Diverse perceptual biases emerge from Hebbian plasticity in a recurrent neural network model

Francesca Schönsberg[*1,2], Davide Giana[3], Yukti Chopra[3],
Mathew Diamond[†§3] and Sebastian Goldt[‡§1]

[1]Department of Physics, International School of Advanced Studies (SISSA), Trieste, Italy
[2]Department of Physics, Ecole normale supérieure, Paris, France
[3]SENSEx Laboratory, International School of Advanced Studies (SISSA), Trieste, Italy

30th May 2024

Abstract

Perceptual biases offer a glimpse into how the brain processes sensory stimuli. While psychophysics has uncovered systematic biases such as contraction and repulsion, a unifying neural network model for how such seemingly distinct biases emerge from learning is lacking. Here, we show that both contractive and repulsive biases emerge from continuous Hebbian plasticity in a single recurrent neural network. We test our model in three experimental paradigms: a working memory task, a reference memory task, and a novel "one-back task" that we design to test the robustness of the model. We find excellent agreement between model predictions and experimental data without fine-tuning the model to any particular paradigm. These results show that apparently contradictory perceptual biases can in fact emerge from a simple local learning rule in a single recurrent region of the brain.

Introduction

The seemingly simple task of categorising the intensity of a stimulus is prone to several perceptual biases. Take for example a tactile intensity working memory task, where the subject is presented with two sequential stimuli on each trial and is required to report whether the second stimulus was weaker or stronger than the first (fig. 1A). Both human and rodent subjects display a *contraction bias* on this task: they tend to overestimate the strength of the first stimulus if it lies below the time-weighted mean of past stimuli, and underestimate the first stimulus if it lies above the mean [1–4]. Another bias is observed in intensity reference memory tasks [5–8], where the subject is presented with one stimulus per trial and reports whether that stimulus was perceived as strong or weak, fig. 1B. Here, a *repulsive bias* emerges: subjects are more likely to characterise a stimulus as strong if the previous trial's stimulus was weak, and vice versa [8].

These biases offer a window into the computational processes underlying perceptual judgements. A number of psychophysical studies, comprising different sensory modalities, species, and details of experimental design, have provided a detailed characterisation of both contraction and repulsion [3, 5–11]. To account for the intriguing biases uncovered, comprehensive phenomenological models have been built [1, 8, 12–16]. Yet, the neural mechanisms that underlie their emergence remain poorly understood. While it was proposed, over 20 years back, that neural circuits could perform working

[*]francesca.schonsberg@phys.ens.fr
[†]diamond@sissa.it
[‡]sgoldt@sissa.it
[§]These authors contributed equally.

memory tasks through a quasi-continuous line of fixed points Machens *et al.* [17] only more recently a neural network model of two static continuous attractors Boboeva *et al.* [18] was shown to account for the specific perceptive biases observed in Akrami *et al.* [3]. The existing theoretical models, however, have thus far failed to capture how perceptual biases arise from short timescale learning. Learning is indeed key in reference memory categorisation, as evidenced by the notable performance disparities of trained subjects between the onset and the end of experimental sessions [8]. Further, while recent phenomenological models [16] point to the existence of general principles underlying diverse perceptual biases, to the best of our knowledge there exists no single neural network model that explains how several biases emerge across different experimental paradigms.

Here, we show that a single recurrent neural network (RNN) with ongoing Hebbian plasticity learns stimulus representations that reflect the perceptual biases of humans and rodents. We first show how the model reproduces contraction and repulsive biases, as well as the history-dependence of choices, as measured in two standard working and reference memory paradigms. Our analysis shows how both effects emerge from a plastic attractor driven by recurrent dynamics and Hebbian plasticity. We then design a new experimental paradigm that tests the generality of the model. In this "one-back task", participants experience a stream of stimuli and have to compare the strength of each stimulus to the previous one, fig. 1C. We find a novel, bi-modal contraction bias in the performance of human subjects and reproduce it in the RNN model. We stress that we do not fine-tune the model to any individual paradigm, nor do we fit the RNN to experimental data or use gradient descent for learning. Instead, we fix hyper-parameters across experiments and use only local Hebbian plasticity to let the weights evolve continuously.

Results

A recurrent neural network model with ongoing Hebbian plasticity

As applied to rat psychophysics [1], the model may be conceived of as a representation of the vibrissal motor cortex (vM1), the main target of somatosensory cortex [20]. vM1 contributes to memory, decision making, and motor planning. Since vM1 is predominantly recurrent, we consider a single recurrent neural network (RNN) model composed of N threshold linear units, fig. 1D, with firing rates $V_i(t)$ and synaptic connections $W_{ij}(t)$, where t is time. For a given connectivity $W_{ij}(t)$, the firing rates of the neurons follow the rate equations [21–23]

$$V_i(t+1) = (1-\gamma)V_i(t) + \gamma g \left[\frac{1}{N} \sum_{j=0}^{N} W_{ij}(t)V_j(t) - T(t) + \eta \xi_i^{\mu}(t) \right]^+, \tag{1}$$

where γ sets the timescale of the neuronal dynamics and $[z]^+ = \max(0, z)$. The neural transfer function within the square brackets, inspired by Treves [24] and Schönsberg *et al.* [25], determines the activity of the neurons through three components: the inputs from the other neurons weighed by the recurrent connectivity W; a time-dependent threshold $T(t)$ to model neural homeostasis, and an external input $\vec{\xi}(t)$. We set the threshold $T(t) \equiv 4\kappa(\langle v_i(t)\rangle - a)^3$, where κ is a hyper-parameter and a is equal to the average of the external inputs.

Rather than optimising the weights of the network using a gradient-based algorithm [26], we endow the network with a biologically plausible continuous Hebbian plasticity [27],

$$W_{ij}(t+1) = (1 - \lambda - V_i(t)^2)W_{ij}(t) + (V_i(t) - a)(V_j(t) - a). \tag{2}$$

The first term implements synaptic homeostasis through an Oja-inspired normalisation term [28], while the firing rates in the second term are thresholded by an average value to enable positive and negative weights. We fix the hyperparameters γ, λ, η, g to values that put the network in a regime where the

[1] The relevant brain region in humans remains unidentified, though new evidence is emerging [19].

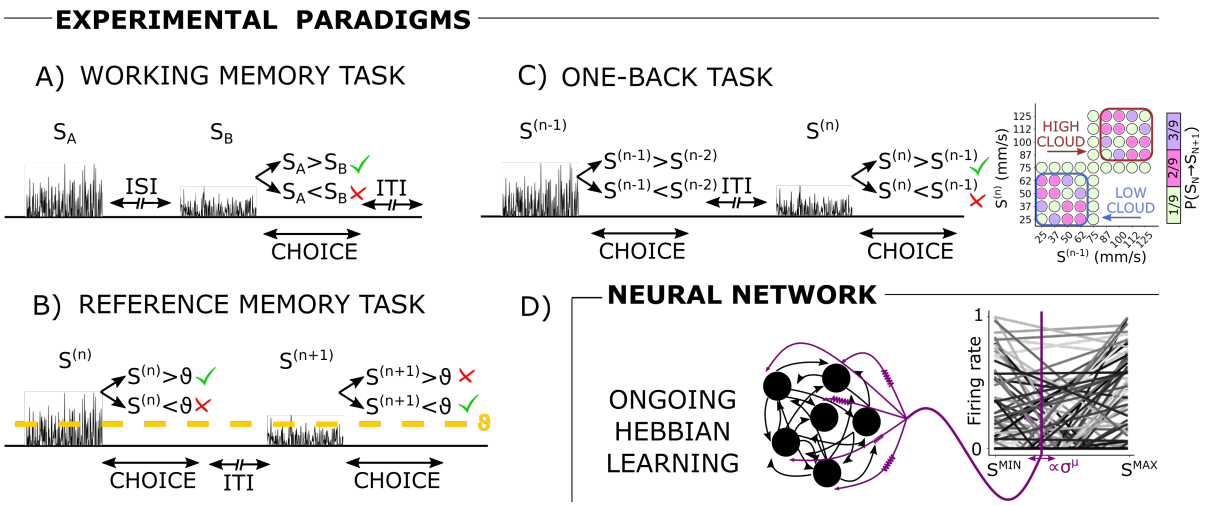

Figure 1: **A single neural network model for three perceptual memory paradigms. A)** In the working memory task, subjects compare the strength of two sequentially presented stimuli separated by an inter stimulus interval (ISI) of various length. **B)** In the reference memory task, subjects compare the strength of one stimulus per trial to a fixed category boundary (orange dashed line). **C)** In the "one-back" task, subjects compare the strength of one stimulus per trial to the strength of the stimulus in the preceding trial. On the right: probabilities governing the strength of the (n)th stimulus given the strength of the $(n-1)$th stimulus for this task. The chosen values for the probabilities lead to long consecutive sequences of stimuli from the low-intensity or high-intensity "clouds". In all three paradigms trials are separated by a variable inter-trial interval (ITI) and subjects receive rewards after correct choices. **D)** The model consists of a fully-connected recurrent neural network (RNN) whose weights are continuously reshaped by Hebbian plasticity. The right panel shows the input profiles for each neuron of the RNN as a function of the strength of the vibrational stimuli, which ranges from S^{MIN} to S^{MAX}. Following the experimental protocol, each time a stimulus of a given strength μ is delivered to the network (violet), its effective value has a stochastic component parameterised by a variability σ^{μ} (see appendix A)

recurrent dynamics and the learning can occur without divergences (see *Methods* for details). We stress that the model will be tested against the experimental results of all three paradigms with the same set of parameters.

