PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Learning efficient representations of environmental priors in working memory

Tahra L. Eissa[1]*, **Zachary P. Kilpatrick**[1,2]

1 Department of Applied Mathematics, University of Colorado Boulder, Boulder, Colorado, United States of America, **2** Institute of Cognitive Science, University of Colorado Boulder, Boulder, Colorado, United States of America

* tahra.eissa@colorado.edu

Citation: Eissa TL, Kilpatrick ZP (2023) Learning efficient representations of environmental priors in working memory. PLoS Comput Biol 19(11): e1011622. https://doi.org/10.1371/journal.pcbi.1011622

Editor: Thomas Serre, Brown University, UNITED STATES

Received: September 22, 2022

Accepted: October 20, 2023

Published: November 9, 2023

Data Availability Statement: The data used to generate the figures and code developed for all proposed models has been made available on github.com/teissa/HeterogeneousWorkingMemory.

Funding: This work was funded by a Collaborative Research in Computational Neuroscience grant NSF-DMS-2207700 (ZPK), and a BRAIN Initiative grant NIH-R01EB029847(ZPK, TLE- salary received). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Experience shapes our expectations and helps us learn the structure of the environment. Inference models render such learning as a gradual refinement of the observer's estimate of the environmental prior. For instance, when retaining an estimate of an object's features in working memory, learned priors may bias the estimate in the direction of common feature values. Humans display such biases when retaining color estimates on short time intervals. We propose that these systematic biases emerge from modulation of synaptic connectivity in a neural circuit based on the experienced stimulus history, shaping the persistent and collective neural activity that encodes the stimulus estimate. Resulting neural activity attractors are aligned to common stimulus values. Using recently published human response data from a delayed-estimation task in which stimuli (colors) were drawn from a heterogeneous distribution that did not necessarily correspond with reported population biases, we confirm that most subjects' response distributions are better described by experience-dependent learning models than by models with fixed biases. This work suggests systematic limitations in working memory reflect efficient representations of inferred environmental structure, providing new insights into how humans integrate environmental knowledge into their cognitive strategies.

Author summary

Working memory is known to play an important role in cognition, allowing us to maintain information in our memory for short periods without a constant stimuli. However, humans display limitations in working memory, such as recalling certain stimuli more frequently and accurately than others. We propose that these recall biases are based on our experience of common stimuli in our environment and driven by goal of efficiently reducing error by remembering common stimuli with more accuracy than rare stimuli. Here, we develop a model that updates an observer's beliefs about the statistics of stimuli in an environment based on experience, biasing working memory recall such that common stimuli are remembered better. We then show that most human subjects' responses from a previously published working memory task are better matched to a model that learns in an experience-dependent way compared to models with fixed biases. Finally, we

Competing interests: The authors have declared that no competing interests exist.

identify a plausible neural mechanism for environmental experience-updating to show how the brain could implement this efficient strategy.

Introduction

Traditional descriptions of working memory, a core feature of cognition [1], conceive of a system that takes in, maintains, and computes information over short timescales without a constant source of input. Knowing the limitations of this system can help identify its role in cognition [2] and provide a bridge to developing relevant neural theories. The limits and biases of working memory can be measured by the statistics of recall errors after a delay, for instance, in a visual delayed response task [3]. In these tasks, humans are asked to recall object features, such as location, color, or shape, a short time after presentation [2, 4–6]. When feature values lie on a continuum, subject responses do as well, giving finely resolved measurements of the direction and magnitude of errors on each trial [7, 8]. For example, people's responses on delayed-response tasks often exhibit error magnitudes that increase roughly linearly with time, comparable to the variance of a diffusion process [9, 10], providing a metric that can guide neural theories for working memory.

Complementary to behavioral studies of working memory, theories describing how the brain encodes information over short periods of time provide mechanistic insight. One well-validated theory associates remembered stimulus values with persistent neural activity in recurrently coupled excitatory neurons that are preferentially tuned to the target values [11]. Broadly tuned inhibitory neurons driven by excitation stabilize this activity into a localized structure called an activity *bump* [12, 13]. Variability in neural tuning and synaptic connectivity can cause this activity bump to wander about feature space, causing trial-by-trial errors and biases often perceived as limitations to the system [14–16]. For example, delayed estimates may exhibit serial bias, whereby stimulus values from previous trials may attract or repel the retained memory of the most recent stimulus value [17]. Analogous attractive biases emerge when subjects retain the values of multiple stimuli within a single trial [18]. Additionally, subjects may exhibit systematic biases that include preferences for focal colors [16, 19], orientations [20] and cardinal directions [21].

While biases are often considered reflections of suboptimality, they can be advantageous when reflecting the structure of the environment or sequences of stimuli the subject might see [22, 23]. There is ample evidence that working memory can be trained, and such biases may be the result of long-term learning [24]. Mechanistically, systematic biases in stimulus coding or delayed estimates could emerge from heterogeneities in synaptic connectivity, so collective neural activity is biased to specific network conformations [25]. Such heterogeneity could also be reflected in variations in the sensitivity of individual neurons' stimulus feature tuning [6, 26, 27]. Heterogeneity in the spatial organization of synaptic connectivity can reduce error by maintaining representations that are less susceptible to noise perturbations [28–30]. Thus, if synaptic heterogeneity reflects the learned or expected distribution of stimulus values, recall of common features could be less error prone, improving cognitive efficiency [31].

Since certain stimulus features may be overrepresented in the natural world (e.g., green/brown colors are more common in a forest; see also [32, 33]), we propose that subjects' systematic biases could result from learning the natural distribution of specific features of the environment, which modulates synaptic connectivity to produce representation biases. Here, we model the effects of environmental feature distributions on delayed estimation in neural circuit models and their low-dimensional reductions, considering both models with network connectivity that

is fixed and those shaped by long-term plasticity. We compare these results to human behavior and find that most subjects exhibit strategies best described by learning models, supporting the hypothesis that long-term representation biases reflect the learning of environmental structure.

Results

We begin with the premise that features of natural environments are distributed such that particular values that are overrepresented and thus, statistically more likely to occur as samples (Fig 1A). Such parametric distributions could take on general forms [27], but for illustration,

