 1, implies that J_e is also the maximum possible input received by any neuron due to the recurrent connections. The constant J_0, instead, contributes to a linear global inhibition term. Its value needs to be chosen depending on J_e and d_0, so that the balance between excitatory and inhibitory inputs ensures that the activity remains localized along the attractor, i.e. it does not either vanish or equal 1 everywhere; together, these three constants set the width of the bump of activity.

The two networks in our model are coupled through excitatory connections from the PPC to the WM network. Therefore, we introduce two equations analogous to *Eq. (2)*, one for each network. The coupling between the two will enter as a firing-rate dependent input, in addition to I^{ext}. The dynamics of the input to a neuron in the WM network writes

$$\tau_h^W \frac{dh_i^W}{dt} + h_i^W = \sum_{j \in j^W} J_{ij} r_j^W + J^{P \to W} r_i^P + I_i^{\text{ext}}\,, \tag{5}$$

where j^W indexes neurons in the WM network, and τ_h^W is the timescale for the integration of inputs in the WM network. The first term in the r.h.s corresponds to inputs from recurrent connections within the WM network. The second term, corresponds to inputs from the PPC network. Finally, the last term corresponds to the external inputs used to give stimuli to the network. Similarly, for the PPC network we have

$$\tau_h^P \frac{dh_i^P}{dt} + h_i^P = \sum_{j \in j^P} J_{ij} r_j^P - \theta_i^P + I_i^{\text{ext}}\,, \tag{6}$$

where j^P indexes neurons in the PPC, and where τ^P is the timescale for the integration of inputs in the PPC network; importantly, we set this to be longer than the analogous quantity for the WM network, $\tau_h^W < \tau_h^P$ (see **Tab. 1**). The first and third terms in the r.h.s are analogous to the corresponding ones for the WM network: inputs from within the network and from the stimuli. The second term instead, corresponds to adaptive thresholds with dynamics specified by

$$\tau_\theta^P \frac{d\theta_i^P}{dt} + \theta_i^P = D^P r_i^P \tag{7}$$

modelling neuronal adaptation, where τ_θ^P and D^P set its timescale and its amplitude. We are interested in the condition where the timescale of the evolution of the input current is much smaller relative to that of the adaptation $(\tau_h^P \ll \tau_\theta^P)$. For a constant τ_θ^P, we find that depending on the value of D^P, the bump of activity shows different behaviors. For low values of D^P, the bump remains relatively stable (**Fig. S1 C** (1)). Upon increasing D^P, the bump gradually starts to drift (**Fig. S1 C** (2-3)). Upon increasing D^P even further, a phase transition leads to an abrupt dissipation of the bump (**Fig. S1 C** (4)).

Note that, while the transition from bump stability to drift occurs gradually, the transition from drift to dissipation is abrupt. This abruptness in the transition from the drift to the dissipation regime may imply that only one of the two behaviors is possible in our model of the PPC (*Sect. 1.3*). In fact, our network model of the PPC operates in the "drift" regime $(\tau_\theta^P = 7.5,\ D^P = 0.3)$. However, we also observe dissipation of the bump, which is mainly responsible for the jumps observed in the model. This occurs due to the inputs from incoming external stimuli, that affect the bump via the global inhibition in the model (**Fig. S1 A**). Therefore external stimuli can allow the network to temporarily cross the sharp drift/dissipation boundary shown in **Fig. S1 B**. As a result, the combined effect of adaptation, together with external inputs and global inhibition result in the drift/jump dynamics described in the main text.

Finally, both networks have a linear geometry with free boundary conditions, i.e. no condition is imposed on the profile activity at neuron 1 or N.

3.2 Simulation

We performed all the simulations using custom Python code. Differential equations were numerically integrated with a time step of dt = 0.001 using the forward Euler method. The activity of neurons in both circuits were initialized to r = 0. Each stimulus was presented for 400 ms. A stimulus is introduced as a "box" of unit amplitude and of width 2 δs around s in stimulus space: in a network with N neurons, the stimulus is given by setting $I_i^{\text{ext}} = 1$ in *Eq. 5* for neurons with index i within $(s \pm \delta s) \times N$, and $I_i^{\text{ext}} = 0$ for all the others. Only the activity in the WM network was used to assess performance. To do that, the activity vector was recorded at two time-points: 200 ms before and after the onset of the second stimulus s_2. Then, the neurons with the maximal activity were identified at both time-points, and compared to make a decision. This procedure was done for 50 different simulations with 1000 consecutive trials in each, with a fixed inter-trial interval separating two consecutive trials, fixed to 5 seconds. The inter-stimulus intervals were set according to two different experimental designs, as explained below.

3.2.1 Interleaved design

As in the study in Ref. [7], an inter-stimulus interval of either 2, 6 or 10 seconds was randomly selected. The delay interval is defined as the time elapsed from the end of the first stimulus to the beginning of the second stimulus. This procedure was used to produce **Figs. 1**, **2**, **3**, **7**, **S2**, **S3**.

3.2.2 Block design

In order to provide a comparison to the interleaved design, but also to simulate the design in Ref. [16], we also ran simulations with a block design, where the inter-stimulus intervals were kept fixed throughout the trials. Other than this, the procedure and parameters used were exactly the same as in the interleaved case. This procedure was used to produce **Figs. 9** and **S6**.

3.3 Human auditory experiment - delayed comparison task

Subjects received, in each trial, a pair of sounds played from ear-surrounding headphones. The subject self-initiated each trial by pressing the space bar on the keyboard. The first sound was then presented together with a blue square on the left side of a computer monitor in front of the subject. This was followed by a delay period, indicated by 'WAIT!' on the screen, then the second sound was presented together with a red square on the right side of the screen. At the end of the second stimulus, subjects had 2 seconds to decide which one was louder, then indicate their choice by pressing the 's' key if they thought that the first sound was louder, or the 'l' key if they thought that the second sound was louder. Written feedback about the correctness of their response was provided on the screen for each individual trial. Every ten trials, participants received feedback on their running mean performance calculated up to that trial. Participants then had to press spacebar to go to the next trial (the experiment was hence self-paced).

The two auditory stimuli, s_1 and s_2, separated by a variable delay (of 2, 4 and 6 seconds), were played for 400 ms, with short delay periods of 250 ms inserted before s_1 and after s_2. The stimuli consisted of broadband noise 2000-20000 Hz, generated as a series of sound pressure level (SPL) values sampled from a zero-mean normal distribution. The overall mean intensity of sounds varied from 60-92 dB. Participants had to judge which out of the two stimuli, s_1 and s_2, was louder (had the greater SPL standard deviation).

We recruited 10 subjects for the negatively skewed distribution and 24 subjects for the bimodal distribution. The study was approved by the University College London (UCL) Research Ethics Committee [16159/001] (London, UK). Each participant performed approximately 400 trials for a given distribution. Several participants took part in both distributions.

4 Supplementary Material

4.1 Computing bump location

In order to check whether the bump is in a target location (**Figs. 3 B**, **S2 B**, and **S3 D**), we check whether the position of the neuron with the maximal firing rate is within a distance of ±5% of the length of the whole line attractor from the target location (**Figs. 3 A**, **S2 A** and **S3 C**). In these figures, we compare the probability that, in a given trial, the activity of the WM network is localized around one of the previous stimuli (estimated from the simulation of the dynamics, histograms) with the probability of this happening due to chance (horizontal dashed line). Here we detail the calculation of the chance probability. In general, if we have two discrete independent random variables, $\hat{X}$ and $\hat{Y}$, with probability distributions p_X and p_Y, the probability of them having the same value is

$$\mathrm{Prob}\{\hat{X}=\hat{Y}\}=\sum_{i,j}\underbrace{\mathrm{Prob}\{\hat{X}=x_i\}}_{p_X^i}\underbrace{\mathrm{Prob}\{\hat{Y}=y_j\}}_{p_Y^j}\mathbb{I}(x_i=y_j)$$

where i, j are the indices for different values of the two random variables and $\mathbb{I}(x_i = y_j)$ equals 1 where $x_i = x_j$ and 0 otherwise. If the two random variables are identically distributed, the above expression writes

$$\mathrm{Prob}\{\hat{X} = \hat{Y}\} = \sum_{i,j} p_X^i \, p_Y^j \, \delta_{i,j} = \sum_i \left(p_X^i\right)^2$$

In our case, the two identically distributed random variables are "bump location at the current trial" and the "target bump location" (that are s_1^{t-2}, s_2^{t-2}, s_1^{t-1}, s_2^{t-1} and $\langle s \rangle$). With the exception of the mean stimulus $\langle s \rangle$, all the other variables are identically distributed, with probability p_m (that is the marginal distribution over s_1 or s_2). We note that the bump location in the WM network follows a very similar distribution to p_m (**Fig. 4 A**). Then, we compute the chance probability with the above relationship, where $p_X \equiv p_m$. For the mean stimulus, instead, we have a probability which is simply equal to 1 for $s = 0.5$ and 0 elsewhere; therefore, the chance probability for the bump location to be at the mean stimulus, then is $p_m(0.5)$.

The excess probability (with respect to chance) for the bump location to equal one of the previous stimuli gives a measure of the correlation between these two; in other terms, of the amount of information retained by the network about previous stimuli.

4.2 The probability to make errors is proportional to the cumulative distribution of the stimuli, giving rise to contraction bias

In order to illustrate the statistical origin of contraction bias consistent with our network model, we consider a simplified mathematical model of its performance (**Fig. 4 B**). By definition of the delayed comparison task, the optimal decision maker produces a label y equal to 1 if $s_1^t < s_2^t$, and 0 if $s_1^t > s_2^t$; the impossible cases $s_1^t = s_2^t$ are excluded from the set of stimuli, but would produce a label which is either 0 or 1 with 50% probability. That is

$$y(s_1, s_2) = \begin{cases} 1 & \text{if } s_1 < s_2 \\ 0 & \text{if } s_1 > s_2 \\ \mathrm{Bernoulli}(1/2) & \text{if } s_1 = s_2 \end{cases} \tag{8}$$

In this simplified scheme, at each trial t, the two stimuli s_1^t and s_2^t are perfectly perceived with a finite probability $1-\epsilon$, with $\epsilon < 1$. Under the assumption that the decision maker behaves optimally based on the perceived stimuli, a correct perception would necessarily lead to the correct label. However, with probability ϵ, the first stimulus is randomly selected from a buffer of stimuli, i.e. is replaced by a random variable $\hat{s}_1$ that has a probability distribution p_m^t .

The probability distribution p_m^t is the statistics of previously shown stimuli. The information about the previous stimulus is given by the activity of the "slower" PPC network. As shown above, after the presentation of the first stimulus of the trial, the bump of activity is seen to jump to the position encoding one of the previously presented stimuli, s_2^{t-1}, s_1^{t-1}, s_2^{t-2} etc. with decreasing probability (**Fig. 3 C**). Therefore, in calculating the performance in the task, we can take p_m^t to be the marginal distribution of the stimulus s_1 or s_2 across trials, as in the histogram (**Fig. 4 A**).