We represent vibrational stimuli using a set of external input patterns that reflect key coding properties of the rat's primary somatosensory cortex (vS1) [29, 30]. In a scheme inspired by the binary morphing patterns of Blumenfeld *et al.* [31], the ith RNN unit receives an input ξ_i^{μ} for a stimulus of strength S^{μ} whose value is a threshold linear function of the stimulus strength with a random slope, as shown in fig. 1D and *Methods*. This choice means that inputs display gradual morphing, i.e. the correlation between to patterns $\vec{\xi}^{\mu}$ and $\vec{\xi}^{\nu}$ decays with the absolute difference in the corresponding stimulus intensities $\Delta S = |S^{\mu} - S^{\nu}|$. To obtain the stimulus intensity encoded by the RNN with state $\vec{V}(t)$, we simply find the input pattern ξ^{μ} that is most similar to $\vec{V}(t)$ in terms of the normalised cosine similarity. We denote the stimulus intensity encoded by the RNN at time t as $\hat{S}(t)$.

Human participants and the RNN model display contraction bias in a working memory task

We first applied the model to a human working memory experiment, fig. 1A. Participants were presented with a sequence of trials, with Inter-Trial Intervals (ITI's) set by the subjects, who initiated each trial at their own selected pace. In each trial, the subject was presented one pair of vibratory stimuli of strengths (S_A, S_B), separated by an Inter-Stimulus Interval (ISI) determined by the experimenter. The task was to report, after a go cue, which of the two stimuli (S_A, S_B), was stronger, earning a reward for correct

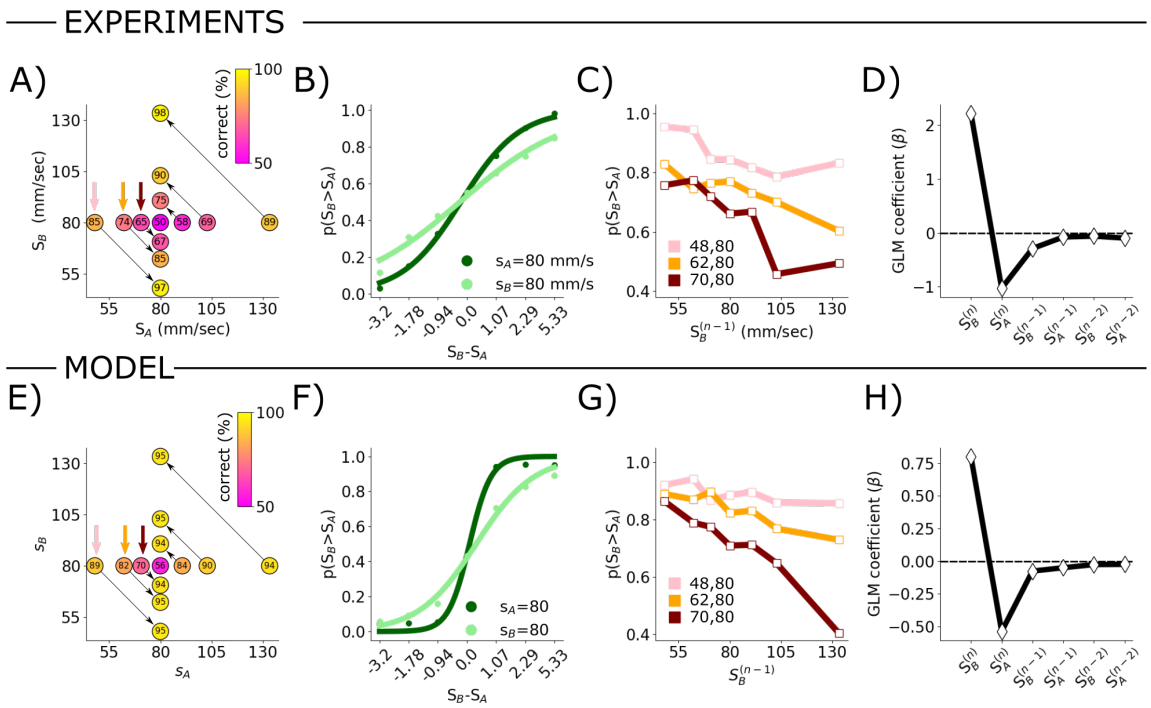

Figure 2: **Contraction bias and history-dependence of choices in a working memory task. A)** Accuracy of human participants in comparing the intensities of two vibrations (S_A, S_B). Arrows explained in main text. **B)** Psychometric curves based on the stimulus difference if either the first or the second stimulus is 80 mm/sec (light and dark green, respectively) **C)** Probability that the second stimulus is (correctly) classified as stronger as a function of the strength of the second stimulus S_B in the preceding trial. This plot is shown for the three stimulus pairs highlighted by coloured arrows in A). **D)** Coefficient values (β) of a GLM fit of the choice at the nth trial as a function of the most recent six perceived stimuli. **E–H)** The same results are reproduced in the RNN model under conditions mimicking those of **A–D)**.

choices (see Methods for details). This paradigm has been used in several forms to explore working memory [2, 3, 17].

The performance of humans shows contraction bias The performance of human participants, fig. 2A, shows intriguing asymmetries. For example, the pair (105, 80) yielded 21 percent lower performance than the pair (80, 105), even if the difference between stimuli is the same (25 mm/s). This asymmetry can be interpreted as an outcome of *contraction bias* [3, 9] as follows: during the ISI the memory of the magnitude of the base stimulus S_A moves towards the average of past stimuli, weighted by their recency. Thus, for the first pair, the memory of $S_A = 105$ moves closer to the stimulus it must be compared to, S_B (80), making the task harder and leading to errors. For the second pair, the memory of $S_A = 80$ moves toward the recency-weighted mean, which is itself in the neighbourhood of 80 mm/sec. Because the base stimulus is in the vicinity of the recency-weighted mean, little contraction occurs, leading to fewer errors. The arrows in fig. 2A point towards the stimulus pair that is easier to discriminate among pairs with the same difference in stimulus strength. The fact that arrows in the upper-right quadrant point leftward while arrows in the lower-left quadrant point rightwards is consistent with contraction bias as an explanation for the differences in stimulus pair difficulty. This bias is also manifest in the average psychometric curves across subjects after separating trials according to whether $S_A = 80$mm/s or $S_B = 80$mm/s, fig. 2B. Unbiased, perfect choices would result in a step function, which is approached by the psychometric curve obtained when the first stimulus $S_A = 80$mm/s. If instead the second stimulus is equal to the mean stimulus, $S_B = 80$mm/s, performance worsens.

Contraction bias in the choices of the RNN To simulate an experiment, both recurrent and Hebbian dynamics, eqs. (1) and (2), start from random initial conditions. We then deliver a sequence of stimulus pairs with the same distribution of intensities and ISI's as in the experiment with human subjects (see *Methods* for details). After the first stimulus has been delivered, the external input $\vec{\xi}(t)$ is set to 0 and the second stimulus is delivered after a delay. If the intensity of the second stimulus is larger than the memory trace of the first stimulus, i.e. the stimulus intensity encoded by the RNN at the onset of the second stimulus, the RNN makes the choice $S_B^{(n)} > S_A^{(n)}$, and vice versa. Reading out choice in this simple way, we find clear evidence of contraction bias in the accuracy of the choices of the RNN fig. 2E, revealing that the memory trace of the RNN contracts towards the average of past stimuli. As seen for human participants, the model's psychometric curves show good accuracy of the RNN on trials where the first stimulus intensity is near the mean stimulus intensity, while accuracy decreases on trials where the second stimulus intensity is near the distribution mean, fig. 2F).

History-dependence of choices in working memory task

A fundamental feature of contraction bias reported in previous studies [3] is recency – the dependence of choices on the stimuli of the immediately preceding trial. We demonstrate this effect in the current data set in fig. 2C. The coloured lines show the probability of making the choice $S_B^{(n)} > S_A^{(n)}$ for the three different pairs in the nth trial as a function of the second stimulus in the $(n-1)$th trial, $S_B^{(n-1)}$. For all three pairs of stimuli shown in fig. 2C, the weaker the preceding stimulus $S_B^{(n-1)}$, the higher the probability that $S_B^{(n)}$ was correctly classified as stronger than $S_A^{(n)}$. This can be interpreted as showing that the stimulus $S_B^{(n-1)}$ is strongly weighted in the continuously updating average towards which $S_A^{(n)}$ is contracted. Tracing the influence of all recently perceived stimuli $S_B^{(n)} - S_A^{(n)} - S_B^{(n-1)} - S_A^{(n-1)} - S_B^{(n-2)} - S_A^{(n-2)}$ on the choice at trial n using the coefficient values β of a generalised linear model (GLM) reveals a positive effect of the last stimulus (meaning that subjects correctly used the current trial's s_B magnitude to make choices) but a negative effect of previous stimuli which decays for stimuli that are further back in time, see fig. 2D. As shown in fig. 2G, choices made based on RNN representations have the same history-dependence on previous choices as do the choices of human subjects, which is further confirmed by a GLM analysis of RNN based choices (H).

Recurrent dynamics and continuous plasticity yield perceptual biases

Continuous Hebbian plasticity creates a dynamical attractor How do the perceptual biases arise in the RNN model? At the beginning of a session, when the connectivity is still close to its random initialisation, the first external input (violet line in fig. 3A) drives the RNN to develop a large overlap with that stimulus, as shown by the overlap profiles in fig. 3A, showing the cosine similarity of a given network state with all the different input patterns (going from light grey to dark grey as time increases). However, as soon as the external input is released, the RNN quickly decorrelates and the overlap profile becomes almost flat, fig. 3B. In other words, no particular stimulus intensity is represented by the network. After a few more trials, the effect of Hebbian plasticity becomes apparent: overlap profiles remain peaked even after the stimulus release, meaning the network represents a uniquely identifiable stimulus intensity every time, the one corresponding to the peak of the bell-shaped curve in fig. 3D. This is the result of the emergence of an attractor state with a peaked overlap profile, given by the combination of continuous Hebbian plasticity and morphing input pattern structure. This attractor can be explicit at each time step t by freezing the connections $W(t)$ and letting the network converge to the fixed point $\vec{V}_t^{FP}$ of the recurrent dynamics eq. (1). The overlap profiles of the attractors at the sample frozen steps are shown in green in fig. 3 A to D (light to dark as time increases).