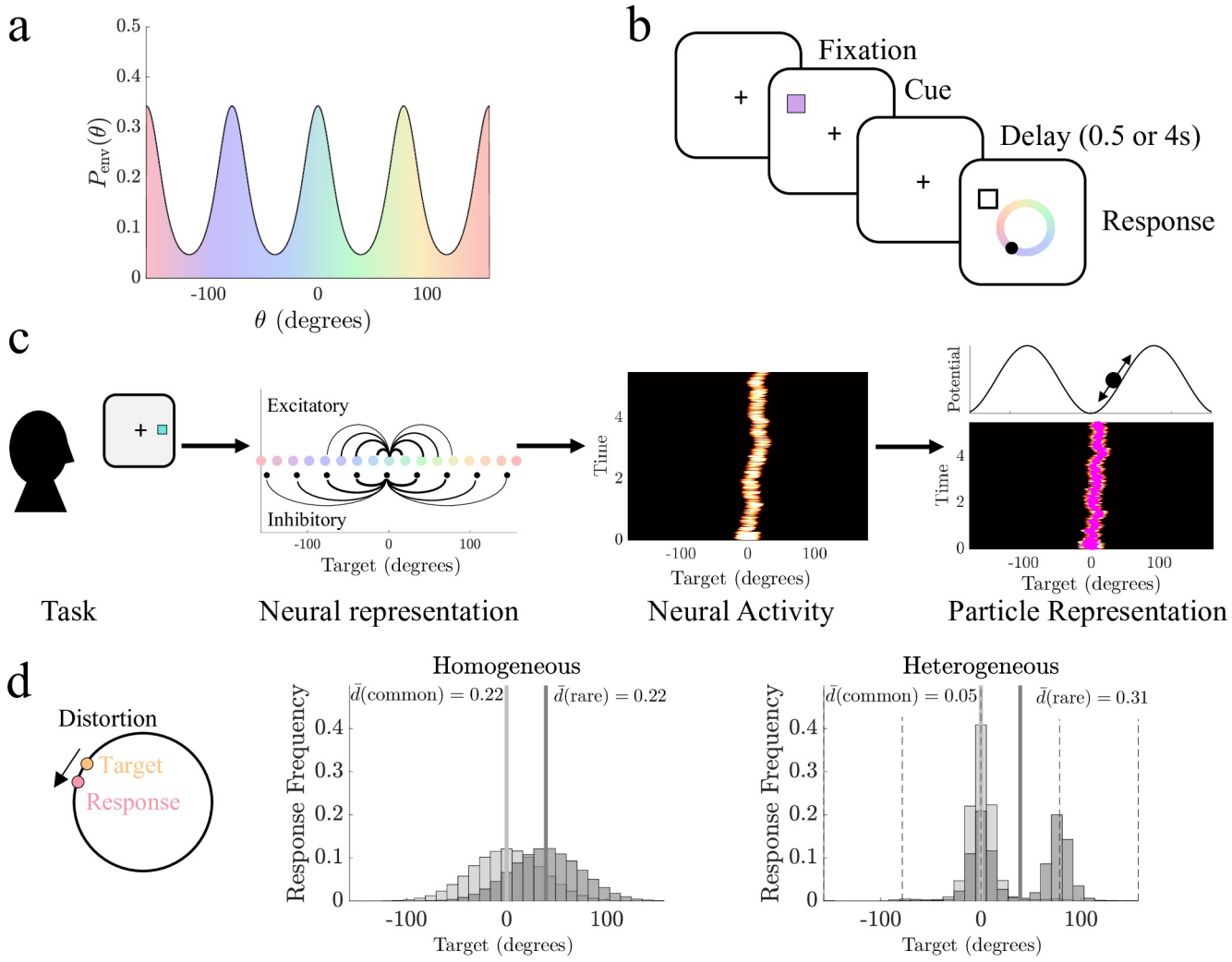

Fig 1. Heterogeneity in the distribution of environmental features is reflected in delayed estimates. a. Natural environments that include overrepresented features, such as certain colors, are described by heterogeneous priors of feature distributions with peaks at the overrepresented features. **b**. Schematic of the delayed estimation task, which requires subjects to remember a target feature (e.g., color) and report it following a short delay period. **c**. The remembered target feature is represented neuromechanistically by a subpopulation of stimulus-specific excitatory neurons with local recurrent excitatory and broad inhibition. The target representation is retained as a bump of sustained neural activity wandering stochastically during the delay. Bump dynamics can be projected to a particle model describing its stochastically evolving position: Spatial heterogeneity in synaptic connectivity is inherited by the particle model as a nontrivial energy landscape with attractors corresponding to regions of enhanced excitation. **d**. Distortion, the circular distance between the target and responses, is influenced by synaptic heterogeneity. In homogeneous networks, response errors at common environmental targets ($\theta = 0$, light grey) and rare targets ($\theta = 45$, dark grey) are equivalent, giving the same local mean distortion ($\bar{d}(\theta)$). With synaptic heterogeneity matched to the environmental prior $P_{\text{env}}(\theta)$, errors are reduced near common stimulus feature values (dashed lines). Parameters used as listed in Methods Table 1.

https://doi.org/10.1371/journal.pcbi.1011622.g001

we assume a parametric prior that is periodic with peaks (dips) at common (rare) stimulus values

$$P_{\text{env}}(\theta) = e^{A\cos(m\theta)},$$

where θ corresponds with a particular feature value described on a ring (e.g., one feature dimension wraps periodically), A describes the amplitude, and m describes the number of peaks in the probability distribution. This periodic function resembles the color biases displayed by humans in [16] as well as cardinal bias common to angle and direction estimates [20, 21]. Note, unless otherwise stated, all subsequent results assume that $m = 4$, so the peaks are centered at cardinal angles of θ, given in radians in formulas, but plotted in degrees in figures for readability.

Our models describe the maintenance of estimates of continuous features [34, 35], arising in tasks where an observer is briefly shown a number of items and, after a delay, probed about remembered stimulus feature values (e.g., location, orientation, or color). These models allow us to speculate on how the environmental priors impact (and potentially bias) how stimulus feature values are remembered (Fig 1B). To illustrate, we focus on examples in which subjects recall colors, though equivalent results can be produced for models of orientation and location recall. Our models are motivated by previous observations that show human performance on delayed estimation tasks degrades over time, such that response variance increases roughly linearly, suggesting a diffusive process drives memory errors [9]. Such diffusive degradation of a stimulus estimate has been modeled in neural circuits as a localized region of persistent activity (bump) that stochastically wanders feature space due to neural and synaptic fluctuations [11, 36]. Activity bumps emerge from strong stimulus-tuned recurrent excitation paired with broad stabilizing inhibition, which generates self-sustained activity [12]. Spatial variation in synaptic connectivity can shape the preferred locations (attractors) of the bump, introducing drift toward attractors [28, 29, 37, 38].

Since the location of the activity bump is a proxy for the remembered stimulus feature value [14, 15], we can simplify our analysis of the impact of activity bump fluctuations by considering low-dimensional models that describe the bump as a particle stochastically moving through an energy landscape (Fig 1C). The (negative) gradient of this energy landscape determines the direction of the drift in the stored estimate [28]. Moreover, asymptotic methods can be used to link the synaptic heterogeneity of neural circuits [39] to a tractable model of the bump position's low-dimensional dynamics [38, 40]. Energy landscapes can be updated to represent an observer's current estimate of the environmental feature distribution $P_{\text{env}}(\theta)$ (see Methods) and can be more easily fit to response data than neural circuit models [14, 16, 23, 41], providing a tractable model for studying the origins of systematic biases in working memory.

We compute our models' average error between a true target value θ and its estimate as the mean distortion $\bar{d}(\theta)$, the circular distance between the target and responses. Overall error across all target values is computed as the total mean distortion $\bar{d}_{\text{tot}} = \int_{-\pi}^{\pi} \bar{d}(\theta) P_{\text{env}}(\theta) d\theta$ [15, 29]. Thus, when synaptic connectivity (and the corresponding energy landscape) is aligned with the environmental prior $P_{\text{env}}(\theta)$, the mean distortion is reduced at common target feature values $\bar{d}(\theta_{\text{common}})$ but increased for rare values $\bar{d}(\theta_{\text{rare}})$. In contrast, purely distance-dependent synaptic connectivity (and a flat energy landscape) produces response distributions and mean distortion that are similar for common and rare target feature values (Fig 1D), making mean distortion a useful metric for quantifying error with respect to changes in synaptic connectivity.

Combining analysis of the energy landscapes with our distortion metric, we now systematically consider the impacts of environmental stimulus distributions on working memory responses, which can guide our understanding of how expectations about the environmental prior can be learned from experience and how these expectations can lead to more efficiently retained memories.

Energy landscapes shape recall distortion

Uniform stimulus priors. We consider a particle model that describes the stochastically evolving estimate of the target feature value in an energy landscape that can incorporate bias, introduced by breaking the symmetry of continuous attractor models of delayed estimation [42]. This low-dimensional model can be derived asymptotically from the stochastic evolution of the position $\theta(t)$ of an activity bump that encodes the estimate and the information about the prior in its network connectivity (see Methods). An energy landscape that reflects an observer's long-term estimate of the periodically-varying environmental prior $P_{\text{env}}(\theta)$ can be generated as

$$U(\theta) = -A_p \cos(n\theta), \tag{1}$$

where A_p describes the well amplitude and n is the number of attractors (each located at the believed common environmental feature values). This simple form for $U(\theta)$ allows us to probe how the alignment of the energy landscape to the true environmental distribution shapes an observer's distortion and produces response biases (Fig 2A).

The movement of the particle through this landscape evolves according to the stochastic differential equation

$$d\theta(t) = -U'(\theta(t))dt + \sigma dW(t), \tag{2}$$

where the particle evolves in response to the systematic drift induced by the energy landscape $U(\theta)$ and dynamic fluctuations generated by the Wiener process $W(t)$. Considering a particle model with a flat energy landscapes ($A_p \equiv 0$), memory of the target stimulus feature evolves according to pure diffusion during the delay period. In contrast, particles evolving along non-trivial energy landscapes ($A_p > 0$) are biased toward the periodically placed attractors at $\theta = \pm(j/n)\pi$ ($j = 0, \ldots, n-1$) (Fig 2B). See also [29, 43] for a detailed account of the *effective diffusion* which can be approximated in a diffusing particle model with a periodic energy landscape.