The probability of a misclassification is then given by the probability that, given the pair (s_1^t, s_2^t), at trial t,

1. the first stimulus is replaced by a random value, which happens with probability ϵ, and

2. the value of $\hat{s}_1$ replaced is larger than s_2^t when s_1^t is smaller and viceversa (**Fig. 4 C**). In summary, the probability of an error at trial *t* is given by

$$\text{Prob}\left\{\text{error} \;\middle|\; s_1^t = s_1,\, s_2^t = s_2\right\} = \epsilon \cdot \begin{cases} p_m^t(s_2)/2 + \sum_{s<s_2} p_m^t(s) & \text{if } s_1 > s_2\ , \\ p_m^t(s_2)/2 + \sum_{s>s_2} p_m^t(s) & \text{if } s_1 < s_2\ . \end{cases} \tag{9}$$

4.3 Bayesian description of contraction bias

We reproduce here the theoretical result from [41], which provides a normative model for contraction bias in the Bayesian inference framework, and apply it to the different stimulus distributions described in *Sect. 1.6.1*.

A stimulus with value *s* is encoded by the agent through a noisy representation $\hat{r} \sim \ell(\cdot\,|\,s)$. Before the presentation of the stimulus, the agent has an expectation of its possible values which is described by the probability π. Assuming that it has access to the internal representation *r*, as well as the probability distributions ℓ and π, the agent can infer the perceived stimulus$\hat{s}$ through Bayes rule:

$$p(\hat{s} = s|r) = \frac{\ell(r|s)\,\pi(s)}{p(r)} \tag{10}$$

where $p(r) = \int ds'\, \ell(r\,|\,s')\,\pi(s')$. In this Bayesian setting, the probability distributions for the noisy representation and for the expected measurement are interpreted as the *likelihood* and the *prior*, respectively.

In the delayed comparison task, at the time of the decision, the two stimuli s_1 and s_2 are assumed to be encoded independently, although with different uncertainties, due to the different delays leading to the time of decision: $\ell(r_1, r_2|\ s_1, s_2) = \ell_1(r_1|\ s_1)\ \ell_2(r_{2|}\ s_2)$, with var[$\ell_1$] > var[$\ell_2$]. Similarly, the expected values of the stimuli are assumed to be independent but also identically distributed: $\pi(s_1, s_2) = \pi(s_1)\ \pi(s_2)$.

The optimal Bayesian decision maker uses the inference of the stimuli through *Eq. (10)* to produce an estimate of the probability that $s_1 < s_2$, given the internal representations,

$$p(\hat{s}_1 < \hat{s}_2|r_1, r_2) = \iint ds_1'\, ds_2'\, \Theta(s_2' - s_1')\, p(s_1'|r_1)\, p(s_2'|r_2) \tag{11}$$

where Θ is the Heaviside function, and yields a label $\hat{y}$= 1 (truth value of "$s_1 < s_2$") when such probability is higher than 1/2, and $\hat{y}$ = 0 otherwise. Therefore, the probability that the Bayesian decision maker yields the response "$s_1 < s_2$" given the *true* values of the stimuli s_1 and s_2 is the average of the label $\hat{y}$ over the possible values of their representations, i.e. over the likelihood:

$$p(\hat{y} = 1|s_1, s_2) = \iint dr_1'\, dr_2'\ \ \Theta\Big(p(\hat{s}_1 < \hat{s}_2|r_1', r_2') - \frac{1}{2}\Big)\, \ell_1(r_1'|s_1)\, \ell_2(r_2'|s_2) \tag{12}$$

4.3.1 Application to our study

In modelling our data, we assume that the likelihood functions $\ell_1(\cdot|\ s_1)$ and $\ell_2(\cdot\,|\,s_2)$ are Gaussian with mean equal to the stimulus, but with different standard deviations, σ_1 and σ_2, respectively, as in [41]. We restrict to the particular case where σ_2 = 0, i.e. there is no uncertainty in the

representation of the second stimulus, since there is negligible delay between its presentation and the decision. We instead assume a finite standard deviation $\sigma_1 = \sigma$, which we use as the only free parameter of this model to produce **Figs.S4 A-D**, panels 2 and 4.

The prior π is chosen to be the marginal distribution of the first stimulus – identical to the marginal of the second stimulus, because of symmetry.

When $\sigma_2 = 0$, $\ell_2(r|s) = \delta(r - s)$ (Dirac delta), and the predicted response probability, *Eq. (12)*, reduces to

$$p(\hat{y} = 1|s_1, s_2) = \int dr_1' \ \Theta\Big(\int_{-\infty}^{s_2} ds_1' p(s_1'|r_1) - \frac{1}{2}\Big)\, \ell_1(r_1'|s_1)\,. \tag{13}$$

4.4 Generalized Linear Model (GLM)

GLM as in Lieder et al

Similarly to Ref. [16], we performed a multivariate logistic regression (an instance of generalized linear model, GLM) to the output of the network in the delayed discrimination task with recent stimuli values as covariates:

$$P(\text{“}s_1^t < s_2^t\text{”}) = \sigma\Big(\alpha\,(s_1^t - s_2^t) + \sum_{i=1}^{h} w_i\,(\overline{s^{t-i}} - s_1^t) + w_{mean}\,(\langle s\rangle - s_1^t)\Big) \tag{14}$$

where σ is the sigmoidal function $\sigma(z) = 1/(1+e^{-z})$, $\overline{s^\tau} = (s_1^\tau + s_2^\tau)/2$ is the mean of the stimuli presented at trial τ, h is the number of “history” terms in the regression, and $\langle s\rangle$ is the mean of the stimuli within and across trials up to the current one. As in Ref. [16], we choose h = 4, i.e. we include in the short-term history, the four trials prior to the current one. The first term in *Eq. (14)*, with weight α, controls the slope of the psychometric curve. The remaining terms, combined linearly with weights w, contribute to biases expressing the long and short-term memory. In Ref. [16], it is shown that subjects on the autistic syndrome (ASD) conserve the higher long-term weights, w_{mean}, while losing the short-term weights expressed by neurotypical (NT) subjects. In contrast, dyslexic (DYS) subjects conserve a higher bias from the recent stimuli, w_1, while losing the higher long-term weights, also expressed by neurotypical subjects.

In order to gain insight into this regression model in terms of our network, we also performed a linear regression of the bump of activity just before the onset of the second stimulus, denoted $\hat{s}_1^t$, versus the same variables:

$$\hat{s}_1^t = s_1^t + \sum_{i=1}^{h} w_i\,(\overline{s^{t-i}} - s_1^t) + w_{mean}\,(\langle s\rangle - s_1^t) \tag{15}$$

In this case, we see that the weights w in the linear regression for $\hat{s}_1^t$ have the same qualitative behavior as the weights for the bias term in the GLM regression for the performance (not shown). This is expected, since the decision-making rule in the network –based on the bump location just before and during the second stimulus, $\hat{s}_1$ and $\hat{s}_2 \simeq s_2^t$ respectively– is deterministic, following $P(\text{“}s_1^t < s_2^t\text{”}) = \Theta(s_2^t - \hat{s}_1^t)$. Therefore, the bias term in the GLM performed in Ref. [16], *Eq. (14)*, corresponds to the displacement of the bump location $\hat{s}_1^t$ with respect to the actual stimulus s_1^t, modelled to be linearly dependent on the displacement of previous stimuli from s_1^t.

Regression model with infinite history

In the regression formulas in *Eqs. (14)* and *(15)*, it is possible to give an interpretation of the parameter w_{mean}, that is the weight of the contribution from the covariate corresponding to the mean of the past stimuli. Let us consider two regression models, one in which, in addition to a regressor corresponding to the mean stimulus, regressors corresponding to the stimulus history are included up to trial h, and another in which $h = \infty$, i.e. infinitely many past stimuli are included as regressors. In this case, *Eq. (15)* rewrites

$$\hat{s}_1^t = s_1^t + \sum_{i=1}^{\infty} w_i \left(\overline{s^{t-i}} - s_1^t\right). \tag{16}$$

If we assume that the weights obtained from the regression have roughly an exponential dependence on time (**Fig. 9 C** and **D**), we can write

$$w_i = \gamma\, w_{i-1} = \gamma^i\, w_0\,. \tag{17}$$

By equating *Eqs. (15)* and *(16)*, we would find that

$$\begin{aligned} w_{mean}\left(\langle s\rangle - s_1^t\right) &= \sum_{i=h+1}^{\infty} w_i \left(\overline{s^{t-i}} - s_1^t\right) \\ &= w_{i+1} \sum_{j=0}^{\infty} \gamma^j \left(\overline{s^{t-(h+1+j)}} - s_1^t\right) \\ &= \frac{w_{h+1}}{1-\gamma}\left(\langle s\rangle_\gamma - s_1^t\right) \end{aligned} \tag{18}$$

where

$$\langle s\rangle_\gamma = \sum_{j=0}^{\infty} g_j\, \overline{s^{t-h-1-j}} \tag{19}$$

that is an average over the geometric distribution $g_j = (1-\gamma)\,\gamma^j$, from time $t-(h+1)$ backward. Since for γ large enough we have $\langle s\rangle_\gamma = \langle s\rangle$, we can identify

$$w_{mean} \propto \frac{w_{h+1}}{1-\gamma}\,. \tag{20}$$

This derivation indicates that the magnitude w_{mean} in the infinite history model, given by *Eq. (15)*, is a function of the discount factor γ as well as the weight of the first trial left out from the finite-history regression (w_{h+1}). A higher γ value, i.e. a longer timescale for damping of the weights extending into the stimulus history, yields a higher w_{mean}. We can obtain γ for each condition (NT, ASD and DYS) by fitting the weights obtained as a function of trials extending into the history (**Fig. 9 C** and **D**). As predicted by *Eq. (20)*, a larger window for short-term history effects (as in the ASD case relative to NT) yields a larger weight for the covariate corresponding to the mean stimulus. Finally, *Eq. (20)* also predicts that w_{mean} is proportional to w_{h+1}, the number of trials back we consider in the regression, h, implying that the number of covariates that we choose to include in the model may greatly affect the results. Both of these predictions are corroborated by plotting directly the value of w_{mean} obtained from the regression (**Fig. 9 E**).

4.5 Supplementary Figures

4.6 Parameters

Acknowledgements

We are grateful to Loreen Hertäg for helpful comments on our figures, and Arash Fassihi for helpful discussions. We also thank Guilhem Ibos for pointing out a typo in our figure legends in a previous version of this manuscript. This work was supported by BBSRC BB/N013956/1, BB/N019008/1, Wellcome Trust 200790/Z/16/Z, Simons Foundation 564408, EPSRC EP/R035806/1, Gatsby Charitable Foundation 562980 and Wellcome Trust 562763.