The relaxation dynamics of the RNN can lead to perceptual biases The constrains of the network state's dynamical flow determined by the attractor, together with the slower Hebbian induced movement of the attractor itself, lead to a categorisation prone to perceptual biases. The argument can be simplified

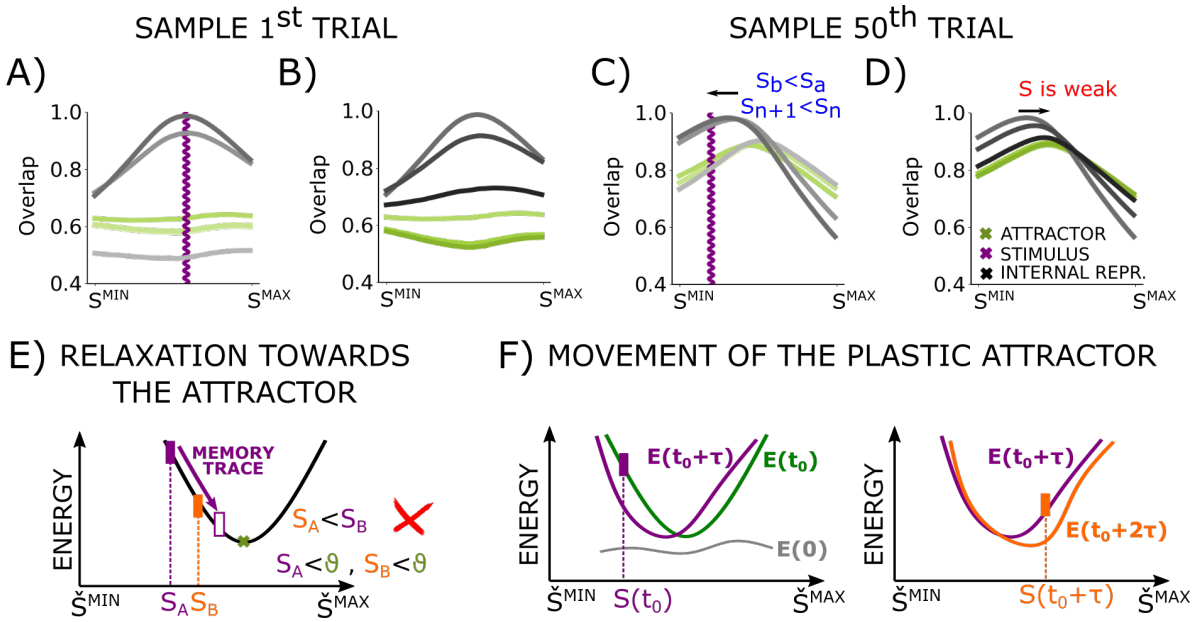

Figure 3: **A neural network model to capture perceptual biases. A)–D)** Dynamics of the RNN during the 1st and the 50th trial visualised via the "overlap profile", i.e. the normalised cosine similarity of the RNN state $\vec{V}(t)$ with every input stimulus (grey lines, increasingly dark as time increases). The green lines represent the overlap profile of the attractor or fixed point of the RNN dynamics, eq. (1), with the weights fixed at $W(t)$. The intensity of the external stimulus is indicated in violet. Note that while the overlap profile quickly decays to an almost uniform profile after the first stimulus (B), the ongoing Hebbian plasticity leads to the emergence of an attractor with peaked overlap profile at the 50th trial (C, D). We indicate the choices made in a reference memory task in red, and in blue for the working and one-back task, in which we consider the plotted stimulus as the second one. **E)&F)** Contraction arises from relaxation towards a plastic attractor, which is reshaped by continuous plasticity. Sketch depicts the energy landscape of the RNN as a function of the internal representation of the stimulus strength near its minimum. **E)** After a stimulus below average S_A is delivered (full violet), the state of the memory, and hence the memory trace of that stimulus, decays towards the bottom of this energy landscape, i.e. the attractor (green cross) corresponding to the recency weighted average of previous stimuli. A second stimulus S_B (orange) with a below-average intensity, can be misclassified if by the time of its delivery the memory representation of the first stimulus (empty violet) has overtaken it. Choices are noted for a working memory and a reference memory task, in the latter consider the two stimuli as part of two different trials. **F)** Sketched energy profiles of the attractor at various times: At $t = 0$, the energy $E(0)$ is not correlated with one particular stimulus representation (grey). As stimuli are delivered, the energy $E(t_0)$ starts having a clear minimum correlated to a specific intensity. Each further stimulation continuously shifts the position of the attractor via Hebbian plasticity. Note that the phenomena sketched separately in E) and F) occur simultaneously.

by first considering a network with fixed weights. The position of the attractor corresponding to these weights, and hence the stimulus intensity encoded by it, can be thought of as the minimum of an effective energy function $E(W)$ [24, 32], sketched in fig. 3E-F in the vicinity of its minimum as a function of the stimulus intensity encoded by the network. At the offset of the first stimulus S_A in a working memory trial (violet in fig. 3E, the state of the RNN evolves to minimise the (effective) energy function. This relaxation of $\vec{V}(t)$ towards the attractor means that the stimulus intensity encoded by the RNN relaxes towards that of the minimum of the effective energy function. In the trial sketched in fig. 3E, this relaxation reduces the distance between the memory trace of the first stimulus and the intensity of the second stimulus S_B (shown in orange). If the inter-stimulus interval is long enough for the memory trace to "overtake" the intensity of the second stimulus, the RNN makes an erroneous choice

by classifying the intensity of the second stimulus as lower than the first.

Continuous Hebbian plasticity results in a dynamic attractor While the location of the attractor will be close to the mean stimulus intensity on average, the ongoing Hebbian plasticity means that the position of the attractor is continuously driven by the external stimuli. In fig. 3F, we illustrate the effect of two stimuli delivered one after the other and which are weaker and stronger than the mean stimulus, respectively. At the beginning, $t = 0$, the energy $E(0)$ is almost flat. After some stimuli have been delivered, for instance at $t = t_0$, the energy has a definable minimum emerged thorough Hebbian learning. If a stimulus is then delivered (violet) this induces a change in the energy, which we illustrate in the energy profile $E(t_0 + \tau)$ after some time τ. The same occurs with consequent stimuli (orange). The movement of the attractor leads to the history-dependence of the choice. In the working memory experiment we are considering, if the second stimulus of the $(n-1)$th trial was strong, the attractor will have shifted towards higher stimulus intensities, so the memory trace of the first stimulus in the next trial will relax towards the new attractor and might "overtake" the second stimulus strength $S_B = 80$, leading to the patterns of misclassification we saw in fig. 2G, where the accuracy of the correct choice for $(70, 80)$ is high if $S_B^{(n-1)}$ was weak and accuracy declines for progressively stronger values of $S_B^{(n-1)}$. If instead the second stimulus of the $(n-1)$th trial was weak, the attractor will have shifted towards lower stimulus intensities, facilitating the correct classification of $S_B^{(n)} > S_A^{(n)}$.

Reference memory

In this section, we show that the concurrent dynamical processes in the RNN model – the movement of the attractor and the relaxation of the RNN representation towards it – can also account for the repulsive effects observed in a reference memory task in humans and rats, without any tuning of the hyper-parameters $\gamma, \kappa, a, \lambda$. In particular, we show that the serial dependence of choices in *contraction* bias is (perhaps counter-intuitively) an instance of *repulsive* effects.

We take experimental data from the reference memory experiment of Hachen *et al.* [8], where in single sessions the rat was tested on a series of trials, separated by ITI's that were self-timed by the subject. In each trial, the rat was tasked with reporting by its choice of left or right reward spout whether a single stimulus was weak or strong, fig. 1B. The threshold θ which separates the weak and strong categories was chosen by the experimenter, and the rat received a reward for correct choices. Hachen *et al.* [8] showed that the internal representation of θ was however primarily influenced by the distribution of stimuli presented, rather than reward contingency. Motivated by this, we simulate the task by driving the RNN using the stimulus distributions akin to the experimental one and we read out the model's choice at the stimulus offset, without modelling reward: if the peak of the overlap profile of the RNN relaxes towards higher intensities, the delivered stimulus is considered weak, whereas if the peak relaxes towards lower intensities, the delivered stimulus is considered strong, see fig. 3D&E.

The importance of learning at the beginning of an experiment Hachen *et al.* [8] found that rats internally construct and apply a representation of θ to categorise stimuli over successive trials from session onset. The rats did not retain this boundary between sessions: the psychometric curves of the animals derived from the first three trials of each session deviates only marginally from chance level across all stimulus intensities, fig. 4A. This poor performance at the onset of each session highlights the importance of learning the network connectivity *dynamically* for any neural network model of this behaviour. In our RNN model, an initial learning phase is necessary to form an attractor that is correlated with the inputs. Consequently, after only the first three stimuli, neuronal configurations assume states that are not yet correlated with the effective inputs, resulting in decision probabilities closer to chance level, as shown in fig. 4E.