We first quantify the total mean distortion $\bar{d}_{\text{tot}}$ of responses from particle models encoding stimuli from a uniform environmental prior. Even given a uniform prior, delayed estimates can be improved due to the stabilizing effects of local attractors that mitigate the wandering from diffusion [28, 29, 44]. However, distortion of the target estimate is also enhanced by the introduction of drift induced by energy landscapes with attractors of varying the strength and number of attractors (Fig 2C). As diffusion increased, total distortion was reduced more by considering energy landscapes with fewer attractors, increasing the strength of energy barrier between attractors and the perturbation needed for particles to 'jump' between them (S1 Fig). Since longer delay times increase the possibility of these jumps, total distortion is best reduced in these cases by having fewer attractors, increasing the energy barrier between them. In this case, the reduction of effective diffusion due to strongly quantizing the space of possible particle positions counteracts the local increases in distortion that can arise due to the strong drift of particles starting close to the saddle [29, 43].

Heterogeneous stimulus priors. In addition to the form of the energy landscape, mean distortion is impacted by the form of the environmental prior $P_{\text{env}}(\theta)$. While the conditional

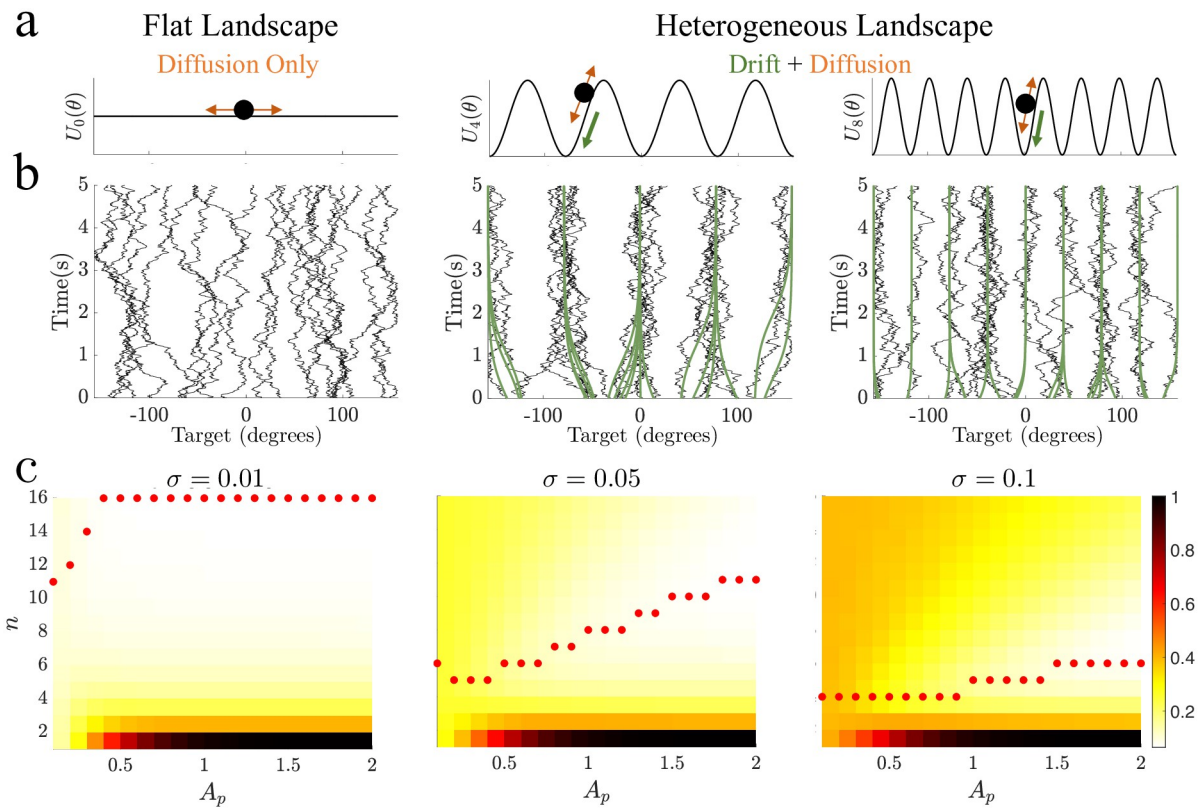

Fig 2. Particles in heterogeneous landscapes are drawn toward attractors. a. Schematics of the flat (homogeneous) landscape, with only diffusion, and heterogeneous landscapes, with potential-driven drift and diffusion. **b**. Example of particle trajectories in the flat and heterogeneous energy landscapes in **a**, sampled from a uniform environmental distribution. In the flat energy landscape, particle motion is driven purely by diffusion. In the heterogeneous energy landscapes, the particle drifts toward attractors over time but diffusion can cause the particle to "jump" wells. Drift-only process shown in green for comparison. Parameters: $A_p = 0, 1$, $n = 4, 8$ **c**. Total mean distortion as amplitude and number of wells was varied for delay of $T_{\text{Delay}} = 5$s and three diffusion values (σ). Optimal particle model identified based on minimum mean distortion (red dots). Parameters not listed for all sub-figures are in in Methods Table 1.

https://doi.org/10.1371/journal.pcbi.1011622.g002

probability of responses is only altered by heterogeneity in the energy landscape, the marginal probability of response is impacted by both the energy landscape and the environmental prior (S2 Fig), confirming that the mean distortion changes with the environmental prior. Matching the number and position of energy landscape wells to the peaks in the prior, we find the mean distortion $\bar{d}(\theta)$ is significantly reduced at common (attractor) locations compared to a model with a flat energy landscape, but shows comparable levels of distortion at rare (saddle) locations (Fig 3A; bootstrapped distortion, $p < 0.05$).

We ask if the total mean distortion typically decreases for periodic energy landscapes Eq (1) as compared to flat landscapes when environmental priors are heterogeneous, and find that it is reduced (relative distortion is negative) as well number increases, especially when the attractor number is matched to the number of peaks in the prior (Fig 3B), though the number of wells does not need to exactly match the number of peaks (S3 Fig) based on the relative contributions of diffusion and drift. Energy landscapes misaligned with the environmental prior (e.g., aligned with rare target locations) generally produced response distributions with higher total mean distortion than aligned models (S4 Fig), confirming that aligning attractors to environmental peaks increases coding accuracy of delayed estimates.