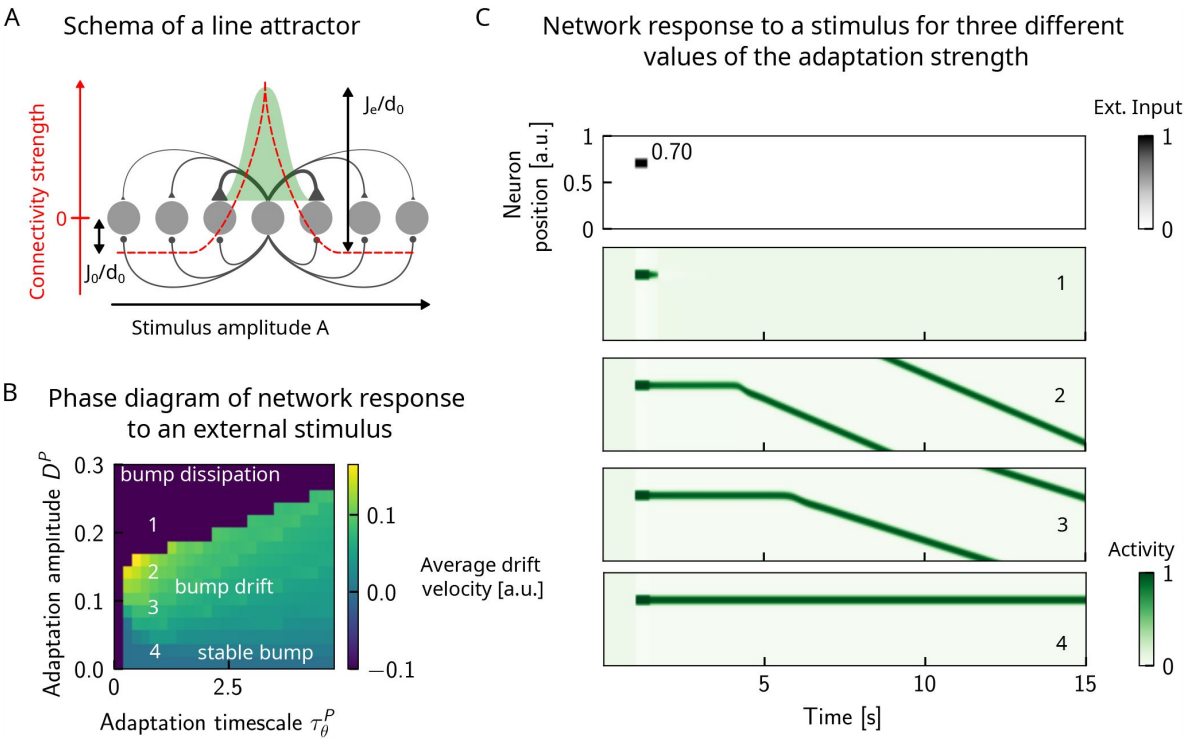

Figure S1

Dynamics of responses in a one-dimensional continuous attractor network, in the presence of adaptation.

(A) We study a one dimensional line attractor in which neurons code for a stimulus feature that varies along a physical dimension, such as amplitude of an auditory stimulus. The connections between pairs of neurons is a decreasing, symmetric function of the distance between their preferred firing locations, allowing for a bump of activity to form and self-sustain when sufficient input is given to the network. However, this self-sustaining activity may be disrupted if neuronal adaptation is present. In particular, drifting dynamics may be observed. **(B)** Left: phase diagram of the average drift velocity as a function of the adaptation timescale τ_θ^P and amplitude D^P. The average drift velocity is simply computed as the distance travelled by the center of the bump in a duration of 50 seconds. Color codes for the average drift velocity (a.u.). Numbers indicate four points for which sample dynamics are shown in (C). **(C)** We observe three main phases: in the first, the activity bump is stable when no or little neuronal adaptation is present (point 4). Larger values of neural adaptation induce drift of the activity bump; the average drift velocity increases upon increasing the neural adaptation (points 2 and 3). Finally, increasing it even further leads to the dissipation of the activity bump (point 1). The boundary between the drift and dissipation phases is abrupt. In these simulations, periodic boundary conditions have been used in order to compute the average average drift velocity over longer durations.

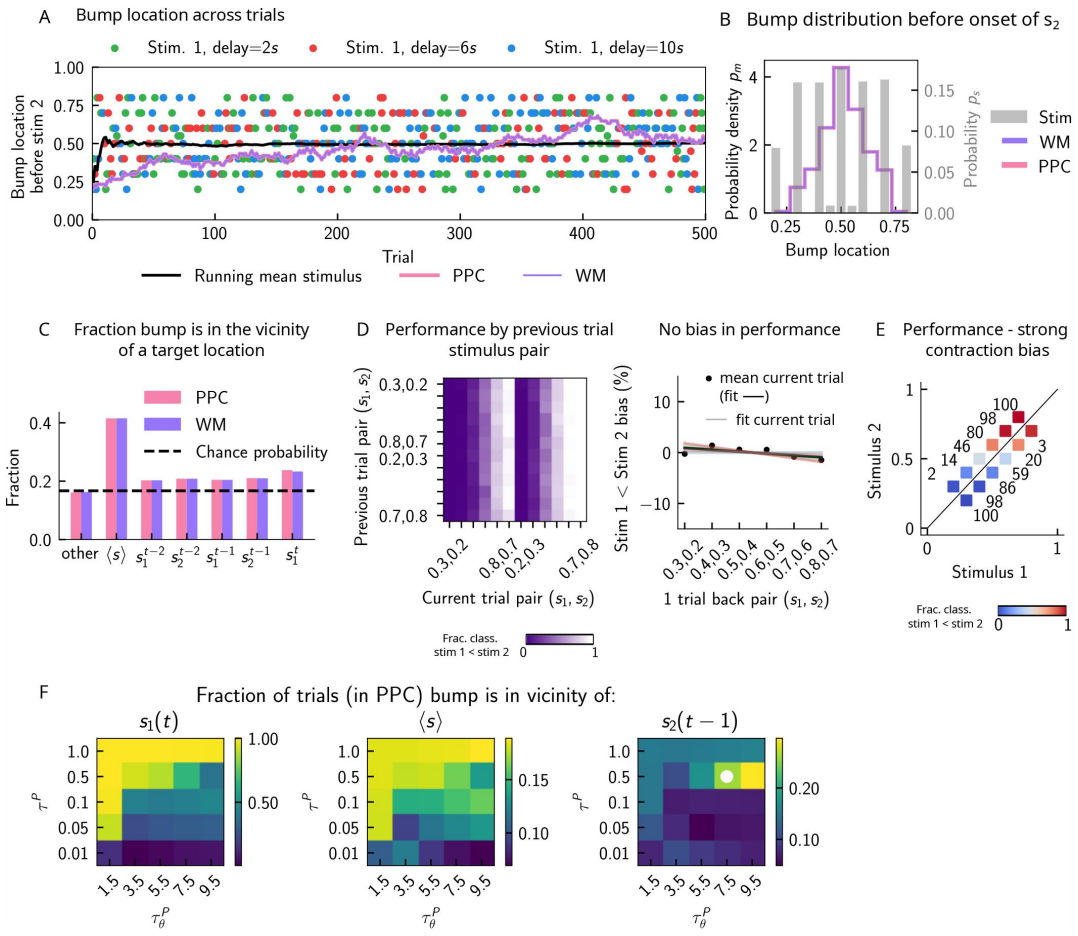

Figure S2

The role of neuronal adaptation in generating short-term history biases.

In order to better understand the network mechanisms that give rise to short-term history effects, we removed neural adaptation in the PPC network and assessed the performance in the WM network. **(A)** As in (**Fig. 3 B**). We track the location of the bump, in the PPC (pink), and in the WM network (purple) before the onset of the second stimulus (the pink curve cannot be seen as the purple curve goes perfectly on top). In this case, the displacement of the bump of activity is smooth and new sensory stimuli (colored dots) induce only a minimal shift in the location of the bump. This behavior is to be contrasted with the case in which there is adaptation in the PPC network, inducing jumps in the bump location (**Fig. 3 A**). An additional effect of no neural adaptation is that the activity in the PPC network, completely overrides the activity in the WM network. **(B)** As in (**Fig. 4 A**). Marginal distribution of the bump location in both networks (pink for PPC, purple for WM) before the onset of s_2 is more peaked than the marginal distribution of the stimuli (gray), as a result of the absence of "jumps". **(C)** As in (**Fig. 3 C**). We compute the fraction of times the bump is in a given location, current trial (s_1^t), four preceding trials $(s_1^{t-2} \text{ to } s_2^{t-1})$, the running mean stimulus, or all other locations (overlapping sets). In this case, in the majority of the trials, the bump is either at the running mean stimulus, or any other location. The fraction of trials in which it is in the position of the four previous stimuli roughly corresponds to chance occurrence (dashed black lines), with only a minor increase for the current stimulus. **(D)** As in (**Fig. 2 D**). Left: The network behavior conditioned on the previous trial stimulus pair does not exhibit any previous-trial attractive dependence (vertical modulation). Colorbar corresponds to the fraction classified as $s_1 < s_2$. Right: This attractive dependence can also be expressed through the bias measure (see main text for how it is computed). Colored lines correspond to current trial pairs, the black dots to the mean over all current trial pairs, and the black line to its linear fit. **(E)** As in (**Fig. 2 C**). Although there are no attractive previous trial effects, the performance expresses a very strong contraction bias, and performance is as if the decision boundary is orthogonal to the optimal decision boundary. Color codes for fraction of trials in which a $s_1 > s_2$ classification is made. **(F)** Phase diagram with τ_θ^P on the x-axis, τ^P on the y-axis, and in color, the fraction of trials in which the bump, before the network is stimulated with the second stimulus, is in the vicinity of a target (specified in the title of each panel). White dot corresponds to parameters of the default network **Tab. 1**.

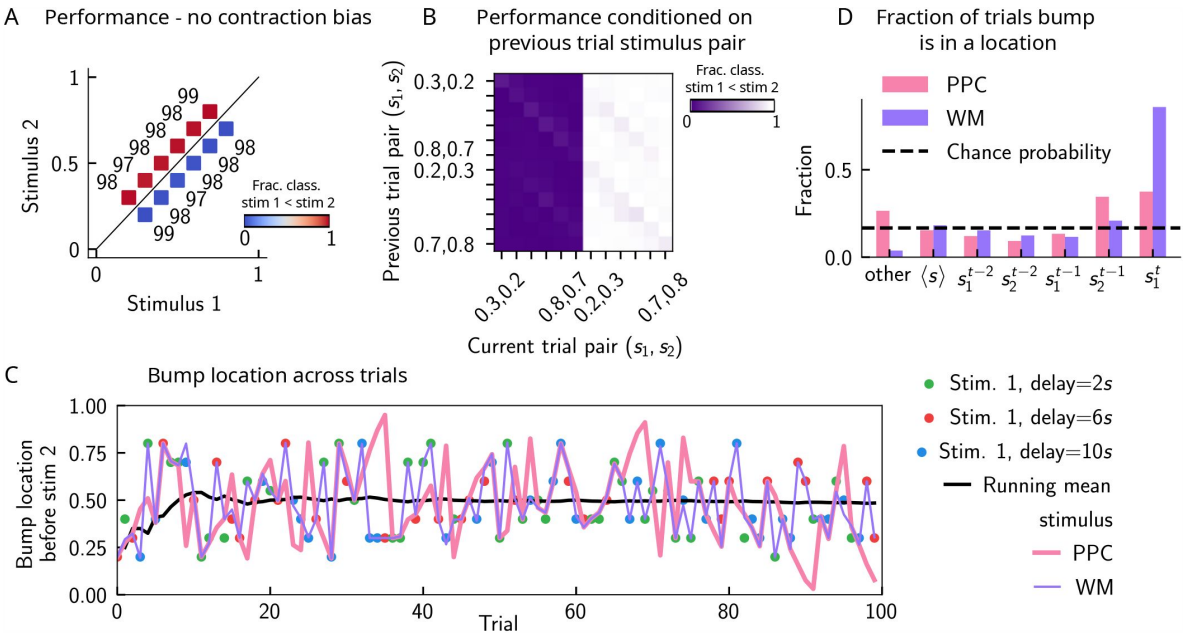

Figure S3

Inactivating the inputs from the PPC network improves performance, in line with experimental findings.