Repulsive effects via different stimulus distributions The repulsive bias in the reference memory task can be seen in two ways. Figure 4B, adapted from Hachen *et al.* [8] , plots the psychometric curves

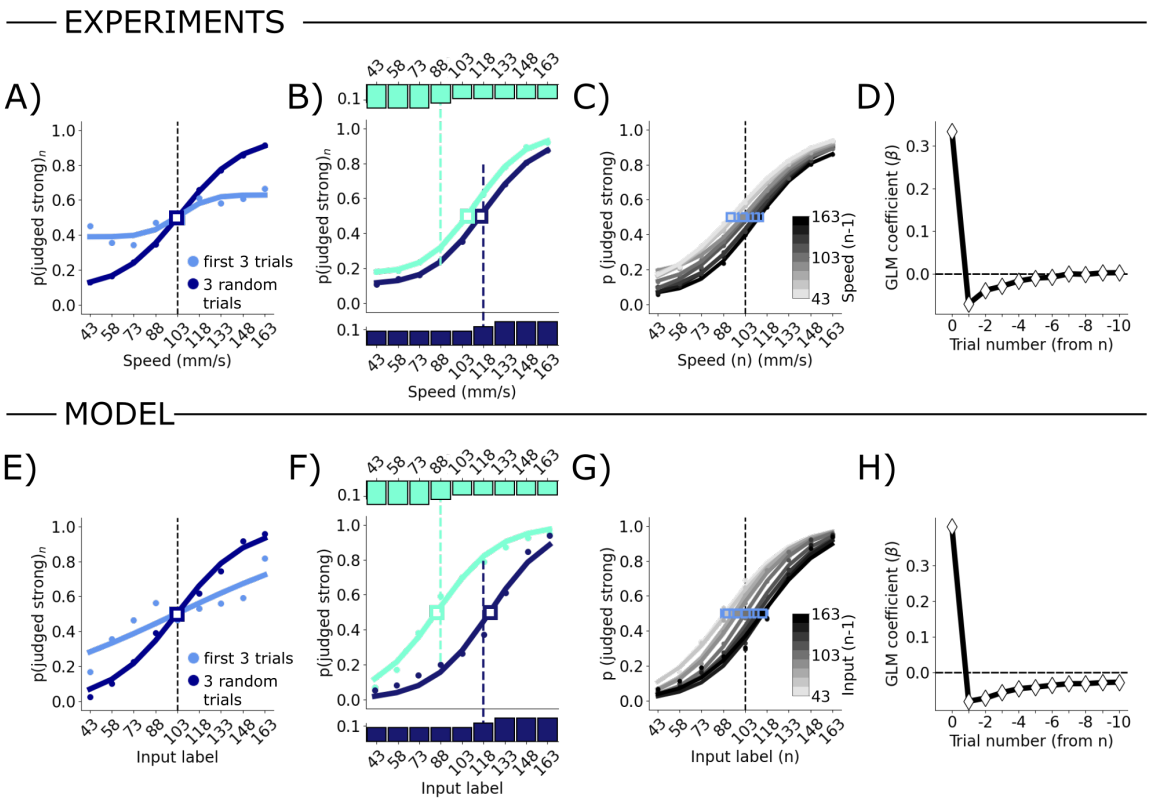

Figure 4: **Psychophysical and RNN results in a reference memory task.** In **A-D)**, the experimental data from Hachen *et al.* [8] are replotted. **A)** Psychometric curves obtained from the first three trials of each session (light blue) and from three randomly selected trials per session (dark blue). The dashed vertical lines in A-C) mark the category boundary. The point of subjective equality (PSE) is denoted by a square. **B)** Psychometric curves for sessions where input distributions had an average stimulus intensity of 88 (light blue histogram) and 118 (dark blue histogram). The input stimuli "attract" the psychometric curves towards the mean of the distribution. **C)** The probability of categorising a stimulus as "strong" conditioned on the intensity of the previous stimulus. Blue squares represent the PSE. **D)** Coefficient values β derived from the generalised linear model (GLM) predicting choice given the stimulus intensities of the ten preceding trials. **E-H)** show results from the RNN under conditions mimicking those of the experiments of Hachen *et al.* [8].

of rats that were exposed to stimulus distributions biased towards weak and strong stimuli (small light and dark blue histograms, respectively). The psychometric curves of fig. 4B show that an increased frequency of strong stimuli shifts the point of subjective equality towards stronger stimuli; likewise, an increased frequency of weak stimuli shifts the point of subjective equality towards weaker stimuli. This does not occur if the distribution of the inputs is kept uniform but the reward is delivered to bias the category boundary [8]. It is as if experiencing many strong stimuli pushes down the representations of stimuli of intermediate strength, leading the rats to underestimate their intensity, and vice versa. We observe the same bias in the RNN model as a consequence of Hebbian plasticity: a biased distribution of stimuli, where high-intensity values are given more frequently, shifts the location of the attractor towards high intensities (on average). This translates into a higher probability of categorising a given stimulus as weak, as shown in fig. 4F. The opposite holds for a stimulus distribution biased towards higher intensities. In appendix C, we discuss how the disparity in the quantitative separation between psychometric curves in the experiment and the model possibly arises due to a small sample effect in the experiments.

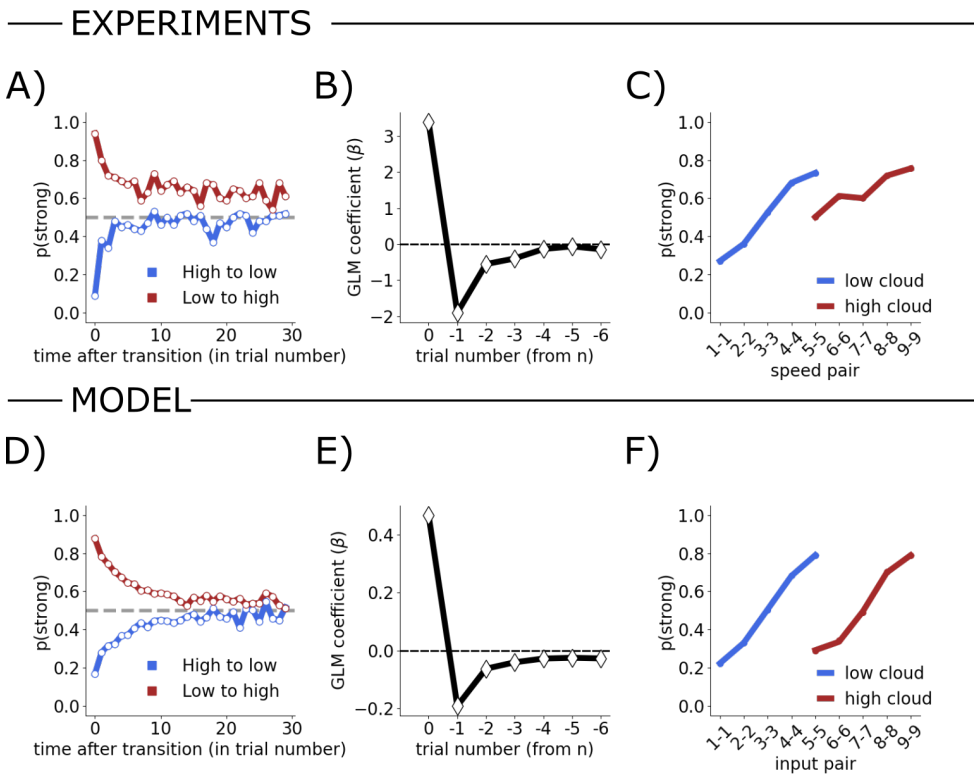

Figure 5: **The one-back task highlights the plasticity of the attractor as it moves between stimulus clouds. A)** Probability of human subjects categorising a stimulus as stronger than the preceding stimulus within each cloud as a function of the time elapsed since the transition between clouds. **B)** Coefficient values (β) of the generalised linear model (GLM) of the choice at trial n given the six preceding stimuli. **C)** Probability of categorising $S^{(n+1)} > S^{(n)}$ when $S^{(n)} = S^{(n-1)}$. If the pair $5, 5$ is preceded and followed by stimuli from the low cloud, it is considered part of the low cloud; and vice versa for the high cloud. **D)** to **F)** Results obtained from the RNN model.

Repulsive effects in the history-dependence The second way of showing repulsion is by considering the effect of the $(n-1)$th stimulus on the judgement of the nth stimulus. The psychometric curves conditioned on the $(n-1)$th stimulus in fig. 4C show clearly that the point of subjective equality of the nth stimulus is pushed up to stronger stimuli after experiencing a strong stimulus at the $(n-1)$th trial, meaning that the strength of intermediate stimuli is underestimated. Likewise, the probability of classifying a stimulus as strong *decreases* if stimulus in the $(n-1)$th trial is strong. The same behaviour is captured by the model as seen in fig. 4G: each stimulus moves the location of the attractor a little bit due to Hebbian plasticity, cf. fig. 3F. This movement of the attractor can then cause misclassification of subsequent stimuli near the average stimulus strength, resulting in a general shift of the psychometric curves. We thus see explicitly that, perhaps counter-intuitively, recency effects in working memory and repulsive effects in reference memory may have the same origin. Finally, we show a GLM analysis of the choice at trial n based on the stimulus strength at previous trials. The importance of previous stimuli decays and there is a large negative weight on the $(n-1)$th stimulus both in the experimental results and in the model, as reported in fig. 4D and H.

The one-back experimental paradigm – plasticity and relaxation at a glance

Our analysis highlighted the importance of the movement of the attractor in the RNN for recency and repulsive effects to emerge in perception. To challenge the generality of the RNN model and, more specifically, to investigate the significance of the sliding of the attractor, we designed a novel **one-back task** which tests for the movement of the attractor as follows. Participants are presented with a sequence

of trials where they receive a single stimulus, similar to the reference memory task. However, triggered by the go cue, they must compare the current trial's stimulus not to a fixed category boundary but to the stimulus of the previous trial. Thus, akin to the working memory task, each stimulus must be compared to the preceding stimulus. Differently from the working memory task, after each judgement the comparison stimulus becomes the base stimulus for the next trial. stronger or weaker than the previous one, akin to the working memory task, see fig. 1C. (If the same stimulus intensity is presented on two consecutive trials, the correctness of the comparison is rewarded randomly.) A crucial feature of the paradigm is that stimulus intensities are not drawn in random order. Instead, the nine equidistant stimulus intensities are grouped into two clouds, a "low" cloud composed of stimuli of intensities labelled from $1-4$ and a "high" cloud composed of stimuli labelled from $6-9$. After a stimulus from the low/high cloud, the next stimulus is either drawn from the same cloud, or else is the intermediate stimulus, 5. From intensity 5, the next stimulus may either transition to the opposing cloud or else "bounce back" to the cloud of the preceding stimulus sequence. The transition probabilities $p(S^{(n)}|S^{(n-1)})$, shown in fig. 1C, can lead to long sequences of stimuli constrained to one of the clouds. This paradigm is thus designed to test the prediction that subjects' psychophysical results will reflect the attractor's continuous movement between two recency weighted averages, "low" and "high".