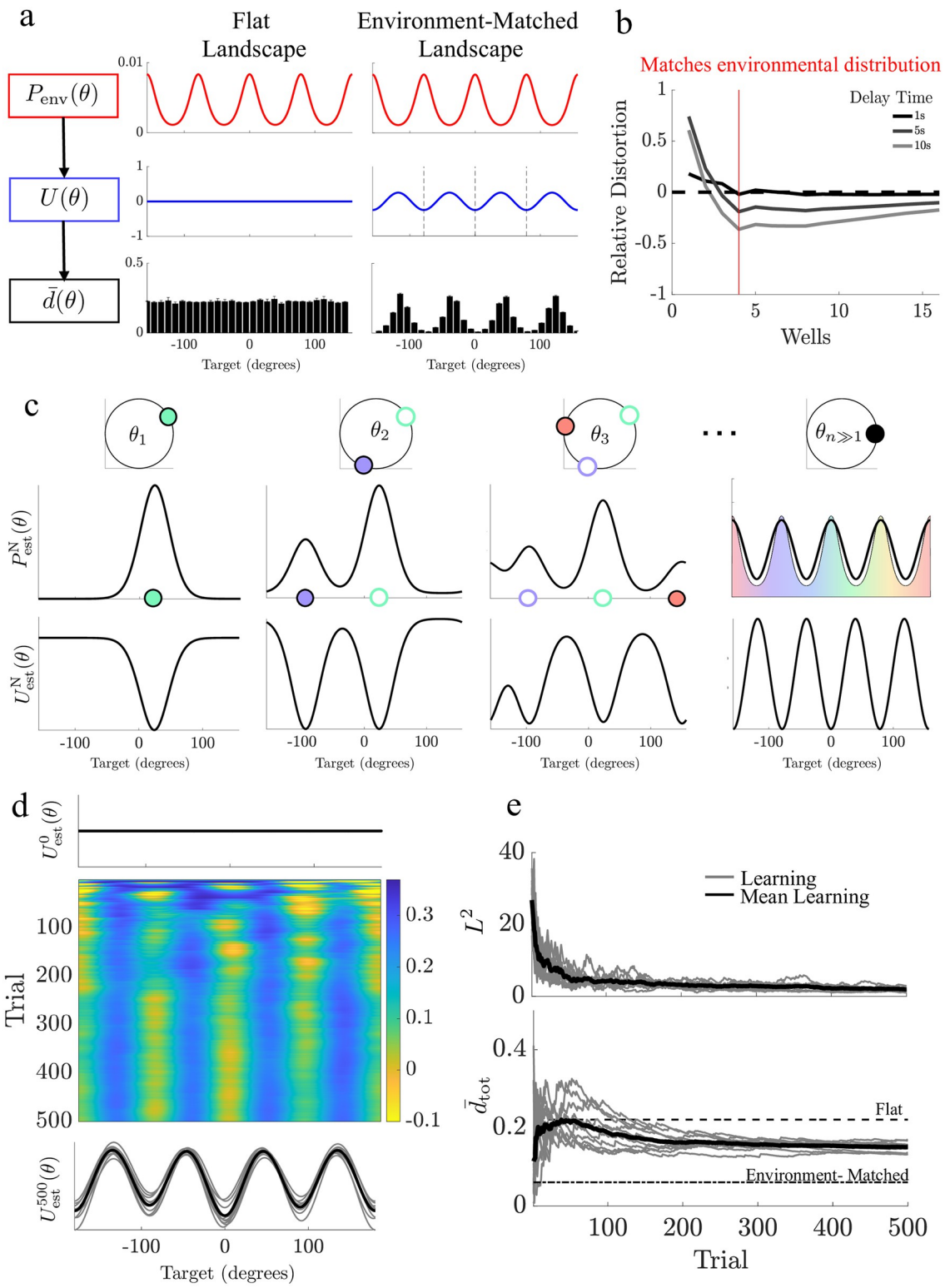

Fig 3. Distortion is reduced when energy landscapes match the environmental prior. a. Top: a heterogeneous environmental distribution ($P_{env}(\theta)$) passes through a heterogeneous energy landscape ($U(\theta)$) and alters the mean bootstrapped distortion ($\bar{d}(\theta)$) at a given target value θ ($N_{boot} = 1e3$). **b**. Relative total mean distortion compared between the flat landscape and heterogeneous landscapes (negative values denote reduced distortion for heterogeneous landscape). Red vertical line denotes where the number (and position) of attractors are aligned to peaks in the environmental prior. **c** Schematic of learning in a particle model. Based on the target

observed on each trial, the estimated environmental distribution $P^N_{est}(\theta) = P(\theta|\theta_{1:N})$ and the particle landscape $U^N_{est}(\theta)$ is updated at the target location. Over the course of many trials, the estimated distribution $P^N_{est}(\theta)$ becomes more similar to the environmental prior $P_{env}(\theta)$, and the energy landscape aligns its wells to its peaks. **d** Heatmap showing the landscape updating over the course of many trials. Top trace shows initial landscape. Bottom trace displays the landscape on the final trial. Grey traces are 10 examples of the learning model, black trace is the average learning model's landscape. **e** Top: L^2–norm for the difference between the experience-dependent belief about the environmental distribution ($P^N_{est}(\theta)$) and the true environmental distribution ($P_{env}(\theta)$). Bottom: Running average of the learning model's total mean distortion ($\bar{d}_{tot}$). Parameters for all sub-figures as listed in Methods Table 1.

https://doi.org/10.1371/journal.pcbi.1011622.g003

Experience-dependent learning in particle models. We next ask whether energy landscapes that model the effects of long-term plasticity can infer a prior based on a long sequence of observations. The effective learning rule assumes subjects sequentially infer the environmental prior from long-term experience: After each trial, the subject's running estimate of the environmental prior is merged with a likelihood function peaked at the current trial's target value. This evolving estimate of the prior can be represented in the energy landscape by updating the landscape such that peaks in the prior estimate are encoded by attractors, corresponding to regions of synaptic potentiation in an equivalent neural circuit description (see Methods and Fig 3C). Over many trials, the energy landscape develops attractors aligned with the common feature locations (Fig 3c and 3d), regardless of observation order (S5 Fig). Thus, the experience-dependent updates generate learning of the environmental prior, and the energy landscape reflects better estimates of the environmental structure, which reduces total mean distortion, trending towards the distortion of a particle model assigned an environment-matched energy landscape (Fig 3E). Note, because learning is continuous, the environmental distribution is only sampled but not precisely represented, and there is decay in learning due to long term depression, the stationary profile of the energy landscape does not precisely match that of the static heterogeneous model, accounting for the differences in total mean distortion.

Subjects' behavior shows hallmarks of learning

We next validate our static and learning particle models against responses from a previously reported data set in which 120 human subjects perform sequences of delayed-estimation trials for target colors drawn from distributions along a one-dimensional ring (see [16] for more details). Subjects were cued with two items, the target and distractor, and asked to respond with the color of one item after a short (0.5s) or long (4s) delay. Item colors on each trial were selected from either an (a) uniform stimulus distribution or (b) heterogeneous distribution with four peaks, offset randomly for each subject (Fig 4A).

We ask if subject responses are best described by particle models with energy landscapes from one of three classes: (a) fixed (flat) and uniform; (b) fixed and heterogeneous; or (c)

Table 1. Parameter values for particle models.