(A) As in (**Fig. 2 C**). The performance of the network when the strength of the inputs from the PPC to the WM network is weakened (modelling the optogenetic inactivation of the PPC) is dramatically improved, and contraction bias is virtually eliminated. The colorbar corresponds to the fraction classified as $s_1 < s_2$. **(B)** As in (**Fig. 2 D**). The performance for each stimulus pair in the current trial is improved and no modulation by the previous stimulus pairs can be observed. The colorbar corresponds to the fraction classified as $s_1 < s_2$. **(C)** As in (**Fig. 3 B**). This improvement of the performance can be traced back to how well the activity bump in the WM network (in purple), before the onset of the second stimulus s_2, tracks the first stimulus s_1 (shown in colored dots, each corresponding to a different value of the inter-stimulus delay interval). Relative to the case in which inputs from the PPC are intact (**Fig. 3 A**), it can be seen that the location of the bump tracks the first stimulus with high fidelity. The activity in the PPC (in pink), instead, is identical to that shown previously (**Fig. 6 A**), as all the other parameters are kept constant. **(D)** As in (**Fig. 2 C**). The bump location can be quantified not only for the stimulus s_1 of the current trial (colored dots, each color corresponding to a given delay interval), but for the four preceding stimuli from the two previous trials (from s_2^{t-1} back to s_1^{t-2}). With weaker inputs from the PPC (pink), the WM (purple) function of the circuit is disrupted less frequently, and in the majority of the trials, the bump of activity corresponds to the first stimulus s_1^t .

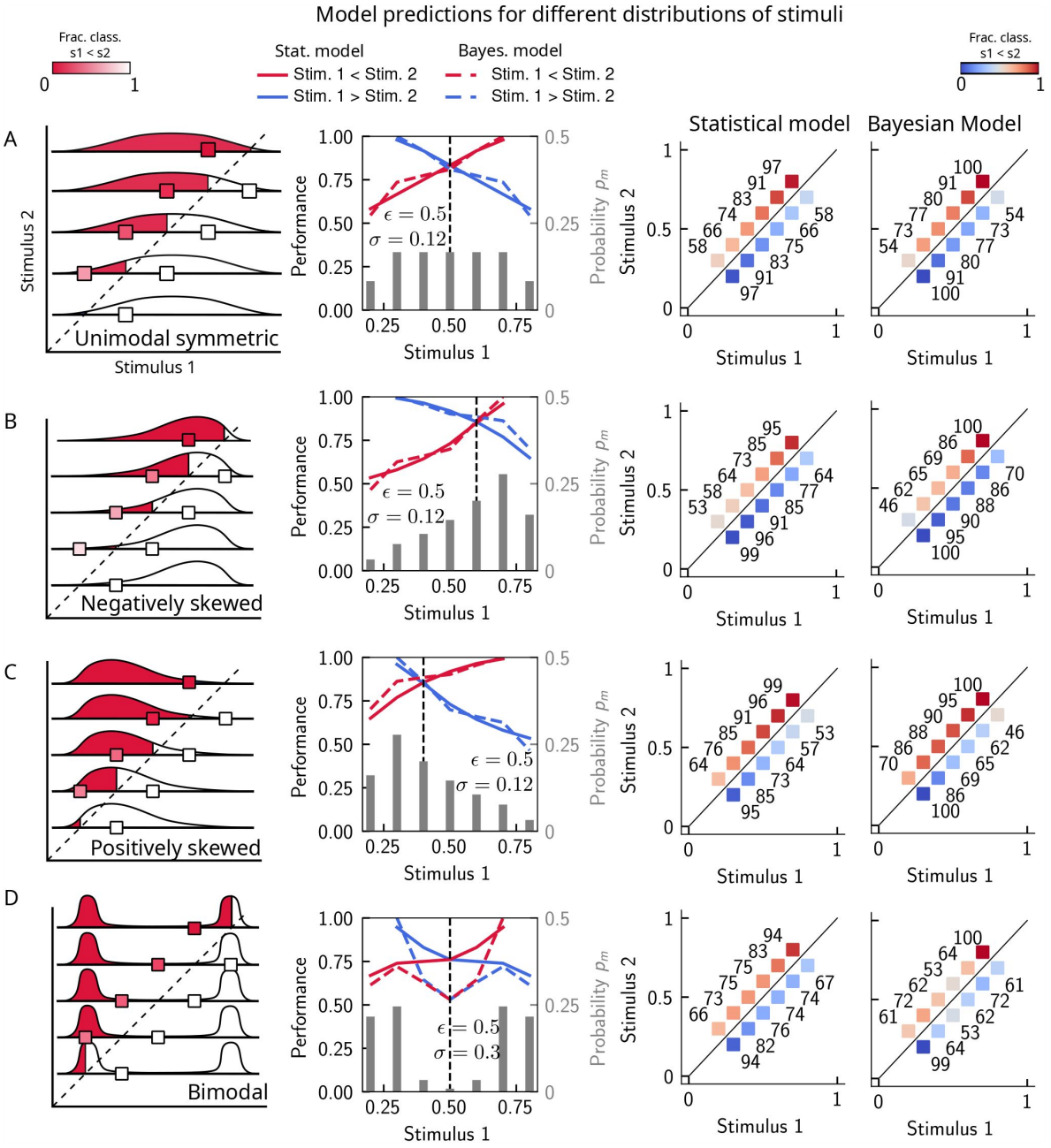

Figure S4

The stimulus distribution impacts the pattern of contraction bias.

The model makes different predictions for the performance, depending on the shape of the stimulus distribution. **(A)** Panel 1: schema of model prediction. Regions shaded in red correspond to the probability of correct comparison, for stimulus pairs above the diagonal, when replacing s_1 with a random value sampled from the marginal distribution with a resampling probability $\epsilon = 0.25$ (see **Fig. 4**). Panel 2: prediction of both models for a unimodal symmetric (in this case quasi-uniform) stimulus distribution, statistical model (solid line) and Bayesian model (dashed line). The marginal stimulus distribution is shown in grey bars (to be read with the right y-axis). The value of s_1 for which there is equal performance for pairs of stimuli below and above the diagonal is indicated by the vertical dashed line, corresponding to the median of the distribution. Panel 3: for each stimulus pair, fraction of trials classified as $s_1 < s_2$ (colorbar), for statistical model. Panel 4: same as panel 3, but for Bayesian model of equal average performance (corresponding to a width of the likelihood of $\sigma = 0.08$ (see *Sect. 1.6.1* and *Sect. 4.3*). **(B)** Similar to A, for a negatively skewed distribution. **(C)** Similar to A, for a positively skewed distribution. **(D)** Similar to A, for a bimodal distribution.

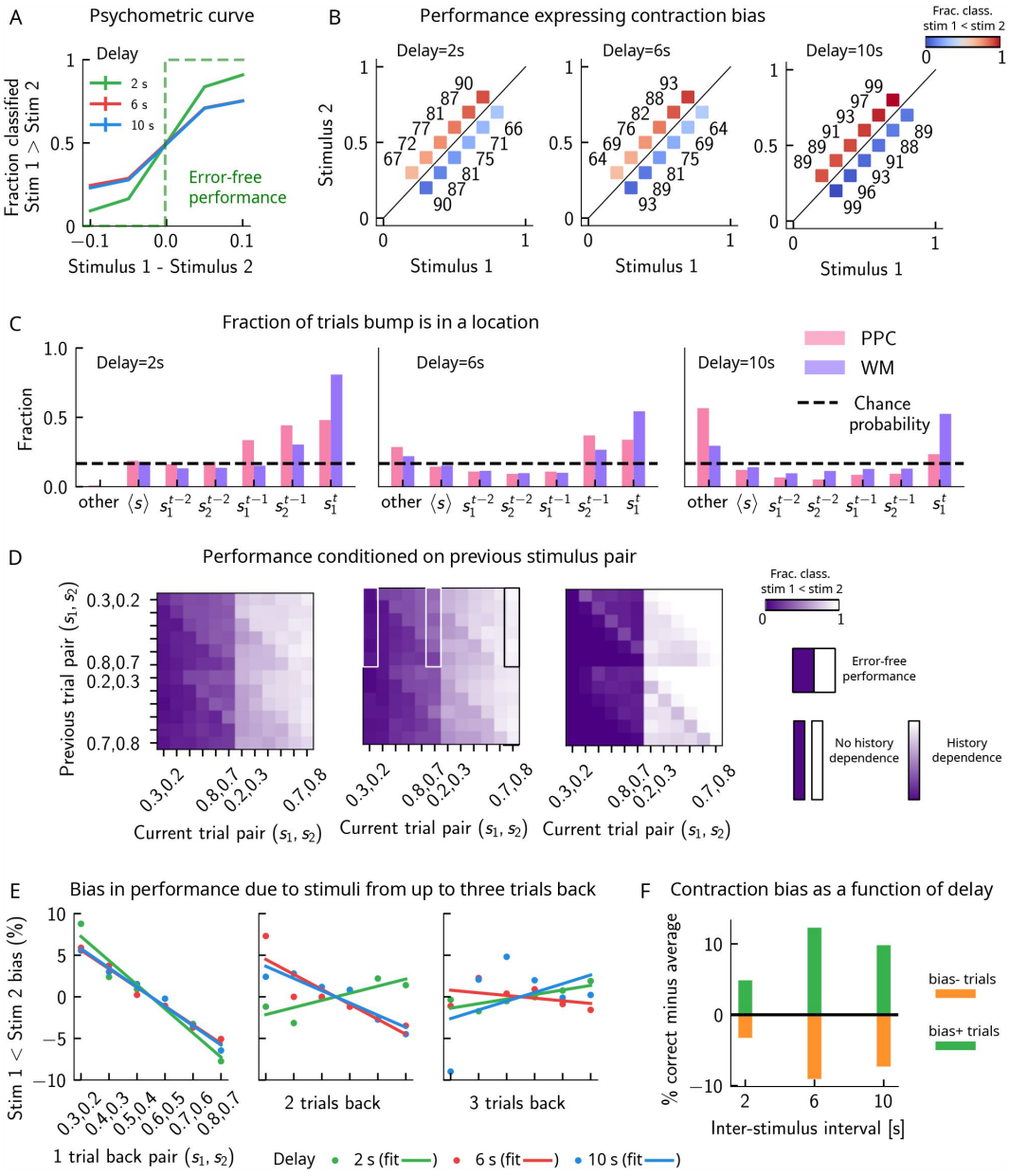

Figure S5

Model predictions for a block design.

(A) As in (**Fig. 2 A**). Performance of the network model for the psychometric stimuli improves with a short delay interval and worsens as this delay is increased. **(B)** As in (**Fig. 2 C**). Performance is affected by contraction bias – a gradual accumulation of errors for stimuli below (above) the diagonal upon increasing (decreasing) s_1. As the delay interval increases, the contraction bias is increased which results in reduced performance across all pairs. Colorbar indicates the fraction of trials classified as $s_1 < s_2$. **(C)** As in (**Fig. 3 C**). The location of the bump that corresponds to the value of s_1 occupies a smaller fraction of trials, as the delay interval increases. **(D)** As in (**Fig. 2 D**). Performance is affected by the previous stimulus pairs (modulation along the y-axis), and becomes worse as the delay interval is increased. The colorbar corresponds to the fraction classified $s_1 < s_2$. **(E)** As in (**Fig. 3 F**). Bias, quantifying the (attractive) effect of the previous stimulus pairs, each color corresponding to a different delay interval. These history effects are attractive: the larger the previous trial stimulus pair, the higher the probability of classifying the first stimulus s_1 as large, and vice-versa. Middle/right panels: same as the left panel, for stimuli extending two and three trials back. **(F)** Quantifying contraction bias separately for Bias+ trials (green) and Bias-trials (orange) yields an increasing bias as the inter-stimulus interval increases.