The task was presented to $n = 24$ participants. fig. 5A shows the probability that a given stimulus was judged as stronger than the previous one as a function of time after a jump between clouds. Within each cloud, the $(n+1)$th stimulus is equally likely to be weaker or stronger than the previous one, but after stimulus intensities switch from the high- to the low-intensity cloud, participants consistently underestimated the strength of the first few stimuli (blue line), before stabilising around the correct value of 0.5. Conversely, after switching from low- to high-intensity stimuli, stimulus intensities were overestimated initially (red). The RNN model generates analogous behaviour, with no external re-setting of the hyper-parameters γ, λ, η, g used previously for working memory and reference memory simulations, see fig. 5D: the memory trace of all the stimuli after the switch tend to converge to the attractor which is at the opposite end of the stimulus strength, leading to misclassifications. At the same time, repeated exposure to weaker/stronger stimuli moves the attractor slowly towards the respective cloud. Averaging over time, the overall influence of previous stimuli on the choice can be estimated in terms of GLM coefficients as done in the previous two paradigms, showing similar results: the choice at trial n depends with a decaying weight on the previous stimuli as found in both experimental and model data in fig. 5B) and E).

A bi-modal contraction bias The one-back paradigm highlights that the contraction of memory traces, which is responsible for contraction bias, does not occur towards the global average of stimulus intensities, which would be the centre of the stimulus range, intensity 5; instead, memory traces contract towards an intensity value that changes in time. The contraction bias in the one-back task is most evident considering choices of participants after experiencing the same input twice, $S^{(n)} = S^{(n-1)}$. The probability of categorising $S^{(n)} > S^{(n-1)}$ deviates from 0.5 in a discontinuous trend, shown in fig. 5C for human subjects. For instance, within a stream of low-intensity stimuli, the second presentation of stimulus of intensity 5 is consistently categorised as stronger than the previous stimulus of intensity 5, while the opposite is true within a stream of high-intensity stimuli. The memory trace of intensity 5 contracts towards a *lower* value when it is part of a low-intensity sequence, and vice versa, consistently biasing the categorisation of the second instance of intensity 5. The RNN model fully reproduces this behaviour, see fig. 5F. Indeed, during the ITI the internal representation of the stimulus relaxes towards the running attractor, replicating the detailed temporal dynamics of actual human subjects. We finally note that the asymmetry in the results for stimuli from the high vs. the low cloud in fig. 5C, as well as the decay of the red curve to a value larger than 0.5 in fig. 5A, arise from a combination of the Weber-Fechner law and the precise stimulation protocol, as we discuss in appendix D.

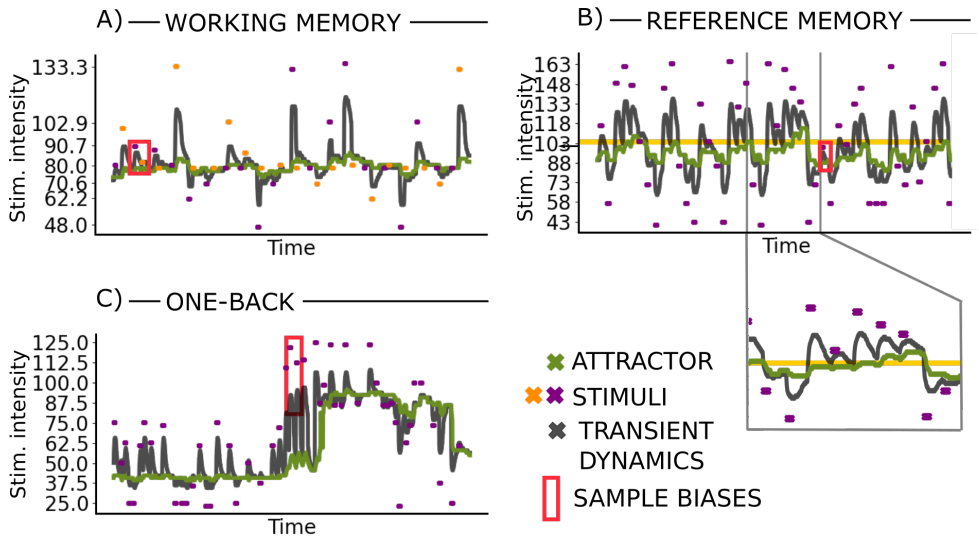

Figure 6: **One neural network reproduces two different biases across three different paradigms:** Sample evolution of the network in a sample working (A), reference (B), one-back (C) task. The external stimulus is depicted in violet (and in orange when corresponding to S_B in B), the internal representation (dark gray) and attractor position (green) at each time step is readout through the maximum of the overlap profile. The red rectangles indicate instances of incorrect choices while the orange horizontal line in Q) represent the overall average.

Discussion

We propose a simple recurrent neural network model with ongoing Hebbian plasticity that reproduces two perceptual biases across three different experimental paradigms without any per-paradigm fine-tuning. Combining Hebbian plasticity with a set of biologically inspired input patterns yields a dynamic point attractor, which embodies the representation of an attractor along the dimension of stimulus intensity. Contraction arises from relaxation towards this plastic attractor, whose location is continuously modified by Hebbian plasticity, leading to both repulsion and recency effects. Additionally, we designed a new behavioural paradigm, the one-back task, that accentuates the movement of this plastic attractor, and found that the RNN model correctly predicted the movement of the attractor with a bimodal stimulus distribution.

Figure 6 reports the intensity values encoded by the transient dynamics and by the attractors in three sample simulations, one per paradigm. While the only difference across simulations is the input distribution, the different choice statistics and the perceptual biases observed in the experimental paradigms are reproduced. The red rectangles highlight wrong choices that reveal the perceptual biases inherent to each paradigm. There are two key timescales in our model: one governs the relaxation of the memory trace of a stimulus after stimulation (black in Figure 6), while the other governs the movement of the plastic attractor (green in Figure 6). While the relaxation occurs on a shorter timescale than the movement of the attractor, the two timescales are comparable, both evolve in time and are only indirectly fixed by the hyper-parameters of the model. A phenomenological model including two different timescales and attractors was proposed by Hachen *et al.* [8] and further developed by Darani *et al.* [14] and Diamond & Toso [15]. Such a model accounts for the existence of different perceptual biases but does not address the neural mechanism underlying their emergence. A neural network model of two tuned static interacting continuous attractors was recently proposed by Boboeva *et al.* [18] to account for contraction bias and its recency characteristics in a working-memory paradigm. Crucially, in their model the connectivity of the networks is imposed through two interaction kernels and then

fixed; in our model, continual reshaping of the connectivity by Hebbian plasticity is key for reproducing perceptual biases across different paradigms, with each have their own stimulus distributions, and for reproducing the different behaviours of subjects in the early vs. later parts of each reference memory session, c.f. fig. 4 A) and E).

Akrami *et al.* [3] demonstrated contraction bias and its history-dependence in auditory and tactile working memory paradigms. Furthermore, they showed that inactivation of the PPC subregion led to a decrease in contraction bias, despite the preservation of memory capabilities, as reproduced by the theoretical model of Boboeva *et al.* [18]. We found that the RNN model is also able to replicate the behaviour of rats after PPC inactivation by simply adjusting the synaptic homeostasis, see supplementary fig. B.1. Changing this hyper-parameter effectively places the model in a regime in which the attractor tends to remain stuck at the intensity of the last perceived stimulus. Thus, the memory capacity is fully maintained, but the effect of contraction bias decreases since there is no relaxation of the memory trace during the inter stimulus interval towards an attractor slowly built by previous experiences. Our model thus opens the possibility for an alternative role of the PPC sub-network in contraction bias.

While the importance of attractors as a mechanism occurring in the brain to code intensity has long been contemplated in the field [8, 17, 18], our model shows how these attractors could be learnt continuously from the input distribution using a local learning rule. Our study highlights the potential significance of *transient* dynamics in encoding and processing information in recurrent neural networks. The mathematical analysis of an RNN with ongoing Hebbian plasticity driven by intermittent, correlated inputs is a significant challenge. Previous work proving the emergence of a unique attractor after one-shot learning of *binary* morphing patterns [31], and the recent work by Clark & Abbott [33] extending dynamical mean-field theory to coupled neuronal-synaptic dynamics without external driving, represent excellent starting points. More generally, it will be interesting to analyse the geometry of the representations learnt by the RNN [34–36] and to explore the extension of this model beyond sensory perception.

Methods

Experimental protocol for working memory task

16 subjects participated in the WM task. All subjects performed one WM session, lasting approximately one and a half hour for a total of 1000 trials. All subjects were volunteers and were paid after the participation, on the basis of how well they performed. The study protocol conformed to international norms and was approved by the Ethics Committee of the International School for Advanced Studies (SISSA).

All experiments took place in the SENSEx Lab at SISSA. The subjects sat in front of a PC screen with their right arm resting on a pillow upon the desk, using their right index fingertip to trigger an infrared sensor in order to begin a trial; when doing so, their fingertip would be in contact with a plastic probe. The probe was attached to a shaker motor (type 4808; Bruel and Kjaer), producing vibrations in the horizontal direction, perpendicular to the fingertip. Subjects' responses were reported by pressing one of two buttons with the right hand. The experiment was automatised via LabVIEW (National Instruments).

Each stimulus was a noisy vibration, obtained by stringing together randomly sampled velocities. Velocities were obtained by sampling a normal distribution with 0 mean and defined by the standard deviation σ, ranging from 3 to 20; the mean stimulus speed is directly proportional to σ, and the motor amplifier gain was set so that the average stimulus speed would be 10 times the value of σ, measured in mm/s. The sequence of velocities formed one seed. There were 50 different seeds for each vibration mean speed, making the signature of a given vibration hard to recognise and making the mean speed a more salient feature.

The subject's finger had to remain in the sensor for the entire trial, comprising: pre- stimulus delay (500ms), base stimulus (334ms), inter-stimulus delay (either 2, 4 or 8s), comparison stimulus (334ms), post-stimulus delay (500ms); after this sequence elapsed, a visual go cue signalled to the subject to report

their choice. If the subject lost contact with the sensor before the go cue, the trial aborted, and a new one had to be initiated.