Variable	Value
σ	0.05
A_p	1
n	4
T_{Delay}	5
β	8
h	0.25
s	5

https://doi.org/10.1371/journal.pcbi.1011622.t001

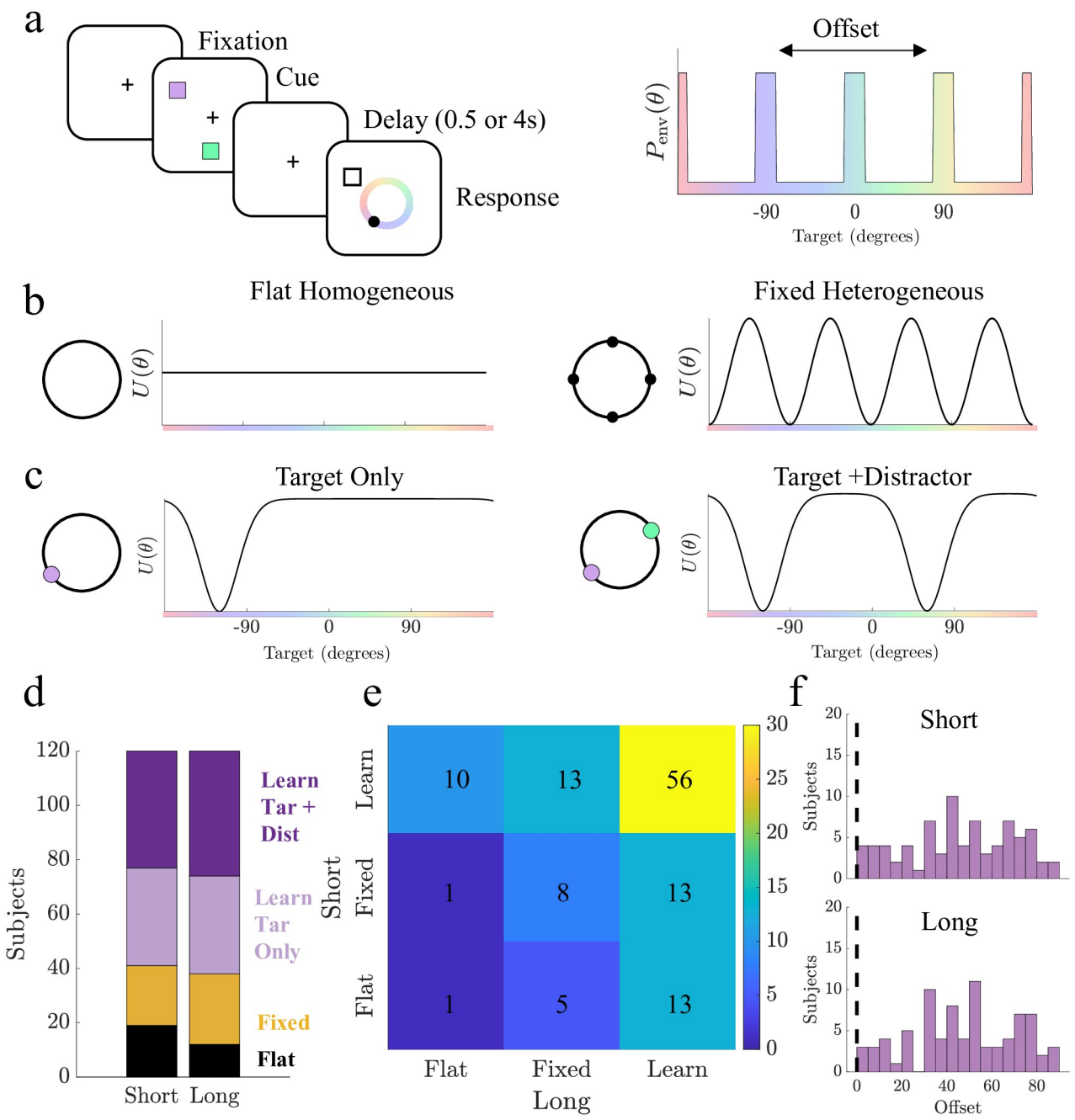

Fig 4. Subject responses based on targets drawn from heterogeneous distributions are best replicated by models governed by heterogeneous landscapes (static and learned). a. Experiment 2 from [16] in which subjects were shown two items, each of which could be drawn from a heterogeneous distribution whose peaks were evenly distributed but randomly offset for each subject. Subjects were prompted to respond with the corresponding color of one item. **b**. Fixed particle models used. Homogeneous landscape includes one free parameter, diffusion. Fixed Heterogeneous models include at least three free parameters: amplitude, number of wells, and diffusion. **c**. Learning particle models. Each model updates iteratively based on three parameters: width of the bump, depth of the bump, and diffusion. Target-Only learning models incorporate only the target prompted for response, and Target+ Distractor models incorporate both items. **d**. Number of subjects best matched to each model type displayed in **b** and **c** for short and long delay periods. **e**. Subjects' best-matched model class for short and long delays. Values show the number of subjects with consistent (upward diagonal) and differing strategy classes across delay times. **f**. Assigned offsets for subjects best matched to Learning models (purple). Dashed lines shows human population bias location.

https://doi.org/10.1371/journal.pcbi.1011622.g004

evolving from each subject's stimulus history. Our fixed and heterogeneous class of models includes three variations: 1. a model with attractors spaced evenly around the ring aligned to each subject's assigned environmental offset (Static Heterogeneous), with free parameters for the amplitude, the number of attractors, and the noise amplitude; 2. a variation allowing the offset of the attractors to deviate from the peaks of the prior (Offset Heterogeneous model); and 3. a variation in which the energy landscape is determined by two Fourier modes (Dual Heterogeneous model) (Fig 4B and S6 Fig).

We also consider four learning models (Fig 4C): one form (two models) updates the energy landscape based only on the target (Target Only), and another form (two models) updates the potential landscape based on both observed items (Target + Distractor). The initial prior (initial landscape) is also varied to account for subjects' potential systematic biases, since subjects can exhibit color biases even given uniform environmental priors [16]. Learning models are initialized either with a flat landscape (Flat Prior) or with a landscape with attractors at the locations of the subject population's biases identified in [16] (Heterogeneous Prior) (S6 Fig). For simplicity, we consider the classes of models (Flat, Fixed Heterogeneous, and Learning) in Fig 4 with additional results for specific model types in S7 Fig.

To identify the model that best matches each subject's responses, we apply cross-validation based on the mean squared error between subject and simulated responses (see Methods) across many possible parameter sets for each model. To consider the possibility that subjects apply different strategies based on the delay period, we analyze short and long trials separately. Nearly all subjects' responses (84% of subjects in short trials and 90% of subjects in long trials) are best described by heterogeneous models, with a majority of subjects applying learning models (66% of subjects in short trials and 68% of subjects in long trials; Fig 4d and S7 Fig). More than half of subjects (54%) apply a consistent strategy type between short and long trials, with 86% of consistent subjects using a learning model (Fig 4e). Of subjects that use the same learning model for both blocks (13 subjects), trial-by-trial analysis reveal that learning models are better matched to trial-specific subject responses once learning occurred, with trial-specific model fits showing a strong likelihood over a fixed heterogeneous model in 92% of subjects (S8 Fig).

We find that many subjects best matched to learning and fixed heterogeneous models have assigned environmental prior offsets centered away from the population biases and not uniformly distributed for both short and long trials ($p < 0.05$, two-sample Kolmogorov-Smirnov test) and that the distribution of assigned offsets for learning model subjects is significantly different than that of subjects best fit to fixed heterogeneous models ($p < 0.05$, two-sample Kolmogorov-Smirnov test), with more learning model subjects having an assigned offset that is far from the population biases (Fig 4f and S9 Fig). Considering subjects who consistently used learning models as compared to those who use different models for each delay or consistently use fixed models, we confirm that learning is more prevalent in subjects with assigned offsets further from population biases (S9 Fig). Given that our learning models implement an experience-dependent updating procedure, these findings suggest that many subjects confronted with observations from an environmental prior that differs from their baseline prior learn the new distribution of stimuli based on this sequence of observations.

Neural mechanism for learning environmental priors

We next study a neural network model capable of implementing experience-dependent inference of environmental priors, comparable to our particle models [22, 45] (see Methods for a demonstration that this model can be asymptotically reduced to our particle models). A neural field model with lateral inhibitory connectivity [29, 46, 47] describes the evolution of neural

activity $u(x, t)$ at locations $x \in [-180, 180]$ corresponding to preferred stimulus value

$$du(x,t) = [-u(x,t) + \int_{-180}^{180} w(x,y) f(u(y,t))dy]dt + \epsilon u(x,t)dW(x,t) + I(x,t)dt. \quad (3)$$

Purely distance-dependent and lateral inhibitory synaptic connectivity $w(x, y) = w_{hom}(x - y)$ is described,

$$w(x-y) = w_{exc}(x-y) - w_{inh}(x-y) = \exp\left[-(x-y)^2\right] - A_{\text{inh}} \exp\left[-\frac{(x-y)^2}{\sigma_{\text{inh}}^2}\right],$$

combining both local excitation and broad inhibition, where A_{inh} is the strength of inhibition and σ_{inh}^2 described the inhibitory spread. Since stimuli lie on a ring, we also pose the neural field upon a periodic ring by wrapping $x, y \in [-180, 180]$ so that the difference $x - y$ should be interpreted as a circular difference. Comparisons with the particle model require converting to radians $180 \mapsto \pi$. This combination of local excitation and lateral inhibition supports the formation of persistent neural activity bumps when a transient input is presented at a particular location [11, 29, 46]. When synaptic connectivity depends only on the difference between neurons' stimulus preferences, bumps have no intrinsically preferred positions in the network and lie along a continuous attractor, establishing an unbiased code for delayed estimation of an input stimulus value. Spatial heterogeneity in connectivity breaks this symmetry to create local attractors as in the particle model. This was either prescribed by a fixed periodic presynaptic function $h(y) = A_n \cdot \cos(ny)$, so that $w(x, y) = (1 + h(y))w_{hom}(x - y)$ or learned via a slowly evolving function $s(y, t)$ that depends on presynaptic neural activity so $w(x, y) = (1 + s(y, t))$ $w_{hom}(x - y)$. Such synaptic heterogeneity can also produce [29] or counteract [28] diverse cellular tuning curves, due to the modulating effects of recurrent architecture. The nonlinear $f(u)$ is a transfer function, $dW(x, t)$ is the increment of a spatiotemporal Weiner process, and $I(x, t)$ is the input representing the cue (See Methods for more details).

Strengthening synaptic efficacy at the peaks of the environmental distribution creates attractors (Fig 5A; right) that bias bumps to drift toward the most common stimulus locations (Fig 5B; right). This relationship between the increase in synaptic efficacy and the formation of attractors can be made mathematically precise via direct asymptotic analysis (see Methods). As such, there is a direct relationship between the stochastic dynamics of a bump's position and the particle models we have discussed already. In short, the introduction of synaptic heterogeneity effectively reshapes an energy landscape that determines the bump position. While this reduces bump wandering when encoding common stimulus values, bumps drift more when instantiated at rare targets, causing larger errors as they are drawn toward attractor locations. As with the particle models, total mean distortion of input stimulus values during the delay is reduced by spatial heterogeneity aligned to the environmental prior (S10 Fig).