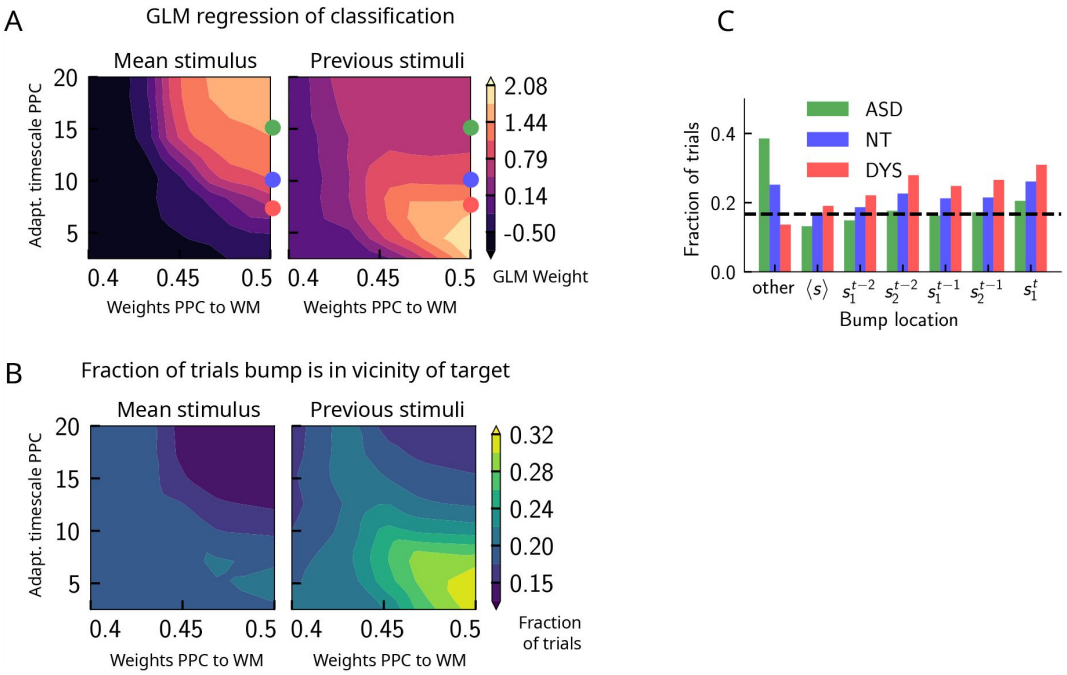

Figure S6

Apparent trade-off between short- and long-term biases, controlled by the timescale of neural adaptation.

(A) Left: GLM weight associated with the regressor corresponding to the mean stimulus across trials (value indicated by colorbar), as a function of the strength of the weights from the PPC to the WM network (x-axis), and the adaptation timescale in the PPC (y-axis). Right: Same as left panel, but displaying the GLM weight associated with the regressor corresponding to the previous trial's stimulus. These two panels indicate that the adaptation timescale *seemingly* exerts a trade-off between the two biases: while decreasing it increases short-term sensory history biases, increasing it increases long-term sensory history biases. The values of the adaptation parameter marked by the three colored dots (in red, blue and green) can mimic behaviors similar to dyslexic, neurotypical, and autistic spectrum subjects (see also **Fig. 9**). **(B)** Left: phase diagram of the fraction of trials in which the activity bump at the end of the delay interval is in the location of the running mean stimulus as a function of the strength of the weights from the PPC to the WM network (x-axis), and the adaptation timescale in the PPC (y-axis). Right: Same as (left), but for the location of any of the two stimuli presented in the previous trial. **(C)** The fraction of trials in which the activity bump at the end of the delay interval corresponds to different locations shown in the x-axis, for three different values of the adaptation timescale parameter, corresponding to qualitatively similar to dyslexic, neurotypical, and autistic spectrum subjects, shown in colors.

Parameter	Symbol	Default value
Number of neurons	N	1000
Neuronal gain	β	5
Range of excitatory interactions [in units of stimulus space length]	d_0	0.02
Strength of inhibitory weights	J_0	0.2
Strength of excitatory weights	J_e	1
Time scale of neuronal integration [s]	τ^P	0.01
Amplitude of external inputs	I_{ext}	1
Duration of stimuli [s]		0.4
Width of box stimulus [in units of stimulus space length]	δs	0.05

Table 2

Simulation parameters Fig. S1.

Parameter	Symbol	Default value
Time-scale of neuronal adaptation in PPC [s]	τ_θ^P	7.5 (DYS), 10 (NT), 15 (ASD)
Amplitude of adaptation in PPC	D^P	0.2
Delay interval [s]		$[2, 6, 10]$
Inter-trial interval [s]		2.2

Table 3

Simulation parameters Fig. 9 and Fig. S6. Other parameters as in Tab. 1

References

[1] Harry Levi Hollingworth (1910) **The central tendency of judgment** *The Journal of Philosophy, Psychology and Scientific Methods* **7**:461–469

[2] Jou Jerwen, Leka Gary E, Rogers Dawn M, Matus Yolanda E (2004) **Contraction bias in memorial quantifying judgment: Does it come from a stable compressed memory representation or a dynamic adaptation process?** *The American journal of psychology* :543–564

[3] Berliner JE, Durlach NI, Braida LD (1977) **Intensity perception. vii. further data on roving-level discrimination and the resolution and bias edge effects** *The Journal of the Acoustical Society of America* **61**:1577–1585

[4] Hellstr öm. Åke (1985) **The time-order error and its relatives: Mirrors of cognitive processes in comparing** *Psychological Bulletin* **97**

[5] Raviv Ofri, Ahissar Merav, Loewenstein Yonatan (2012) **How recent history affects perception: the normative approach and its heuristic approximation** *PLoS Comput Biol* **8**

[6] Fischer Jason, Whitney David (2014) **Serial dependence in visual perception** *Nature neuroscience* **17**:738–743

[7] Akrami Athena, Kopec Charles D, Diamond Mathew E, Brody Carlos D (2018) **Posterior parietal cortex represents sensory history and mediates its effects on behaviour** *Nature* **554**:368–372

[8] Kiyonaga Anastasia, Scimeca Jason M, Bliss Daniel P, Whitney David (2017) **Serial dependence across perception, attention, and memory** *Trends in Cognitive Sciences* **21**:493–497

[9] Cicchini Guido Marco, Mikellidou Kyriaki, Burr David (2017) **Serial dependencies act directly on perception** *Journal of vision* **17**:6–6

[10] Czoschke Stefan, Fischer Cora, Beitner Julia, Kaiser Jochen, Bledowski Christoph (2019) **Two types of serial dependence in visual working memory** *British Journal of Psychology* **110**:256–267

[11] Alais David, Kong Garry, Palmer Colin, Clifford Colin (2018) **Eye gaze direction shows a positive serial dependency** *Journal of vision* **18**:11–11

[12] Manassi Mauro, Liberman Alina, Kosovicheva Anna, Zhang Kathy, Whitney David (2018) **Serial dependence in position occurs at the time of perception** *Psychonomic Bulletin & Review* **25**:2245–2253

[13] Manassi Mauro, Liberman Alina, Chaney Wesley, Whitney David (2017) **The perceived stability of scenes: serial dependence in ensemble representations** *Scientific reports* **7**:1–9

[14] arez-Pinilla Marta S ú, Seth Anil K, Roseboom Warrick (2018) **Serial dependence in the perception of visual variance** *Journal of Vision* **18**:4–4

[15] Papadimitriou Charalampos, Ferdoash Afreen, Snyder Lawrence H (2015) **Ghosts in the machine: memory interference from the previous trial** *Journal of neurophysiology* **113**:567–577

[16] Lieder Itay, Adam Vincent, Frenkel Or, Jaffe-Dax Sagi, Sahani Maneesh, Ahissar Merav (2019) **Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia** *Nature neuroscience* **22**:256–264

[17] Barbosa J õ ao, Compte Albert (2020) **Build-up of serial dependence in color working memory** *Scientific reports* **10**:1–7

[18] Fritsche Matthias, Spaak Eelke, De Lange Floris P (2020) **A bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception** *Elife* **9**

[19] Ashourian Paymon, Loewenstein Yonatan (2011) **Bayesian inference underlies the contraction bias in delayed comparison tasks** *PloS one* **6**

[20] Romo Ranulfo, Salinas Emilio (2003) **Flutter discrimination: neural codes, perception, memory and decision making** *Nature Reviews Neuroscience* **4**:203–218

[21] Seung H Sebastian (1998) **Continuous attractors and oculomotor control** *Neural Networks* **11**:1253–1258

[22] Wang Xiao-Jing (2001) **Synaptic reverberation underlying mnemonic persistent activity** *Trends in neurosciences* **24**:455–463

[23] Zhong Weishun, Lu Zhiyue, Schwab David J, Murugan Arvind (2020) **Nonequilibrium statistical mechanics of continuous attractors** *Neural Computation* **32**:1033–1068

[24] Spalla Davide, Cornacchia Isabel Maria, Treves Alessandro (2021) **Continuous attractors for dynamic memories** *Elife* **10**

[25] Wu Si, Amari Shun-ichi (2005) **Computing with continuous attractors: stability and online aspects** *Neural computation* **17**:2215–2239

[26] Wu Si, Wong KY Michael, Fung CC Alan, Mi Yuanyuan, Zhang Wenhao (2016) **Continuous attractor neural networks: candidate of a canonical model for neural information representation** *F1000Research* **5**

[27] Fung CC Alan, Wong KY Michael, Wu Si (2008) **Dynamics of neural networks with continuous attractors** *EPL (Europhysics Letters)* **84**

[28] Fung CC Alan, Wong KY Michael, Wu Si (2010) **A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks** *Neural Computation* **22**:752–792

[29] Trappenberg Thomas P (2005) **Continuous attractor neural networks** *In Recent developments in biologically inspired computing* :398–425

[30] Murray John D *et al.* (2014) **A hierarchy of intrinsic timescales across primate cortex** *Nature neuroscience* **17**:1661–1663

[31] Siegle Joshua H *et al.* (2021) **Survey of spiking in the mouse visual system reveals functional hierarchy** *Nature* **592**:86–92

[32] Gao Richard, van den Brink Ruud L, Pfeffer Thomas, Voytek Bradley (2020) **Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture** *Elife* **9**

[33] Wang Xiao-Jing, Jiang Junjie, Pereira-Obilinovic Ulises (2023) **Bifurcation in space: Emergence of function modularity in the neocortex** *bioRxiv* :2023–6

[34] Mej ías Jorge F, Wang Xiao-Jing (2022) **Mechanisms of distributed working memory in a large-scale network of macaque neocortex** *Elife* **11**

[35] Ding Xingyu, Froudist-Walsh Sean, Jaramillo Jorge, Jiang Junjie, Wang Xiao-Jing (2022) **Predicting distributed working memory activity in a large-scale mouse brain: the importance of the cell type-specific connectome** *bioRxiv* :2022–12

[36] Hernandez Adrian, Salinas Emilio, Garcia Rafael, Romo Ranulfo (1997) **Discrimination in the sense of flutter: new psychophysical measurements in monkeys** *Journal of Neuroscience* **17**:6391–6400

[37] Sinclair Robert J, Burton Harold (1996) **Discrimination of vibrotactile frequencies in a delayed pair comparison task** *Perception & psychophysics* **58**:680–692

[38] Fassihi Arash, Akrami Athena, Esmaeili Vahid, Diamond Mathew E (2014) **Tactile perception and working memory in rats and humans** *Proceedings of the National Academy of Sciences* **111**:2331–2336

[39] Fassihi Arash, Akrami Athena, Pulecchi Francesca, Sch önfelder Vinzenz, Diamond Mathew E (2017) **Transformation of perception from sensory to motor cortex** *Current Biology* **27**:1585–1596