Nominal base and comparison pair difficulty does not depend on the value of the stimuli, but on the relative difference between them: the closer they were, the more difficult the discrimination. We expressed this difficulty using an index, the Normalised Speed Difference (NSD) (1), according to Weber's Law:

The stimuli pairs were arranged in two sets: a vertical set, in which the base stimulus σ was fixed at 8, and a horizontal set, in which the comparison stimulus σ was fixed at 8. For both sets, NSD values ranged from -0.25 to 0.25, including 0. Stimuli pairs from both sets were randomly presented during the WM sessions.

Experimental protocol for one-back task

Twenty-four human subjects (ages 19-38), recruited from the online research participation system (SONA), participated in the One Back Experiment. The study adhered to frame of rules specified by international norms for human behaviour experiments and was approved by the Ethics Committee of the International School for Advanced Studies (SISSA).

The experiment was conducted in Human labs at SENSEx Lab in SISSA. The subjects were seated in a dimly lit room. The experiment was executed in LabView software. Subjects viewed a monitor screen for cues and feedback related to the task. They wore headphones that presented white noise and eliminated any external sounds from the environment or motor. The vibrotactile stimuli were delivered to the tip of their right-hand index finger through a shaker motor (type 4808; Bruel and Kiaer). The responses were recorded using keyboard button press using left hand. The subjects saw a blue dot as fixation on the screen. They received feedback (green/red dot for correct /incorrect) on each trial through the screen.

The subjects performed a one-back memory task where the goal was to compare stimulus n with n-1. The stimuli were generated by sampling a sequence of velocity values. The probability of velocity values was drawn from a normal distribution (μ = 0, SD = σ in mm/s). The gain of the motor was amplified 10 times. The velocity sequence was taken randomly from 50 seed values for a given trial. These were transmitted as voltage values to the shaker motor producing motion in horizontal direction using a plastic knob delivered perpendicular to the finger. The stimuli had a fixed duration of 350 ms. The judgement is based on the mean speed of the stimulus, which is subjectively reported by subjects to be "strength", "amplitude", or "intensity". The mean speed of the vibration was proportional to the SD values used. Nine SD values [2.5, 3.75, 5, 6.25, 7.5, 8.75, 10, 11.25 and 12.5] were used.

Each trial began with stimulus delivery followed by response time window (timed on the screen by a graphical motion of blue colour filling a vertical bar) followed by Inter-Trial Interval (ITI). The subjects had two seconds available to report their decision by pressing keys [A or D]. After every response, the feedback was displayed on the screen. Failure to respond within the given response time resulted in a timeout. Three ITI values [0, 2 and 6 sec] were used, resulting in total ITI of 2, 4 or 8 seconds between two consecutive stimuli.

The experiment lasted one hour and was divided into 3 blocks (250 trials each). Each subject performed a total of 750 trials. The average performance of subjects in One Back task is 76.63 percent.

RNN model

Motivation

The three psychophysical experiments that we reproduce with our model – reference memory, working memory, and one-back memory – are comprised of judgements made on vibratory stimuli delivered to the fingertip of human participants, or to the whiskers of the rats. As pertains to rats, the model may be considered a representation of the frontal cortical region known as vibrissal motor cortex (vM1), the main target of somatosensory cortex in rats [20]. vM1 is involved in memory, decision making, and motor planning and is a recurrent network that receives afferent inputs from other subareas of frontal

cortex, and is connected with posterior parietal cortex [20]. In humans, the corresponding cortical regions that fulfil these tasks are currently unknown.

Input structure

Our model for the inputs to the RNN units is inspired by physiological data from Arabzadeh *et al.* [29] and Fassihi *et al.* [30]. They found that the majority of vS1 neurons in behaving rats (the modelled input to the recurrent neural network) show monotonically increasing firing rates as vibration intensity increases, with heterogeneity across neurons in regard to sensitivity (minimum vibration to cause excitation) and slope. We set the ratio of positive versus negative slopes to 50% to account for the possible presence of interneurons. Half of the neurons receive inputs whose magnitude increases with stimulus strength (positive slope), while the other half receives inputs that decreases with stimulus strength. We provide additional details about the generation of input patterns used in our model and an analysis of their statistics in appendix A. Following the experimental setup, any time a stimulus intensity μ is delivered, we sample an input from a Gaussian distribution with mean μ and a standard variance σ^{μ} which increases with stimulus strength, see appendix A.

Simulations

All simulations were conducted in Python and are accessible online. We employed the following hyperparameters for eqs. (1) and (2): $\gamma = 0.1$, $\lambda = 0.014$, $\eta = 2$, and $g = 1.2$, $\kappa = N$ facilitating a homeostatic regime where learning and dynamic evolution proceed continuously without stagnation. Variations of these parameters are also feasible. We simulated an RNN with $N = 200$ units and parametrise the spectrum of weak to strong vibrational stimuli with $P = 100$ input patterns. For each input ξ^{μ}, we generated 5000 random numbers from a Gaussian distribution with a mean of zero and the corresponding standard deviation σ_x. This quantity of 5000 was chosen to approximate the average intensity perceived over approximately 500ms in the experimental setup (see Supp. Material A). We then calculated the average absolute value s_x and selected the input pattern $\mu = \text{int}(s_x P/\sigma_{\text{max}})$, where σ_{max} represents the highest standard deviation employed in the experiments. In alignment with the experimental results, where we typically reference the effective average velocity 10σ in the results, we maintain the same value in the model figures 2 and 4. We initialised the RNN weight matrix with random weights ranging from -1 to +1 at the beginning of each "experimental session", i.e. at the beginning of each run. The temporal parameters of the experimental protocol are configured as follows.

Working Memory Task 5 time steps before each stimulus, 5 time steps for stimulus presentation, followed by randomly selected intervals of 5, 10, or 20 time steps before the second 5-time step stimulus, followed by 10-15 time steps before the following trial.

Reference Memory Task 10 time steps before the stimulus, 10 time steps for stimulus presentation, and a randomised interval of between 10 to 20 time steps after the stimulus.

One-Back Task 10-time steps for stimulus presentation, followed by randomly selected intervals of 20, 40, or 80-time steps from one stimulus to the next. In figs. B.1 and E.1 we used a different stimulation protocol as reported in the capitation.

The specific standard deviation values and their respective extraction probabilities used in both the experimental sessions and the model are as follows:

- Working Memory Task: Standard deviation values (σ) were 4.8, 6.22222, 7.05888, 8, 9.06667, 10.28577, and 13.33333. The S_A and S_B couples were organised as depicted in fig. 2A)E), with each pair extracted with equal probability.

- Reference Memory Task: Standard deviation values (σ) were 4.3, 5.8, 7.3, 8.8, 10.3, 11.8, 13.3, and 14.8. Each value was extracted with equal probability, except in fig. 4 B)F), where biased extraction probabilities were indicated in the plot.

- One-Back Task: Standard deviation values (σ) were 2.5, 3.75, 5, 6.25, 7.5, 8.75, 10, 11.25, and 12.5. Extraction probabilities were as reported in fig. 1D).

Acknowledgements We thank Iacopo Hachen for valuable discussions and for providing the data used in fig. 4. We thank Alessandro Treves for valuable discussions on Blumenfeld *et al.* [31]. SG acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR "Future Artificial Intelligence Research". This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22].

Author contributions F.S, M.D. and S.G. designed the study. F.S. proposed the theoretical model, developed the simulations and performed the experimental-theoretical comparisons. D.G. performed the experiments on the Working Memory paradigm and Y.C. performed the experiments on the One-Back paradigm. F.S. and S.G. wrote the first draft of the paper and all authors contributed to and approved the final manuscript.

References

1. Ashourian, P. & Loewenstein, Y. Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one* **6,** e19551 (2011) (cit. on p. 1).
2. Fassihi, A., Akrami, A., Esmaeili, V. & Diamond, M. E. Tactile perception and working memory in rats and humans. *Proceedings of the National Academy of Sciences* **111,** 2331–2336 (2014) (cit. on pp. 1, 4).
3. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554,** 368–372 (2018) (cit. on pp. 1, 2, 4, 5, 12, 19, 21).
4. Tal-Perry, N. & Yuval-Greenberg, S. Contraction bias in temporal estimation. *Cognition* **229,** 105234 (2022) (cit. on p. 1).
5. Gibson, J. J. & Radner, M. Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of experimental psychology* **20,** 453 (1937) (cit. on p. 1).
6. Magnussen, S. & Johnsen, T. Temporal aspects of spatial adaptation. A study of the tilt aftereffect. *Vision research* **26,** 661–672 (1986) (cit. on p. 1).
7. Fritsche, M., Mostert, P. & de Lange, F. P. Opposite effects of recent history on perception and decision. *Current Biology* **27,** 590–595 (2017) (cit. on p. 1).
8. Hachen, I., Reinartz, S., Brasselet, R., Stroligo, A. & Diamond, M. Dynamics of history-dependent perceptual judgment. *Nature communications* **12,** 6036 (2021) (cit. on pp. 1, 2, 7, 8, 11, 12, 18, 20).
9. Hollingworth, H. L. The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods* **7,** 461–469 (1910) (cit. on pp. 1, 4).
10. Ashourian, P. & Loewenstein, Y. Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one* **6,** e19551 (2011) (cit. on p. 1).
11. Karim, M., Harris, J. A., Langdon, A. & Breakspear, M. The influence of prior experience and expected timing on vibrotactile discrimination. *Frontiers in neuroscience* **7,** 255 (2013) (cit. on p. 1).
12. Fritsche, M., Spaak, E. & De Lange, F. P. A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. *Elife* **9,** e55389 (2020) (cit. on p. 1).
13. Oña-Jodar, T. *et al.* NEURAL MECHANISMS UNDERLYING WORKING MEMORY ERRORS. *IBRO Neuroscience Reports* **15,** S672–S673 (2023) (cit. on p. 1).
14. Darani, Z. Y., Hachen, I. & Diamond, M. E. Dynamics of the judgment of tactile stimulus intensity. *Neuromorphic Computing and Engineering* **3,** 014014 (2023) (cit. on pp. 1, 11).