We next identify a neuromechanistic learning rule that can modulate synaptic strength based on experience, reshaping the effective energy landscape along which the bump's position evolves: Synapses emanating from activated neurons (those encoding the stimulus value) are potentiated [22], while a weak decay term reduces synaptic strength in regions that have not recently been observed (Fig 5C). This rule comes from ample evidence for physiological mechanisms supporting long-timescale presynaptic potentiation throughout the nervous system [48, 49]. Such synaptic modulation leads to an increase in connectivity strength at the target location of each trial and a reduction of synaptic efficacy for targets that have not been recently observed (Fig 5D). Updates occur iteratively, so synaptic plasticity modulates weight functions across long timescales to reflect the environmental prior (Fig 5E). As with our particle models, once the neural network learns the environmental prior through experience, it maintains delayed

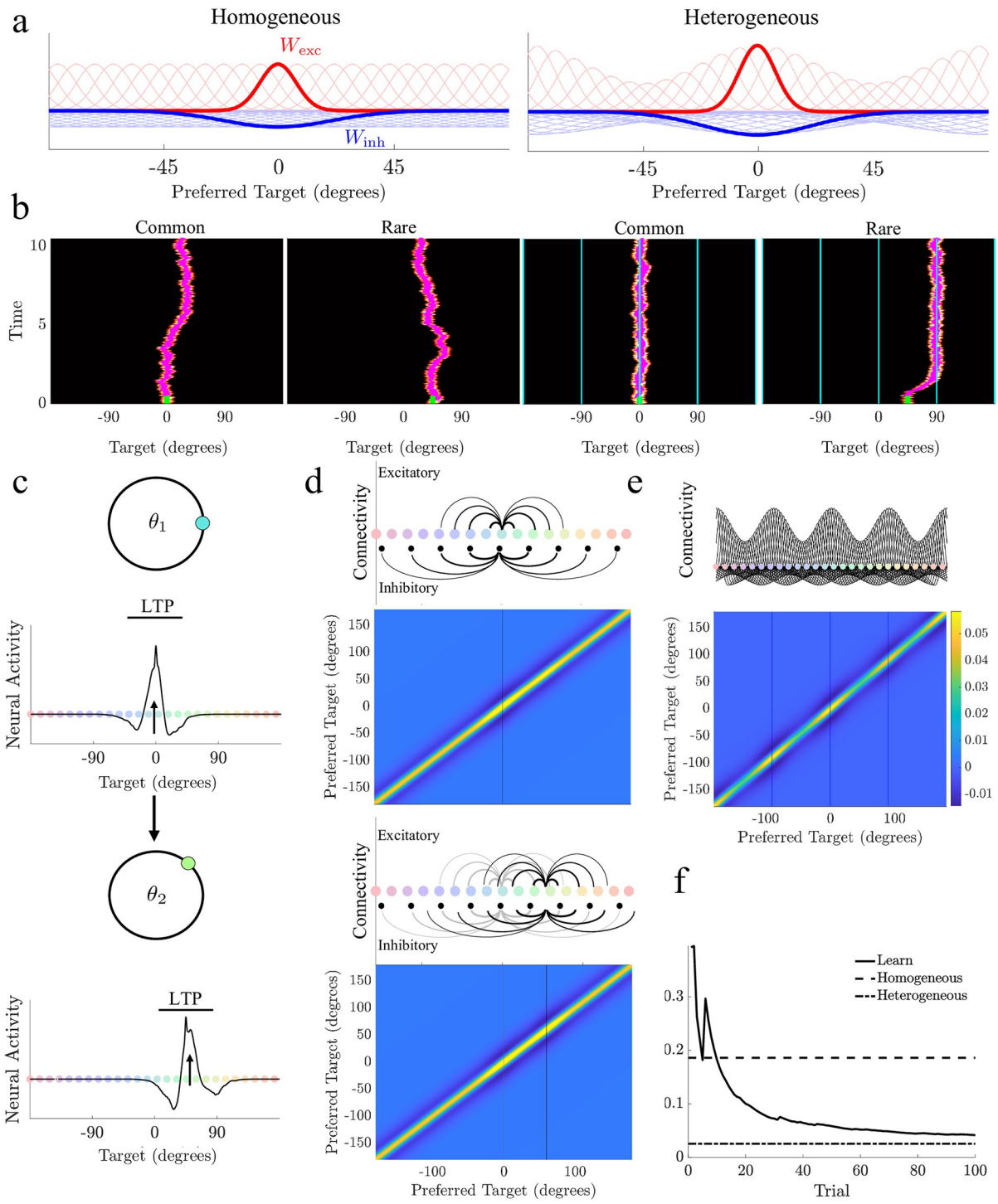

Fig 5. Experience-dependent modulation via long-term potentiation and leak reduces distortion of encoded stimulus values. a. Localized excitatory (W_{exc}) and inhibitory (W_{inh}) synaptic weights associated with preferred item features in a neural field encoding a delayed estimate. Example of of synaptic weight originating from neuron with preference $\theta = 0$ shown in bold. Heterogeneous neural networks are modulated so synaptic footprints originating from peaks of $P_{env}(\theta)$ are stronger. **b**. Example bumps of sustained neural activity over 10s delay period originating at common and rare targets in a homogeneous (left) and heterogeneous (right) case. Cyan traces in heterogeneous plots denote attractor locations with enhanced synaptic weights. Stimulus input duration shown in green. **c**. Experience-dependent learning results from pre-synaptic long-term potentiation (LTP) of neurons with a preference for the previous target value. **d**. Learning breaks the symmetry of the spatially-dependent weight kernel, creating enhanced peaks originating from neurons activated across trials. With inactivation over time, connectivity weakens (grey traces from previous trial). **e**. After many trials, the weight matrix recovers the static heterogeneous synaptic

structure that matches the current environment. **f**. Average total distortion over time decreases as the experience-dependent neural field model learns the environmental distribution. Parameters for all sub-figures as listed in Methods Table 3.

https://doi.org/10.1371/journal.pcbi.1011622.g005

estimates with reduced total mean distortion compared to homogeneous networks with fixed synaptic structure (Fig 5F). Moreover, the weak decay term allows the network to adapt to new heterogeneous environments, updating connectivity to identify the new attractor locations (S11 Fig). Finally, we compared the rate of learning, approximated by the average total distortion over time, between the neural field model and particle models parameterized by subjects who were consistently matched to learning models. We found that the rate of learning between subjects' best fit particle models and the neural field model were qualitatively the same (S12 Fig). The performance and dynamics of the particle and neural models are thus well aligned.

Discussion

We have demonstrated that systematic biases observed in human subjects' delayed estimates can be attributed to environmental experience, specifically corresponding systematic variations in the frequency of stimulus feature values. Our work identifies a potential learning mechanism that can be implemented in reduced models and physiologically motivated neural circuit models and is in line with human response data. This moves beyond prior work, which primarily proposed analogous attractor-based models with fixed energy landscapes [29] and did not propose mechanistic motivations for the attractors [16, 17], by incorporating a Bayesian inference framework and mechanistic basis. Our analysis identifies a mechanism by which sequential inference can be implemented in a particle and neural circuit model with a plausible synaptic plasticity rule.

Beginning with a simplified model of delayed-estimate degradation, we confirm that systematic response biases can be induced by breaking the symmetry of the energy landscape that shapes the evolution of the delayed estimate over time. Such symmetry-breaking stabilizes memories at attractor locations that, when aligned to peaks in the environmental prior, reduce response error at common stimulus values at the expense of larger errors for rare feature values. Overall, total mean distortion of responses is reduced in models with aligned heterogeneous energy landscapes, compared to those with flat landscapes, given the higher propensity of common input stimulus values. Experience-dependent learning of the environment can be implemented neuromechanistically via long-term potentiation that enhances recurrent excitation in neurons encoding common stimulus values and weak synaptic decay for stimuli that have not been observed recently. Responses from human subjects are better matched to models that learn environmental priors than those that are fixed, particularly if the task environment does not well match their baseline beliefs. Thus, subjects confronted with environments that deviate from their priors appear to dynamically update their beliefs based on experience, supporting our hypothesis that systematic biases are learned via experience-dependent plasticity.