[40] Esmaeili Vahid, Diamond Mathew E (2019) **Neuronal correlates of tactile working memory in prefrontal and vibrissal somatosensory cortex** *Cell reports* **27**:3167–3181

[41] Loewenstein Yonatan, Raviv Ofri, Ahissar Merav (2021) **Dissecting the roles of supervised and unsupervised learning in perceptual discrimination judgments** *Journal of Neuroscience* **41**:757–765

[42] Jazayeri Mehrdad, Shadlen Michael N (2010) **Temporal context calibrates interval timing** *Nature neuroscience* **13**:1020–1026

[43] Fritsche Matthias, Mostert Pim, de Lange Floris P (2017) **Opposite effects of recent history on perception and decision** *Current Biology* **27**:590–595

[44] Li Lux, Chan Arielle, Iqbal Shah M, Goldreich Daniel (2017) **An adaptation-induced repulsion illusion in tactile spatial perception** *Frontiers in human neuroscience* **11**

[45] I Hachen S Reinartz, R Brasselet A Stroligo, Diamond ME (2021) **Dynamics of history-dependent perceptual judgment** *Nature communications* **12**:1–15

[46] Knapen Tomas, Rolfs Martin, Wexler Mark, Cavanagh Patrick (2010) **The reference frame of the tilt aftereffect** *Journal of Vision* **10**:8–8

[47] Boi Marco, ğmen Haluk Ö, Herzog Michael H (2011) **Motion and tilt aftereffects occur largely in retinal, not in object, coordinates in the ternus–pikler display** *Journal of Vision* **11**:7–7

[48] Mathôt Sebastiaan, Theeuwes Jan (2013) **A reinvestigation of the reference frame of the tilt-adaptation aftereffect** *Scientific reports* **3**:1–7

[49] Jaffe-Dax Sagi, Kimel Eva, Ahissar Merav (2018) **Shorter cortical adaptation in dyslexia is broadly distributed in the superior temporal lobe and includes the primary auditory cortex** *ELife* **7**

[50] Jaffe-Dax Sagi, Frenkel Or, Ahissar Merav (2017) **Dyslexics' faster decay of implicit memory for sounds and words is manifested in their shorter neural adaptation** *Elife* **6**

[51] Daniel Algom (1992) **8 memory psychophysics: An examination of its perceptual and cognitive prospects** *In Advances in psychology* :441–513

[52] Eustace Christopher Poulton and Simon Poulton (1989) **Bias in quantifying judgements**

[53] Preuschhof Claudia, Schubert Torsten, Villringer Arno, Heekeren Hauke R (2010) **Prior information biases stimulus representations during vibrotactile decision making** *Journal of Cognitive Neuroscience* **22**:875–887

[54] Olkkonen Maria, McCarthy Patrice F, Allred Sarah R (2014) **The central tendency bias in color perception: Effects of internal and external noise** *Journal of vision* **14**:5–5

[55] Olsen Grethe M, Hovde Karoline, Kondo Hideki, Sakshaug Teri, Haaland Sømme Hanna, Whitlock Jonathan R, Witter Menno P (2019) **Organization of posterior parietal–frontal connections in the rat** *Frontiers in systems neuroscience*

[56] Tong Ke, Chad Dub é. (2022) **A tale of two literatures: A fidelity-based integration account of central tendency bias and serial dependency** *Computational Brain & Behavior* :1–21

[57] Karim Muhsin, Harris Justin A, Langdon Angela, Breakspear Michael (2013) **The influence of prior experience and expected timing on vibrotactile discrimination** *Frontiers in neuroscience* **7**

[58] Kerst Stephen M, Howard James H (1978) **Memory psychophysics for visual area and length** *Memory & Cognition* **6**:327–335

[59] Rahnev Dobromir, Denison Rachel N (2018) **Suboptimality in perceptual decision making** *Behavioral and Brain Sciences* **41**

[60] Maes Amadeus, Barahona Mauricio, Clopath Claudia (2023) **Long-and short-term history effects in a spiking network model of statistical learning** *Scientific Reports* **13**

[61] Fuster Joaquin M, Alexander Garrett E (1971) **Neuron activity related to short-term memory** *Science* **173**:652–654

[62] Miyashita Yasushi, Chang Han Soo (1988) **Neuronal correlate of pictorial short-term memory in the primate temporal cortex** *Nature* **331**:68–70

[63] Funahashi Shintaro, Bruce Charles J, Goldman-Rakic Patricia S (1989) **Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex** *Journal of neurophysiology* **61**:331–349

[64] Funahashi Shintaro, Bruce Charles J, Goldman-Rakic Patricia S (1990) **Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms** *Journal of neurophysiology* **63**:814–831

[65] Romo Ranulfo, Brody Carlos D, Hernández Adrián, Lemus Luis (1999) **Neuronal correlates of parametric working memory in the prefrontal cortex** *Nature* **399**:470–473

[66] Salinas Emilio, Hernandez Adrian, Zainos Antonio, Romo Ranulfo (2000) **Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli** *Journal of neuroscience* **20**:5503–5515

[67] Zhang Xiaoxing *et al.* (2019) **Active information maintenance in working memory by a sensory cortex** *Elife* **8**

[68] Hopfield John J (1982) **Neural networks and physical systems with emergent collective computational abilities** *Proceedings of the national academy of sciences* **79**:2554–2558

[69] Amit Daniel J, Amit Daniel J (1992) **Modeling brain function: The world of attractor neural networks**

[70] Battaglia Francesco P, Treves Alessandro (1998) **Stable and rapid recurrent processing in realistic autoassociative memories** *Neural Computation* **10**:431–450

[71] Barak Omri, Sussillo David, Romo Ranulfo, Tsodyks Misha, Abbott LF (2013) **From fixed points to chaos: three models of delayed discrimination** *Progress in neurobiology* **103**:214–222

[72] Mongillo Gianluigi, Barak Omri, Tsodyks Misha (2008) **Synaptic theory of working memory** *Science* **319**:1543–1546

[73] Barak Omri, Tsodyks Misha (2007) **Persistent activity in neural networks with dynamic synapses** *PLoS Comput Biol* **3**

[74] Barbosa Joao, Stein Heike, Martinez Rebecca L, Galan-Gadea Adrià, Li Sihai, Dalmau Josep, Adam Kirsten, Valls-Solé Josep, Constantinidis Christos, Compte Albert (2020) **Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory** *Nature neuroscience* **23**:1016–1024

[75] Stern Merav, Istrate Nicolae, Mazzucato Luca (2021) **A reservoir of timescales in random neural networks** *bioRxiv*

[76] Barak Omri, Tsodyks Misha, Romo Ranulfo (2010) **Neuronal population coding of parametric working memory** *Journal of Neuroscience* **30**:9424–9430

[77] Miller Paul, Brody Carlos D, Romo Ranulfo, Wang Xiao-Jing (2003) **A recurrent network model of somatosensory parametric working memory in the prefrontal cortex** *Cerebral Cortex* **13**:1208–1218

[78] Machens Christian K, Romo Ranulfo, Brody Carlos D (2005) **Flexible control of mutual inhibition: a neural model of two-interval discrimination** *Science* **307**:1121–1124

[79] Barak Omri, Tsodyks Misha (2014) **Working models of working memory** *Current opinion in neurobiology* **25**:20–24

[80] Machens Christian K, Romo Ranulfo, Brody Carlos D (2010) **Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex** *Journal of Neuroscience* **30**:350–360

[81] Salinas Emilio (2011) **Prior and prejudice** *Nature neuroscience* **14**:943–945

[82] Dalgleish Henry WP, Russell Lloyd E, Packer Adam M, Roth Arnd, Gauld Oliver M, Greenstreet Francesca, Thompson Emmett J, Häusser Michael (2020) **How many neurons are sufficient for perception of cortical activity?** *Elife* **9**

Article and author information

Vezha Boboeva
Sainsbury Wellcome Centre, University College London, Department of Bioengineering, Imperial College London
ORCID iD: 0000-0002-2476-8714

Alberto Pezzotta
Gatsby Computational Neuroscience Unit, University College London, The Francis Crick Institute
ORCID iD: 0000-0002-8998-7942

Claudia Clopath
Department of Bioengineering, Imperial College London
For correspondence: c.clopath@imperial.ac.uk
ORCID iD: 0000-0003-4507-8648

Athena Akrami
Sainsbury Wellcome Centre, University College London
For correspondence: athena.akrami@ucl.ac.uk
ORCID iD: 0000-0001-5711-0903

Copyright

Editors

Reviewing Editor
Tatyana Sharpee
Salk Institute for Biological Studies, La Jolla, United States of America

Senior Editor
Laura Colgin
University of Texas at Austin, Austin, United States of America

Reviewer #1 (Public Review):

This paper aims to explain recent experimental results that showed deactivating the PPC in rats reduced both the contraction bias and the recent history bias during working memory tasks. The authors propose a two-component attractor model, with a slow PPC area and a faster WM area (perhaps mPFC, but unspecified). Crucially, the PPC memory has slow adaptation that causes it to eventually decay and then suddenly jump to the value of the last stimulus. These discrete jumps lead to an effective sampling of the distribution of stimuli, as opposed to a gradual drift towards the mean that was proposed by other models. Because these jumps are single-trial events, and behavior on single events is binary, various statistical measures are proposed to support this model. To facilitate this comparison, the authors derive a simple probabilistic model that is consistent with both the mechanistic model and behavioral data from humans and rats. The authors show data consistent with model predictions: longer interstimulus intervals (ISIs) increase biases due to a longer effect over the WM, while longer intertrial intervals (ITIs) reduce biases. Finally, they perform new experiments using skewed or bimodal stimulus distributions, in which the new model better fits the data compared to Bayesian models.

The mechanistic proposed model is simple and elegant, and it captures both biases that were previously observed in behavior, and how these are affected by the ISI and ITI (as explained above). Their findings help rethink whether our understanding of contraction bias is correct.

On the other hand, the main proposal - discrete jumps in PPC - is only indirectly verified. The majority of the behavioral predictions stem from the probabilistic model, which is consistent with the mechanistic one, but does not necessitate it.
The revised submission uses the self-paced nature of the experiments to confirm the systematic change in bias with inter-trial-interval, as predicted by the model. This analysis strengthens the hypothesis.

- https://doi.org/10.7554/eLife.86725.2.sa1

Reviewer #2 (Public Review):

Working memory is not error free. Behavioral reports of items held in working memory display several types of bias, including contraction bias and serial dependence. Recent work from Akrami and colleagues demonstrates that inactivating rodent PPC reduces both forms of bias, raising the possibility of a common cause.

In the present study, Boboeva, Pezotta, Clopath, and Akrami introduce circuit and descriptive variants of a model in which the contents of working memory can be replaced samples from recent sensory history. This volatility manifests as contraction bias and serial dependence in simulated behavior, parsimoniously explaining both sources of bias. The authors validate their model by showing that it can recapitulate previously published and novel behavioral results in rodents and neurotypical and atypical humans.

Both the modeling and the experimental work is rigorous, providing convincing evidence that a model of working memory in which reports sometimes sample past experience can produce both contraction bias and serial dependence, and that this model is consistent with behavioral observations across rodents and humans in the parametric working memory (PWM) task.