15. Diamond, M. E. & Toso, A. Tactile cognition in rodents. *Neuroscience & Biobehavioral Reviews,* 105161 (2023) (cit. on pp. 1, 11).

16. Hahn, M. & Wei, X.-X. A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nature Neuroscience,* 1–12 (2024) (cit. on pp. 1, 2).

17. Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307,** 1121–1124 (2005) (cit. on pp. 2, 4, 12).

18. Boboeva, V., Pezzotta, A., Clopath, C. & Akrami, A. From recency to central tendency biases in working memory: a unifying network model (2023) (cit. on pp. 2, 11, 12).

19. Lee, H. & Lee, S.-H. Boundary updating as a source of history effect on decision uncertainty. *Iscience* **26** (2023) (cit. on p. 2).

20. Schwarz, C. & Chakrabarti, S. in *Scholarpedia of Touch* 751–769 (Springer, 2015) (cit. on pp. 2, 13, 14).

21. Grossberg, S. On learning, information, lateral inhibition, and transmitters. *Mathematical Biosciences* **4,** 255–310 (1969) (cit. on p. 2).

22. Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences* **81,** 3088–3092 (1984) (cit. on p. 2).

23. Sompolinsky, H., Crisanti, A. & Sommers, H.-J. Chaos in random neural networks. *Physical review letters* **61,** 259 (1988) (cit. on p. 2).

24. Treves, A. Graded-response neurons and information encodings in autoassociative memories. *Physical Review A* **42,** 2418 (1990) (cit. on pp. 2, 6).

25. Schönsberg, F., Monasson, R. & Treves, A. Continuous quasi-attractors dissolve with too much-or too little-variability. *bioRxiv,* 2023–08 (2023) (cit. on p. 2).

26. Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63,** 544–557 (2009) (cit. on p. 2).

27. Hebb, D. *The organization of behavior: A neuropsychological theory* (John Whisking Wiley, 1949) (cit. on p. 2).

28. Oja, E. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology* **15,** 267–273 (1982) (cit. on p. 2).

29. Arabzadeh, E., Petersen, R. S. & Diamond, M. E. Encoding of whisker vibration by rat barrel cortex neurons: implications for texture discrimination. *Journal of Neuroscience* **23,** 9146–9154 (2003) (cit. on pp. 3, 14).

30. Fassihi, A., Akrami, A., Pulecchi, F., Schönfelder, V. & Diamond, M. E. Transformation of perception from sensory to motor cortex. *Current Biology* **27,** 1585–1596 (2017) (cit. on pp. 3, 14).

31. Blumenfeld, B., Preminger, S., Sagi, D. & Tsodyks, M. Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron* **52,** 383–394 (2006) (cit. on pp. 3, 12, 15).

32. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters* **55,** 1530 (1985) (cit. on p. 6).

33. Clark, D. G. & Abbott, L. Theory of coupled neuronal-synaptic dynamics. *arXiv preprint arXiv:2302.08985* (2023) (cit. on p. 12).

34. Bernardi, S. *et al.* The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183,** 954–967 (2020) (cit. on p. 12).

35. Chung, S. & Abbott, L. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology* **70,** 137–144 (2021) (cit. on p. 12).

36. Nogueira, R., Rodgers, C. C., Bruno, R. M. & Fusi, S. The geometry of cortical representations of touch in rodents. *Nature Neuroscience* **26,** 239–250 (2023) (cit. on p. 12).

37. Fechner, G. T. *Elemente der psychophysik* (Breitkopf u. Härtel, 1860) (cit. on p. 20).

Appendix

A Input distribution

Here, we provide additional details about the generation of input patterns utilised in our model and an analysis of their statistics. The inputs are represented by a matrix $\bar{\boldsymbol{\xi}}$ of size $N \times P$, where each row $\vec{\xi_i}$ denotes the input profile received by neuron i for every intensity value. Each column $\vec{\xi^\mu}$ represents the pattern of activity delivered to the network when a specific input μ is present, with ξ_i^μ denoting the intensity input received by neuron i for input μ. To discretize intensity values, we divide the range from a minimum s^{MIN} to a maximum s^{MAX} into P bins, each coding for the intensity $((S^{MAX} - S^{MIN})/P)$. To generate the input patterns, we employ the following procedure: we begin with a vector of length N containing random numbers ranging from -1 to $+1$; we create a second vector by replacing each negative entry in the first vector with a random number between 0 and 1, and each positive entry with a random number between -1 and 0; we linearly interpolate $P-2$ patterns from the first to the last vector, ensuring continuity between intensity levels; finally, we set any negative values to zero to maintain non-negativity in the generated patterns. Through this process, we obtain a set of input patterns ranging from ξ_i^0 to ξ_i^P. The pattern statistics can thus be summarised as follows: Approximately 50% of units receive positively coded inputs, with the remaining units receiving no input for each intensity (fig. A.1A, displaying 50 sample profiles $\vec{\xi_i}$ with different colors). The average input pattern activity exhibits a slight decay towards $\xi^{P/2}$ before increasing again (fig. A.1 B). The fraction of units receiving a non-zero input remains roughly constant across patterns (fig. A.1 C).The correlation structure between patterns exhibits a monotonic shape, which explains why the dynamics observed in the network after progressive Hebbian learning can be likened to a bell shape when considering the overlap with the inputs. In fig. A.1 D, we show, in a sample with $P = 50$, the overlap $\vec{\xi^{20}}$ (light gray), $\vec{\xi^{25}}$ (black), and $\vec{\xi^{30}}$ (dark gray) with all other patterns.
Based on the stimulation protocol in the three different tasks, the input $\xi(t)$ in eq. (1) is defined as follows:

$$\begin{cases} \xi(t) = \vec{\xi^\mu} & \text{if a stimulus } \mu \text{ is present} \\ \xi(t) = \vec{0} & \text{otherwise} \end{cases} \tag{A.1}$$

Two stochastic components determine which specific stimulus is delivered at each time:

1. The **stimuluation protocol**, defined experimentally, dictates the stimuli delivered and the probabilities of delivering each of them (see methods).

2. The **stimulation setup**.

   - **Experimentally**,inputs are delivered through a motor as a speed of vibration in a stochastic manner. For each specific "average speed" $\bar{v}_x$, there corresponds a certain standard deviation σ_x. The effective velocity v delivered to the subject changes every 0.1 ms (10 kHz), drawn from a Gaussian distribution with standard deviation σ_x. This is illustrated in the distributions for the stimuli set of the reference memory paradigm studied in [8] in fig. A.1 E. Negative values are converted to positive ones, resulting in a distribution with an effective mean $m(v_x) = \sqrt{2/\pi}\sigma_x$ (vertical line in fig. A.1 F). Through a motor-gain parameter y, the effective average velocity is then converted to $\bar{v}_x = ym(v_x) = 10\sigma_x$. An example distribution of inputs observed in one session of [8] is illustrated in fig. A.1 G).

   - **In the model**, we replicate the same bias in the distribution of the stimuli as follows: whenever an input x is delivered, we compute the average of 5000 random numbers from the respective distribution, we estimate the average μ and delver the corresponding pattern $\vec{\xi^\mu}$ (vertical line in fig. A.1 H).

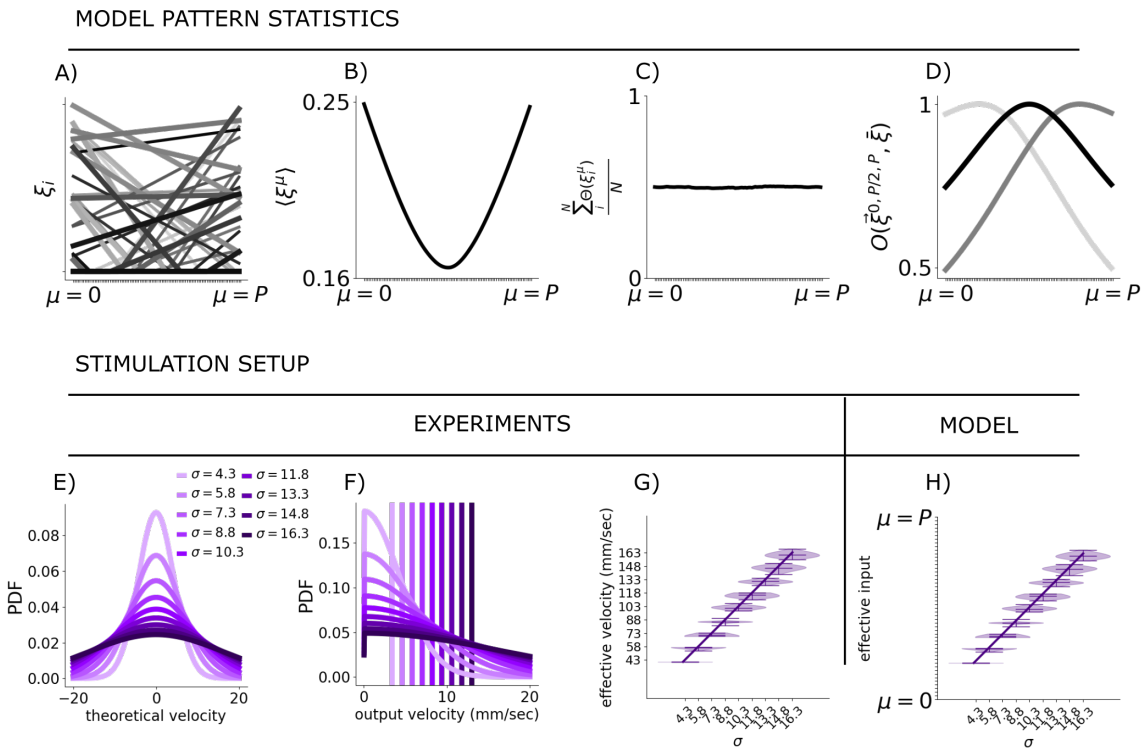

Figure A.1: **Pattern Statistics and Stimulation Setup:** A) Fifty sample input profiles ξ_i (different colours) depicted for different intensities. B) Average activity across input profiles averaged over 100 realisations. C) Fraction of active units in the input profiles averaged over 100 realisations. D) Normalized overlap (cosine similarity) between $\vec{\xi}^{20}$ (light gray), $\vec{\xi}^{25}$ (black), $\vec{\xi}^{30}$ (dark gray), and all other patterns. E) Gaussian distribution with mean zero and standard deviation as labelled. F) Same as E), but showing absolute values; the vertical line indicates the effective average activity. G) Violin plot representing the actual average obtained by sampling 5000 random numbers from the distributions in F. H) Identical to G), but with the input intensities discretized into P bins, illustrating the transition between experiments and model.