Our work supports previous findings showing response variability can be reduced in neuronal networks with spatial heterogeneous synaptic connectivity, even for uniformly distributed stimulus value probabilities [16, 28–30], and extends these findings to determine the efficient tuning of such codes for non-uniform stimulus priors. Our models generate attractive biases whereby delayed estimates tend towards common stimulus values. In related direct perceptual reporting tasks, estimates can be repelled from common stimulus values [50], and such effects have been referred to as "anti-Bayesian" biases. We think one reason for this discrepancy may be that delayed estimation relies on attractor dynamics, which are not necessary in direct reports. In neuronal networks, these attractors are usually created via recurrent architecture,

making synaptic heterogeneity and plasticity a plausible mechanism for representing long-term features of the environment. Direct perceptual reports can be modeled by a neural population with no synaptic connectivity, but heterogeneities can be instituted by varying the amplitude, spacing, and width of neural tuning curves across the population [50]. This particular form of model produces anti-Bayesian biases, where responses are repelled from common stimulus locations. It would be interesting to merge these two considerations in future work, to determine how the effects of synaptic and firing rate function heterogeneity combine. The form of systematic working memory biases is quite varied [21, 51–53], as shown in the diversity of best fit models presented in our work, so there are likely multiple neural and synaptic mechanisms that subserve these biases.

For example, subjects' use of the distractor item as part of their updating procedure suggests that experience-dependent updates could occur during stimulus observation, rather than after subjects' response as suggested by work on short-term serial biases [8, 22]. Representations of memoranda in multiple item working memory tasks have also been shown to interact, sometimes causing additional errors in memory [18, 54, 55] or reducing cardinal biases [56]. Notably, multi-items were presented sequentially in [56], implying that both short-term plasticity rules [22] and multi-item interactions, such as swapping errors [57], may work in conjunction to produce suboptimal strategies. Future work may consider how multi-item working memory tasks impact experience-dependent learning of task environments.

We had hypothesized that our fixed heterogeneous models would better represent subject responses when environmental priors were more aligned to the population biases (offsets closer to the population bias peaks), because these environments would require less updating to subjects' environmental beliefs. In contrast, the population of subjects whose strategies are best described by fixed heterogeneous models have a wide range of assigned offsets, while most subjects described by learning models have assigned offsets that deviated from the original population biases. It is unclear whether subjects matched to static models were resistant to learning or were not given sufficient experience (i.e., trials) to adapt. Likewise, it is possible that subjects could apply a dynamic update procedure that is not based on the specific stimuli shown but generally morphs the stimulus distribution to align with the new environment. However, our trial-by-trial analysis of subjects consistently matched to learning models shows that subjects appear to update their environmental distribution in an experience-dependent way. Future studies could investigate more extensively the rate and form of learning when subjects are presented with stimuli drawn from heterogeneous environmental priors to identify the causes for subject-model variability.

Our work has established and validated a novel mechanistic hypothesis to describe how people infer the distribution of environmental stimuli and its impacts on their delayed estimates. Our results support recent findings on training-induced changes in prefrontal cortex [58], suggesting learning over longer timescales can have substantial stimulus-specific impacts in working memory. Moreover, our work posits that limitations and biases in working memory are not necessarily suboptimal, but can be motivated by efficient coding principles and modulated by environmental inference processes. These findings establish a correspondence between environmental inference and working memory that reveals a deeper understanding on the role of working memory in cognitive processes.

Materials and methods

Particle model

We described the models here in radian coordinates (i.e., the distance around the ring is 2π rather than 360 degrees), but all figures were plotted by rescaling to degrees deg = $(180/\pi) \cdot$

rad. All particle models with fixed energy landscapes used

$$U(\theta) = -\frac{A_p}{n} \cdot \cos(n\theta),$$

in which A_p described the amplitude and n described the number of wells (attractors). The homogeneous model was recovered when taking $A_p = 0$. Particle movement was simulated using a stochastic differential equation

$$d\theta(t) = -U'(\theta(t))dt + \sigma dW(t),$$

which incorporated noise as a Wiener process with increment $dW(t)$. Numerical simulations were performed using the Euler-Maruyama scheme in which the values for θ were discretized to 1 degree (π/180 radian) bins and time was discretized to 10ms bins. All parameter values listed in Table 1 were used unless otherwise stated.

Distortion. The mean distortion for a given input stimulus value θ was computed as

$$\bar{d}(\theta) = \int_{-\pi}^{\pi} P(\theta'|\theta)(1 - \cos(\theta - \theta'))d\theta',$$

where $P(\theta'|\theta)$ indicates the model or subject's probability of responding θ' given the stimulus was θ. Computations are performed using Monte Carlo sampling. To compute stimulus-specific distortion for a given particle model and environment, θ was binned and simulations were used to compute $\bar{d}(\theta)$ for the bin ($N_{\text{sim}} = 10^5$ per bin). Bootstrapping procedures were used to resample distortion and compute the standard deviations ($N_{\text{boot}} = 1e3$).

Total mean distortion across all stimulus values in an environmental prior was computed

$$\bar{d}_{\text{tot}} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} P(\theta)P(\theta'|\theta)(1 - \cos(\theta' - \theta))d\theta d\theta',$$

which can be approximated by Monte Carlo simulations with initial conditions sampled from the environmental distribution $P_{\text{env}}(\theta)$ ($N_{\text{sim}} = 10^5$).

Conditional and marginal distributions. The conditional probability $P(\theta'|\theta)$ was computed by simulating the distribution of responses (particle end locations) for each discretized value θ ($N_{\text{sim}} = 10^5$). Marginal distributions of the response $P_{\text{resp}}(\theta)$ were computed by averaging the discrete conditional probability solutions relative to the known environmental distribution $P_{\text{env}}(\theta)$.

Relating the energy landscape to an experience-based posterior. An experience-based posterior can be related to the stationary distribution of a particle on an energy landscape associated with Eq (2). The equivalent Fokker-Planck equation describing the evolution of the distribution $p(\theta, t)$ of possible particle positions θ at time t assuming a potential function $U(\theta)$ was

$$\partial_t p(\theta, t) = \partial_\theta[U'(\theta)p(\theta, t)] + \frac{\sigma^2}{2}\partial_\theta^2 p(\theta, t). \quad (4)$$

We derived the form of $U(\theta)$ that led to a stationary density that corresponds to a particular posterior $L(\theta)$ in the limit $t \to \infty$ in Eq (4). The stationary density $\bar{p}(\theta)$ was analogous to a posterior $L(\theta)$ since it is the probability density that the system represents when there is no information about the current trial's target remaining. Thus, we derived an association between $\bar{p}(\theta)$ and $U(\theta)$ to identify how the energy landscape function $U(\theta)$ should be tuned so that

$\bar{p}(\theta) \approx L(\theta)$. In the limit $t \to \infty$, we found Eq (4) becomes

$$\partial_\theta^2 \bar{p}(\theta) = -\frac{2}{\sigma^2} \partial_\theta [U'(\theta) \bar{p}(\theta)], \tag{5}$$

a second order ordinary differential equation with solution

$$\bar{p}(\theta) = \chi \exp\left[-\frac{2U(\theta)}{\sigma^2}\right], \tag{6}$$

where χ is a normalization factor. Thus, to match $\bar{p} \approx L$, we need

$$U(\theta) \approx \frac{\sigma^2}{2} \log \frac{\chi}{L(\theta)}.$$

We assumed that we were in the limit of weak heterogeneities, so the deviation of the function $L(\theta)$ from flat will be weak, and $L(\theta) \approx \frac{1}{2\pi} + \epsilon l(\theta)$ (where $\int_{-\pi}^{\pi} l(\theta) d\theta = 0$), which allowed us to make a linear approximation

$$U(\theta) \approx \frac{\sigma^2}{2} [\log 2\pi\chi - 2\pi l(\theta)] \propto -l(\theta),$$

Thus, we removed the constant shift and were only concerned about the proportionality of the energy landscape to the negative of the variation $l(\theta)$ in the posterior.

Experience-dependent particle model. To incorporate learning into the particle model, we updated its energy landscape based on the history of experienced stimulus values according to the equation

$$U_{est}^N(\theta) = \frac{N-1}{N} U_{est}^{N-1}(\theta) + \mathcal{N}_U \frac{h - se^{\beta \cos(\theta - \theta_N)}}{N} \tag{7}$$

which incorporated a von Mises distribution centered at the location of the true stimulus value θ_N on trial N with s as the scaling factor, h the shift, and β the spread. This energy landscape update was meant to represent the trial-by-trial probabilistic update to the stimulus distribution estimate. The mean distortion and the particle landscape were updated iteratively on each trial, such that for each stimulus, the distortion was computed and included in the running average. All parameter values listed in Table 1 were used unless otherwise stated.