These efforts constitute an admirable initial validation of the proposed model, and the authors present several novel predictions that will allow for further tests in future experiments. First, the authors note that their circuit model predicts a bimodal error distribution in delayed estimation paradigms (due to noisy sampling of sensory history on a

subset of trials) that merges into a unimodal distribution when recent sensory history and the current to-be-reported stimulus have very similar values (Fig. 5c). Analysis of extent delayed estimation datasets (e.g., https://osf.io/jmkc9/) or new experiments will provide the opportunity for a straightforward test of this hypothesis.

Second, the bulk of the modeling efforts presented here are devoted to a circuit-level description of how putative posterior parietal cortex (PPC) and working-memory (WM) related networks may interact to produce such volatility and biases in memory. This effort is extremely useful because it allows the model to be constrained by neural observations and manipulations in addition to behavior, and the authors begin this line of inquiry here (by showing that the circuit model can account for effects of optogenetic inactivation of rodent PPC). As the authors note, population electrophysiology in PPC and WM-related areas and single-trial analyses will play an important role in the ultimate validation of this model.

Finally, it is noteworthy that, in the spirit of moving away from an overreliance on p-values (e.g., Amrhein et al., PeerJ 2017), the authors eschew conventional hypothesis testing when reporting their experimental results. The p-values aren't missed, in large part thanks to excellent visualizations and apparently large effect sizes. It's unclear how well this approach would generalize to other questions and datasets; nevertheless, this study provides an interesting data point in the ongoing conversation around reproducibility and rigor.

- https://doi.org/10.7554/eLife.86725.2.sa0

Author Response

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *This is an important study about the mechanisms underlying our capacity to represent and hold recent events in our memory and how they are influenced by past experiences. A key aspect of the model put forward here is the presence of discrete jumps in neural activity with the posterior parietal region of the cortex. The strength of evidence is largely solid, with some weaknesses noted in the methodology. Both reviewers suggested ways in which this aspect of the model can to be tested further and resolve conflicts with previously published experimental results, in particular the study by Papadimitriou et al 2014 in Journal of Neurophysiology.*

We thank the editors for their assessment. As mentioned in the cover letter, we have addressed all the reviewers' concerns and would like to request and update of the assessment to reflect the revisions we have made.

We thank both reviewers for their careful reading and feedback that helped clarify many aspects of the model. Below, we address their comments.

> ***Reviewer #1 (Public Review):***
>
> *This paper aims to explain recent experimental results that showed deactivating the PPC in rats reduced both the contraction bias and the recent history bias during working memory tasks. The authors propose a twocomponent attractor model, with a slow PPC area and a faster WM area (perhaps mPFC, but unspecified). Crucially, the PPC memory has slow adaptation that causes it to eventually decay and then suddenly jump to the value of the last stimulus. These discrete jumps lead to an effective sampling of the*

> ***Public Reviews:***

We thank both reviewers for their careful reading and feedback that helped clarify many aspects of the model. Below, we address their comments.

> ***Reviewer #1 (Public Review):***
>
> *This paper aims to explain recent experimental results that showed deactivating the PPC in rats reduced both the contraction bias and the recent history bias during working memory tasks. The authors propose a twocomponent attractor model, with a slow PPC area and a faster WM area (perhaps mPFC, but unspecified). Crucially, the PPC memory has slow adaptation that causes it to eventually decay and then suddenly jump to the value of the last stimulus. These discrete jumps lead to an effective sampling of the distribution of stimuli, as opposed to a gradual drift towards the mean that was proposed by other models. Because these jumps are single-trial events, and behavior on single events is binary, various statistical measures are proposed to support this model. To facilitate this comparison, the authors derive a simple probabilistic model that is consistent with both the mechanistic model and behavioral data from humans and rats. The authors show data consistent with model predictions: longer interstimulus intervals (ISIs) increase biases due to a longer effect over the WM, while longer intertrial intervals (ITIs) reduce biases. Finally, they perform new experiments using skewed or bimodal stimulus distributions, in which the new model better fits the data compared to Bayesian models.*
>
> *The mechanistic proposed model is simple and elegant, and it captures both biases that were previously observed in behavior, and how these are affected by the ISI and ITI (as explained above). Their findings help rethink whether our understanding of contraction bias is correct.*
>
> *On the other hand, the main proposal - discrete jumps in PPC - is only indirectly verified.*

We agree with the reviewer that the evidence for discrete jumps in PPC has been provided in behavioural results (short-term, n-back trial biases), and not from neural data. However, we believe electrophysiological investigations are out of the scope of the current manuscript and future works are needed to further verify the results.

> *The model predicts a systematic change in bias with inter-trial-interval. Unless I missed it, this is not shown in the experimental data. Perhaps the self-paced nature of the experiments allows to test this?*

We thank the reviewer for this great suggestion.

We had not previously looked at this in the data for the reason that in the simulations, the ITI is set to either 2.2, 6 or 11 seconds, whereas the experiment is self-paced. Therefore, any comparison with the simulation should be made carefully.

However, after the reviewer's suggestion, we did look at the change in the bias with the inter-trial interval, by dividing trials according to ITIs lower than 3 seconds ("short" ITI), and higher than 3 seconds ("long" ITI). This choice was motivated by the shape of the distribution of ITIs, which is bimodal, with a peak around 1 second, and another after 3 seconds (new Fig 8F). Hence, we chose 3 seconds as it seemed a natural division. However, 3 seconds also happens to be approximately the 75th percentile of the distribution, and this means that there is much more data in the "short" ITI than the "long" ITI set. In order to have sufficient data in the "long" ITI for clearer effects we used all of our dataset – the negatively skewed, and also two bimodal distributions (of which only one was shown in the manuscript, for

succinctness). This larger dataset allows us to clearly see not only a decreasing contraction bias with increasing ITI (Fig 8G), but also a decreasing onetrial-back attractive bias with increasing ITI (Fig 8H). We have uploaded all the datasets as well as scripts used to analyze them to this repository: https://github.com/vboboeva/ParametricWorkingMemory_Data.

> *The data in some of the figures in the paper are hard to read. For instance, Figure 3B might be easier to understand if only the first 20 trials or so are shown with larger spacing. Likewise, Figure 5C contains many overlapping curves that are hard to make out.*

We have limited the dynamics in Fig 3B to the first 50 trials for better visibility. Likewise, as suggested, we report the standard error of the mean instead of the standard deviation in old Fig 5C (new Fig 6C) – this allows for the different curves to be better discernible.

> *There is a gap between the values of tau_PPC and tau_WM. First - is this consistent with reports of slower timescales in PFC compared to other areas?*

Recent studies by Xiao-Jing Wang and colleagues (Refs. 1-3 below) suggest that may be the case. In Wang et al 2023, Ref 1 below), the authors use a generative model to study the concept of bifurcation in space in working memory, that is accompanied by an inverted-V shape of the time constants as a function of cortical hierarchy.

Briefly, they propose a generative model of the cortex with modularity, incorporating repeats of a canonical local circuit connected via long-range connections. In particular, the authors define a hierarchy for each local circuit. At a critical point in this hierarchy axis, there is a phase transition from monostability to bistability in the firing rate. This means that a local circuit situated below the critical point will only display a low activity steady state, while those above the critical point additionally display a persistent activity steady state.

The model predicts a critical slowing down of the neural fluctuations at the critical point, resulting in an inverted-V shape of the time constants as a function of the hierarchy. They test the predictions of their model – the bifurcation in space and that inverted-V-shaped time constants as a function of the hierarchy - on connectome-based models of the macaque and mouse cortex. Interestingly both datasets show similar behavior. In particular, during working memory, frontal areas (higher in the hierarchy, e.g. area 24c in macaques) has a smaller time constant relative to posterior parietal areas (lower in the hierarchy, like LIP or f7). We have now cited this new work.

[1] https://www.biorxiv.org/content/10.1101/2023.06.04.543639v1

[2] https://elifesciences.org/articles/72136

[3] https://www.biorxiv.org/content/10.1101/2022.12.05.519094v3.abstract

> *Second - is it important for the model, or is it mostly the adaptation timescale in PPC that matters?*

We have run simulations producing a phase diagram with tau_theta^P on the x-axis, tau^P on the y-axis, and in color, the fraction of trials in which the bump is in the vicinity of a target (Fig S2 F), before the network is presented with the second stimulus. This target can be the first stimulus s_1 (left), mean over stimuli (middle) and previous trial's stimulus (right)). White point corresponds to parameters of the default network.

In this phase diagram, the lowest value that tau_P takes is tau_WM=0.01. When tau_P=tau_WM, the bump is rarely in the vicinity of 1-trial-back stimulus, and we can see that

tau_PPC should be greater than tau_WM in order for the model to yield 1-trial back effects. We conclude that it is indeed important for tau_PPC > tau_WM.

We have included this in Fig S2 F of the manuscript.

> *Regarding the relation to other models, the model by Hachen et al (Ref 45) also has two interacting memory systems. It could be useful to better state the connection, if it exists.*

The model proposed by Hachen et al is conceptually different in that one module stores the mean of the sensory stimulus; it could be related to a variant of our model where adaptation is turned off in the PPC network (Fig S2 A). However, the task they model is also different: subjects have to learn the location of a boundary according to which the stimulus is classified as 'weak' or 'strong', set by the experimenter. Hence, it is a task where learning is needed - this contrasts with the task we are modelling, where only working memory is required. How task demands reconfigure existing circuits via dynamics and/or learning to perform different computations is a fascinating area of research that is outside the scope of this work.

> ***Reviewer #2 (Public Review):***
>
> *Working memory is not error free. Behavioral reports of items held in working memory display several types of bias, including contraction bias and serial dependence. Recent work from Akrami and colleagues demonstrates that inactivating rodent PPC reduces both forms of bias, raising the possibility of a common cause.*
>
> *In the present study, Boboeva, Pezzotta, Clopath, and Akrami introduce circuit and descriptive variants of a model in which the contents of working memory can be replaced by previously remembered items. This volatility manifests as contraction bias and serial dependence in simulated behavior, parsimoniously explaining both sources of bias. The authors validate their model by showing that it can recapitulate previously published and novel behavioral results in rodents and neurotypical and atypical humans.*
>
> *Both the modeling and the experimental work is rigorous, providing compelling evidence that a model of working memory in which reports sometimes sample past experience can produce both contraction bias and serial dependence, and that this model is consistent with behavioral observations across rodents and humans in the parametric working memory (PWM) task.*
>
> *Evidence for the model advanced by the authors, however, remains incomplete. The model makes several bold predictions about behavior and neural activity, untested here, that either conflict with previous findings or have yet to be reported but are necessary to appropriately constrain the model.*
>
> *First, in the most general (descriptive) formulation of the Boboeva et al. model, on a fraction of trials items in working memory are replaced by items observed on previous trials. In delayed estimation paradigms, which allow a more direct behavioral readout of memory items on a trial-by-trial basis than the PWM task considered here, reports should therefore be locked to previous items on a fraction of trials rather than display a small but consistent bias towards previous items. However, the latter has been reported (e.g., in primate spatial working memory, Papadimitriou et al., J Neurophysiol 2014). The ready availability of delayed estimation datasets online (e.g., from Rademaker and colleagues, https://osf.io/jmkc9/) will facilitate in-depth investigation and reconciliation of this issue.*

As pointed out by the reviewer, in the PWM task that we are modelling here, the activity in the network is used to make a binary decision. However, it is possible to directly analyse the network activity before the onset of the second stimulus.

In their manuscript, Papadimitriou et al. study a memory-guided saccade task in nonhuman primates and argue that the animals display a small but consistent bias towards previous items (Fig 2). In that figure, the authors compute the error as the difference between the saccade direction and target direction in each trial. They compute this error for all trials in which the preceding trial's target direction is between 35° and 85° relative to the current trial (counterclockwise with respect to the current trial's target). They discover that the residual error distribution is unimodal with a mode at 1.29° and a mean at 2.21° (positive, so towards the preceding target's direction), from which they deduce a small but systematic bias towards previous trial targets.