B Working memory paradigm in Akrami *et al.* [3]

We utilise the same model as described in the main text, including the same parameters, but with stimuli and inter-stimulus interval distributions similar to those presented in Akrami et al. (2018). Our simulations reproduce the observed contraction bias effects (fig. B.1 A), the repulsive effect of previous stimulus pairs (fig. B.1 B), and the dependence of contraction bias on inter-stimulus intervals (fig. B.1 C) (to be compared with Fig. 1B, Fig.2A-C, Fig. 1D of [3] respectively). Furthermore, Akrami *et al.* [3] showed that inactivating the posterior parietal cortex (PPC) leads to a reduction in contraction bias while maintaining memory capability. In our model, which normally operates in a homeostatic regime where learning and transient neural evolution occur continuously, we observe a similar effect if we modify the synaptic homeostatic parameter λ in eq. (2). This adjustment leads to a regime where synapses quickly forget previous stimuli. As a result, the internal representation in the network tends to become stuck to the last perceived stimulus, which effectively becomes the new fixed point on average. The attractor is thus very closed to the last perceived stimulus and the drift of the memory trace during the ISI is minimal, consequently diminishing contraction bias while preserving memory until the next stimulus, see fig. B.1 D (to be compared with Fig. 3C of [3]).

C Asymmetric shift of psychometric curves in Hachen *et al.* [8]

Figure 4B) in the main text reproduces a plot of Hachen *et al.* [8]. The reported shift in the psychometric curves is asymmetric: the psychometric curves obtained when more high stimuli were delivered during the session (dark blue) has a point of subjective equality (square) corresponding to the new average; instead, the point of subjective equality of the psychometric curves obtained from these sessions where more low stimuli were delivered (light blue), do not coincide to the new average and is actually even slightly shifted to the right as compared to the PSE for uniform distributions. On the contrary, the result of the model is a symmetric shift, as reported in fig. 4 F). We asked further details to the authors of Hachen *et al.* [8]. First they reinforced that the major result reported by their figure is the effective separation between the psychometric curves, which we reproduce, more than the quantitative distance of this shift. Second, they reported that the seemingly absence of movement of the light blue psychometric curve, with respect to an hypothetical uniform distribution of stimuli, is attributed to the specific subset of subjects which were considered for these experiments and which showed, by chance, a strong tendency of underestimation, i.e. of psychometric curves with PSE shifted towards the right. Thus, it remain open the question on whether a larger subject pool would actually lead to a symmetric shift in the psychometric curves as that we predict through the model. Note that in the model results we can use a large sample size, thus obtaining smooth results, but if we would focus only on about 10 picked seeds which give rise to right-shifted psychometric we would also reproduce an asymmetric shift. Otherwise, we hypothesise that if even with a larger pool the shift would keep being asymmetric another component of the model which could contribute in reproducing the behaviour is the distribution of the input pattern profiles. Indeed, in all results we presented so far, a ratio of 50% is considered between positively and negatively coding stimuli. Changing this percentage may have some effect of this type on the learned patterns.

D Performance asymmetries in the one-back task

We note that the asymmetry in the results for stimuli from the high vs. the low cloud in fig. 5C), as well as the decay of the red curve to a value larger than 0.5 in fig. 5A, arise from two independent factors. Firstly, the Weber-Fechner law [37] which suggests that the subjective sensation of stimulus intensity is proportional to its logarithm. This phenomenon facilitates the comparison of two equally distanced quantities when they are small compared to when they are large. Secondly, the specific stimulation setup generates a broader distribution of inputs when stimuli are high, cf. fig. A.1. In our model, we can eliminate this second stochastic component, thereby rectifying the asymmetries in the results, as shown in fig. E.1. This allows us to estimate the specific effect of Weber-Fechner through subtraction.

E The effect of the stochasticity in the experimental setup

In our one-back experimental results, we notice an asymmetric response when the stream of stimuli is presented in the low vs high cloud. This asymmetry arises from both the well-known Weber-Fechner's effect [37] and the specific characteristics of the stimulation setup used in our experiments. In the model, we have the capability to eliminate the stochasticity inherent in the inputs, exposing the network solely to the average stimuli without considering the positive Gaussian distribution. By doing so, we effectively remove the asymmetry observed in the results, as depicted in fig. E.1.

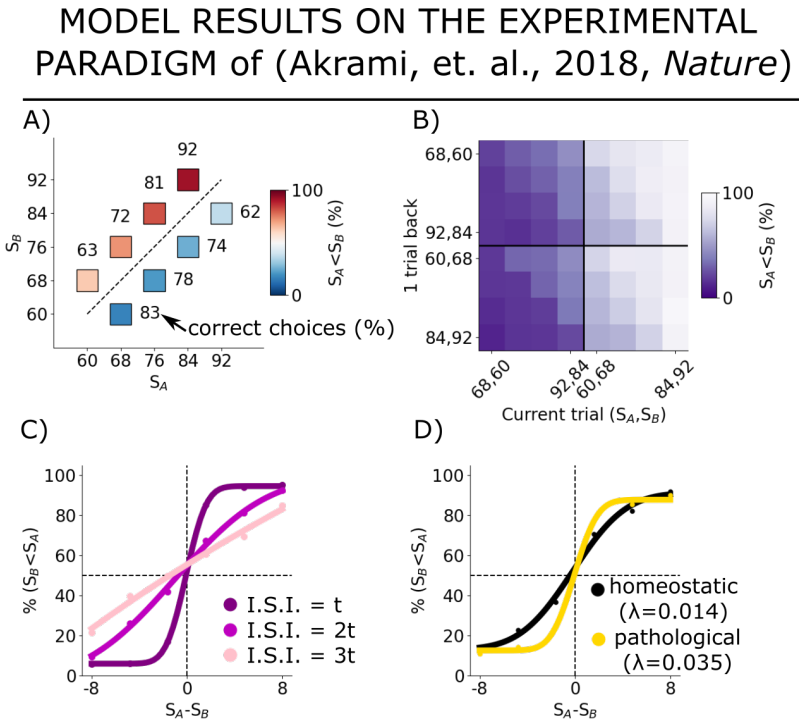

Figure B.1: **The RNN model replicates contraction bias and history-dependence demonstrated by Akrami *et al.* [3].** A) Probability of correct choices for 8 stimulus pairs (S_A, S_B). Variations in the probability of correct choice among different pairs illustrate the presence of contraction bias, aligning with results depicted in Fig. 1B of [3]. B) Probability of correct choice influenced by the stimulus pair presented in the previous trial, showcasing a trend where lower intensity pairs result in a higher probability of S_B being chosen stronger than S_A. This mirrors the findings in Fig. 2A-C of [3]. C) Performance across a subset of 6 stimulus pairs, each with a constant $S_B = 76$ and S_A ranging from 66 to 84. Results demonstrate the impact of ISI intervals on performance, with shorter intervals leading to better performance, consistent with Fig. 1E of [3]. D) Performance variations with changes in the synaptic homeostasis parameter λ. Increasing λ induces higher performance by accelerating synaptic forgetting of previous stimuli and facilitating faster attractor updates. This pathological regime replicates effects observed during the optogenetic inactivation of PPC, akin to Fig. 3C of [3]. Model parameters: Each stimulus is preceded by 5 time steps, followed by a presentation lasting 10 time steps. Subsequent to the initial stimulus, random intervals of 15, 30, or 45 time steps occur before the second 10 steps stimulus, with a random interval between 10-15 time steps between subsequent trials. Network dimensions: A) and B) $N = 204$, $P = 102$; C) and D) $N = 196$, $P = 98$. Hyperparameters γ, λ, η, and g remain consistent with the main model, except for D) (yellow), where $\lambda = 0.035$.

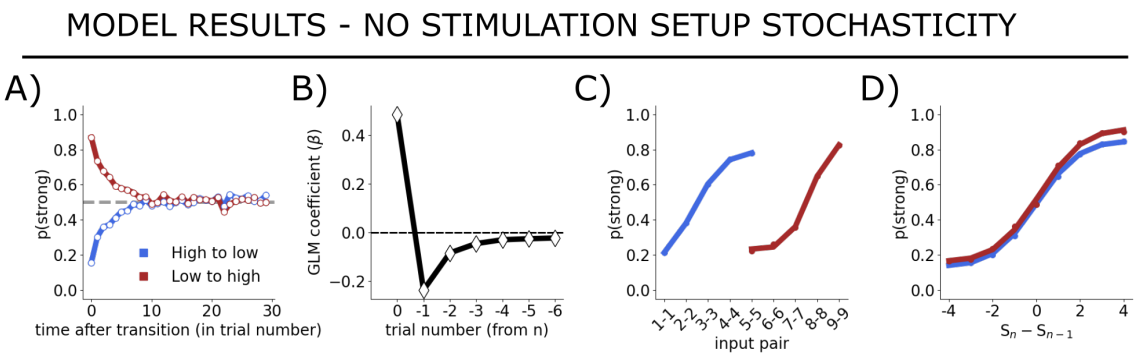

Figure E.1: **One-back Model without Input Stochasticity:** The model and figures remain consistent with those reported in fig. 5E-H. The difference lies in the effective inputs, where the stochastic component of the stimulation setup is removed.