We have computed a similar measure for our network with default parameters (Table 1), by subtracting the location of the bump at the end of the delay interval (s_hat(t), 'saccade') from the initial location of the first stimulus in the current trial (s1(t) or the 'target'). We have done this for all trials where s1(t)=0.2, and where s2(t-1) takes specific values. These distributions are characterized by two modes. The first corresponds to those trials where the bump is not displaced in WM (i.e. mean of zero). We can also see the appearance of a second mode at the location of s1(t) - s2(t-1), corresponding to the displacements towards the preceding trial's stimulus described in the main text. If, instead, we limit the analysis to a small range of previous trials close to s1(t) (similar to Papadimitriou et al) then the distribution of residual errors will appear unimodal, as the two modes merge. Importantly, note that there is a large variability around the second mode, expressing a more complex dynamics in the network. As can be seen in Fig 3B, the location of the bump is not always slaved to the one in the PPC in a straightforward way -- due to the adaptation in the PPC, the global inhibition in the connectivity kernel, as well as interleaved design for various delay intervals, the WM bump can be displaced in nontrivial ways (see also Recommendation no 4), yielding the dispersion around the second peak. It remains to be seen whether such patterns can be observed in the data from previous works on continuous working memory recall (including Papadimitriou et al). However, to our knowledge, such detailed and full analysis of errors at the level of individual trials has not been done.

In summary, this analysis shows that the type of dynamics in our network is not one of the two cases: 1) small and systematic bias in each and every trial or 2) large error that occurs only rarely; rather, the dispersion around both modes suggests that the dynamics in our model are a mixture of these two limit cases.

We have also performed another typical analysis, reported in several continuous recall tasks (e.g. Jazayeri and Shadlen 2010) where contraction bias has been reported. We plot WM bump locations after the delay period for every trial (s_hat(t)), and their averages, against the nominal value of s1(t). We see that the mean WM location deviates from the identity line toward the mean values of s1(t), again showing contraction bias as an average effect, while individual trials follow the dynamics explained above.

We have now included a new section on continuous recall (Sect. 1.5 and a new figure (Fig 5)), which details the two above-mentioned analyses. The analysis of freely available datasets of delayed estimation tasks, unfortunately, is out of the scope of this work, and we leave such analyses to future studies.

> *Second, the bulk of the modeling efforts presented here are devoted to a circuit-level description of how putative posterior parietal cortex (PPC) and working-memory (WM) related networks may interact to produce such volatility and biases in memory. This effort is extremely useful because it allows the model to be constrained by neural observations and manipulations in addition to behavior, and the authors begin this line of inquiry here (by showing that the circuit model can account for effects of optogenetic inactivation of rodent PPC).*

> *Further experiments, particularly electrophysiology in PPC and WM-related areas, will allow further validation of the circuit model. For example, the model makes the strong prediction that WM-related activity should display 'jumps' to states reflecting previously presented items on some trials. This hypothesis is readily testable using modern high-density recording techniques and single-trial analyses.*

As mentioned in response to the previous comment, we note again that in the WM network, the bump 'displacement' has a complex dynamics -- the examples we have provided in Fig 1A and 2B mainly show the cases in which jumps occur in the WM network, but this is not the only type of dynamics we observe in the model. We do have instances in which the continuity of the model causes drift across values, and we have now replaced the right panel in Fig 2B with one such instance, in order to emphasize that this displacement towards the previous trial's stimulus (s2(t-1)) can occur in various ways. For a more thorough analysis, we have analyzed the distance between s1(t) and the position of the bump in the WM network at the end of the delay period s_hat(t), conditioned on specific values of s1(t) and s2(t-1) (Fig 5C). In this figure, we can see the appearance of two modes: one centered around 0, corresponding to the correct trials where the stimulus is kept in WM (s1(t) = s_hat(t)), and another mode centered around s2(t-1), the location of the second stimulus of the previous trial, where the bump is displaced. Note, as we explain in Sect. 1.5, the large dispersion around this second mode, which suggests that the bump is not always displaced to that specific location and may undergo drift.

We agree with the reviewer that future electrophysiological experiments (or analysis of existing datasets) are necessary for validation of these results.

> *Finally, while there has been a refreshing movement away from an overreliance on p-values in recent years (e.g., Amrhein et al., PeerJ 2017), hypothesis testing, when used appropriately, provides the reader with useful information about the amount of variability in experimental datasets. While the excellent visualizations and apparently strong effect sizes in the paper mitigate the need for p-values to an extent, the paucity of statistical analysis does impede interpretation of a number of panels in the paper (e.g., the results for the negatively skewed distribution in 5D, the reliability of the attractive effects in 6a/b for 2- and 3- trials back).*

We share the reviewer's criticism towards the misuse of p-values – in order for a clearer interpretation of old Fig 5D (new Fig 7E), we have looked at the 2 and 3 trials-back biases by using all of our dataset – the negatively skewed, and also two bimodal distributions (of which only one was shown in the manuscript). This larger dataset of 43 subjects (approximately 17,200 trials) allows us to clearly see the 2 and 3 trial back attractive biases, and the effect that the delay interval exerts on them.

> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Fig 5 A&C - It might be beneficial to separate the distribution of stimuli from the performance. It is hard to read the details of the performance, especially with error bars.*

Following the next recommendation, we have exchanged the standard deviation to standard errors of the mean, hopefully this allows to better read the performance.

> *Fig 5C. The number of participants should be written. Perhaps standard errors instead of standard deviation?*

We have now changed the standard deviation to standard errors of the mean and included the number of participants in the figure.

> *Fig 2B - hard to understand, because there is no marking of where "perfect" memory of s1 would be.*

The perfect memory of s1 is shown in the upper panel as black bars.

> *Fig 3B. dot number 9 (blue, around 0.7) - why is WM higher than stimulus?*

This trial has a long ISI (blue means 10s). During this delay, the bump in the PPC, under the influence of adaptation, drifts far below the first stimulus (note that the previous trial also had its first stimulus in the same location, as a result of which the adaptative thresholds have built up significantly, causing the bump to move away from that location). During this delay period, neurons in the WM network receive inputs from the PPC network: if this input is strong enough, it can disrupt an existing bump; if not, this input still exerts inhibiting influence on the existing bump via the global inhibition in the connectivity. This can cause an existing bump to slowly drift in a random direction, and finally dissipate. Note that the lines in Fig 2B represent the neuron with the maximal activity, this activity may be a stable bump, or an unstable bump that may soon dissipate.

Other examples with similar dynamics include trials 43 and 54.

> *L167 fewer -> smaller*

We have now corrected this.

> *Fig 3C - bump can also be in between. Is this binned?*

We have not binned the length of the attractor; to produce that figure, we check whether the position of the neuron with the maximal firing rate is within a distance of ±5% of the length of the whole line attractor from the target location.

> *L221 Lapse at the boundary of attractor. This seems very different from behavior. Specifically, if it is in the boundaries, it should be stimulus dependent.*

Very sorry, we did not manage to understand the reviewer's comment.

> *L236 are -> is*

We have now corrected this.

> *Fig S4 - should be mostly in main text.*

Part of this figure is in Fig 6A, but given the amount of detail, we think Supplementary Material is better suited.

> *L253-254. Differences across all distributions - very minor except the bimodal case.*

That is correct, this is why we conducted the experiment with the bimodal distribution, to better differentiate the predictions of the two models.

> *L273 extra comma after "This probability"*

We have now corrected this.

> *ITI was only introduced in section 1.5.2. Perhaps worth mentioning the default 5s value earlier in the paper.*

We have now mentioned this in line 97-98.

> *Fig S6B title: perhaps "previous stimuli"?*

We have now corrected this.

> *L364 i"n A given trial"*
>
> *Equation 2 - no decay term?*

Thank you for pointing out this error, we have now corrected this.

> *Equation 5,6 are j^W and j^P indices of neurons in those populations?*

Yes, j^W indexes neurons in the WM network, and j^P those in the PPC. We have now added this in the text for clarity.

> *Bump with adaptation - other REFs? Sandro?*

We are aware of continuous bump attractors implementing short-term synaptic plasticity in various studies (including by Sandro Romani), but not in the form we have described. May the reviewer kindly point us towards the relevant literature.

> *Free boundary - what is the connectivity for neurons 1 and N? Is it weaker than others? Is the integral still 1? Does this induce some bias on the extreme values?*

The connectivity of the network is all-to-all. However, as expressed by Eq. (3), the distance-dependent contribution to the weights, K, decreases exponentially as we move from neuron 1 onwards, and from neuron N down. The sum (or integral, in the large-N limit) of the K_ij for j on either side of neuron i is unity only when i is sufficiently far from 1 or N. We have rephrased the paragraph starting in line 516 to make this clearer.

The presence of a boundary could introduce a bias in theory, but in practice, it affects the dynamics only when the bump drifts sufficiently close to it. The smallest stimulus in the simulated task has amplitude 0.2, with width 0.05, which implies the activation of 50 neurons on either side of neuron 400. If one compares this with the width of the kernel K in stimulus space (d_0 = 0.02), which spans ~10 neurons, we can see that the bump of activity stays mostly far from the boundary. It is possible, though it is observed rarely, when several consecutive long delay intervals happen to occur, that the bump in PPC drifts beyond the location corresponding to either the minimum or maximum stimulus.

> *Code availability?*

Code simulating the dynamics of the network as well as analysing the resulting data can be found in the following repository: https://github.com/vboboeva/ParametricWorkingMemory Code used to analyse human behavioural data and fit them with our statistical model can be found in this repository: https://github.com/vboboeva/ParametricWorkingMemory_Data Code used to run the auditory PWM experiments with human subjects (adapted from Akrami et al 2018) can be found here: https://github.com/vboboeva/Auditory_PWM_human

> *L547 stimuli*

We have now corrected this.

> *Equation 14 uses both stimuli. Was this the same for the rest of analysis in the paper (first figures for instance)?*

This equation was used for all GLM analyses (Figs 9 and S6).

> *D0 is very small (0.02). Does this mean that activity is essentially discrete in the model? Fig 1A & 2B - the two examples of model activity suggest this is the case. In other words - are there cases where the continuity of the model causes drift across values? Can you show an example (similar to Fig 1A)?*

Since this point has been raised beforehand, we refer to the first comment, Fig 2B and Sect. 1.5 for the response to this question.

> *Table 1 - inter trial interval 6. Text says 5*

We have now corrected this in the text.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *In addition to my review above, I just have a few minor comments:*
>
> - *If I understood correctly, the squares inside the purple rectangle in Figure 1B are meant to show a gradation from red to blue, but this was hard to make out in the pdf.*

Actually the squares are all on one side or the other of the diagonal, therefore they do not have any gradation.

> - *line 164: "The resulting dynamics... [are]?"*

We have corrected this in the text.

> - *Fig 7B legend: "The network performance is on average worse for longer ITIs" – correct?*

This was a mistake, we have replaced worse with better.

> *Other comments*

We realized that the colorbar reported the incorrect fraction classified in Figs 1B, 2C, 7B (new 8B), S2C, S3A, S5B. We have corrected this in the new version of the manuscript.

We also found a minor mistake in one of our analysis codes that computed the n-trial back biases for different delay intervals. This did not change our results, actually made the effects clearer. The figures concerned are Fig 3F and new Fig 7E.