The additive update of the particle landscape linearly approximated the typical multiplicative scaling of posterior updating based on successive independent observations. To demonstrate how the updating rule for the energy landscape is related to Bayesian sequential updating of the posterior, recall that that enforcing $U(\theta) \propto -l(\theta)$ ensured an energy landscape aligned with the learned posterior. Thus, we derived an approximate inferred distribution of possible future stimulus target values, updated based on the observed history $\theta_{1:N}$. We assumed that when an observer sees a target value θ_N, they inferred that subsequently similar values are more likely, according to the von Mises distribution

$$f_{\theta_N}(\theta) = \mathcal{N} e^{\beta \cos(\theta - \theta_N)},$$

where $\mathcal{N}$ was a normalization factor. We noted this was self-conjugate ($f_{\theta'}(\theta) \equiv f_\theta(\theta')$). We will also assumed that $0 < \beta \ll 1$, so the variation in $f_{\theta'}(\theta)$ was weak, which allowed us to approximate $f_{\theta_N}(\theta) \approx \mathcal{N}[1 + \beta \cos(\theta - \theta_N)]$. Sequential analysis then could determine how a posterior for future observations should be updated based on each observed target. Take $p_N(\theta) = p(\theta|\theta_{1:N})$ to be the posterior based on past observations $\theta_{1:N}$ which can be computed directly

as the product of probabilities

$$p_N(\theta) = \frac{\bar{p}}{p(\theta_{1:N})} \prod_{j=1}^{N} f_{\theta_j}(\theta),$$

where $\bar{p}$ is the uniform distribution and we have utilized the self-conjugacy of $f_{\theta'}(\theta) \equiv f_{\theta}(\theta')$. We used the linearization of the likelihood function and truncated to linear order in β to find

$$p_N(\theta) \approx \frac{1}{2\pi} \left[1 + \beta \sum_{j=1}^{N} \cos(\theta - \theta_j) \right],$$

which, with the approximate formula for $f_{\theta_N}(\theta)$, can be written as

$$p_N(\theta) \approx \frac{N-1}{N} p_{N-1}(\theta) + \frac{1}{N} f_{\theta_N}(\theta).$$

Lastly, noting the proportional relationship of the desired energy landscape to the posterior, $U_N(\theta) \propto -l_N(\theta)$, we found that the appropriate update for the energy landscape to match this iterative additive update of the posterior was

$$U_N(\theta) \propto \frac{N-1}{N} U_{N-1}(\theta) - \frac{1}{N} f_{\theta_N}(\theta),$$

which we could rewrite using the full form of $f_{\theta_N}(\theta)$ plus a shift to obtain Eq (7).

Thus, in the long-term limit (as $N \to \infty$), the energy landscape convolved the environmental prior $P_{\text{env}}(\theta)$ against the negative of the likelihood function:

$$U_\infty(\theta) \propto - \int_{-\pi}^{\pi} P_{\text{env}}(\theta - \theta') \exp[\beta \cos\theta'] d\theta'.$$

Given that the environmental prior had the form $P_{\text{env}}(\theta) = \mathcal{N} e^{A\cos(m\theta)}$, we then made the approximation
$P_{\text{env}}(\theta - \theta') \approx \frac{1}{2\pi} + A\cos(m(\theta - \theta')) = \frac{1}{2\pi} + A\cos(m\theta)\cos(m\theta') + A\sin(m\theta)\sin(m\theta')$, so we could compute

$$U_\infty(\theta) \propto -\mathcal{A}\cos(m\theta),$$

where $\mathcal{A} = A\int_{-\pi}^{\pi} \cos(m\theta') \exp[\beta\cos\theta'] d\theta'$ and the other term vanished due to its odd symmetry. This was consistent with the form of the fixed heterogeneity we used to align with this environmental prior.

Human data

Response data from a delayed estimation task was taken from [16], experiment 2, with permission. The task was administered to 120 consenting subjects with normal color vision in Amazon Mechanical Turk who performed and achieved minimal engagement. Each trial within the task presented a subject with two colored squares simultaneously for 200ms after which time they disappeared and a delay of 500 ms (100 short trials) or 4000 ms (100 long trials) ensued prior to a response being cued by presenting an outlined square in the location of one of the two previous prompt (implicit identification of the target object). Participants then provided an estimate of the cued color by using a mouse to drag a small circle around a ring of colored continuum. Each item had a 50% chance of being drawn from the biased distribution. The biased distribution included 4 peaks spanning 20 degrees, equally spaced about the circle. The offset of the stimulus peaks were picked uniformly and randomly and assigned independently

to each subject. The location of the population bias was identified based on the peaks in response frequency across the population of human subjects observed in experiment 1 from [16], which probed subjects to report a color drawn from a uniform distribution but subject showed preferences in the reports.

Subject model fitting

We fit subject responses to 8 different particle models, separated by trial duration to account for possible changes in strategy for each trial length, and identified the most likely model using cross-validation:

1. **Flat potential** (1 free parameter) in which the particle dynamics were only influenced by diffusion

$$d\theta(t) = \sigma dW(t).$$

2. **Static Heterogeneous** (3 free parameters) in which the particle was subject to drift and diffusion, parameterized by the A_p (amplitude), n (number of wells), and noise σ,

$$d\theta(t) = -A_p \sin(n\theta - \theta_{\text{off}})dt + \sigma dW(t),$$

where θ_{off} was the offset assigned to a subject by the experiment (not fit).

3. **Offset Heterogeneous** (4 free parameters) included all of the above parameters but incorporated a free parameter for the offset value θ^s_{off}, such that a subject could use a model not aligned to their assigned offset θ_{off}.

4. **Dual Heterogeneous** (5 free parameters) assumed that subject response were governed by an energy landscape determined by two frequencies (n_1 and n_2) with amplitudes A_1 and A_2, and assuming the offset to be at the assigned location. The stochastic dynamics of the particle were described

$$d\theta(t) = -A_1 \sin(n_1\theta - \theta_{\text{off}})dt - A_2 \sin(n_2\theta - \theta_{\text{off}})dt + \sigma dW(t).$$

5–6. **Target-only Learning** (3 free parameters) assumed the energy landscape was updated on each trial as described in the experience-dependent particle model section above, by adding an inverted von Mises distribution centered at the target location with free parameters for β and s. Noise was parameterized by σ as before. The initial landscape was either chosen to be (a) flat $U_0(\theta) = 0$, or (b) heterogeneous $U_0(\theta) = -A_p \cos(4(\theta - \theta_{\text{off}}))$ with $A_p = 1$ and θ_{off} aligned to the offset of the established population biases described above.

7–8. **Target + Distractor Learning** (3 parameters) models were implemented equivalently to the "Target-only" model but were updated by adding two inverted von Mises distributions to the energy landscape on each trial, one centered at the target and the other centered at the distractor stimulus value.

Models were fit to each subject's set of responses using 5-fold cross-validation performed for short and long delay trials separately. For each cross-validation iteration, we sub-selected a unique 20% of the trials uniformly from across the trial block for testing and used the other 80% of trials for training. For each subject-model, we tested 100 parameter sets, selected randomly from a bounded domain for each parameter (Table 2), on the 80% of the trials (training

Table 2. Parameter ranges for human response model fitting.

Model Class	Variable	Bounded Domain
Flat, Fixed, Learn	σ	[0.01, 0.2]
Fixed	$A_p/A_1/A_2$	[0.1, 2]
Fixed	$n/n_1/n_2$	[1, 12]
Fixed	θ^s_{off}	[0, π/2]
Learn	β	[1, 10]
Learn	s	[1, 10]

https://doi.org/10.1371/journal.pcbi.1011622.t002

trials), running 100 simulations with each set of parameters for each trial. The mean squared error (MSE) between each simulated and subject response were computed, and the parameter set with the lowest MSE was selected for that subject-model pair. We then simulated responses for the final 20% of trials using the selected parameter set and computed the MSE for these trials. This process was performed 5 times, testing all trials for a given delay length. The test-set MSEs were then averaged, and the model with lowest mean testing-set MSE was selected for each subject. Our selection of 100 parameter sets was based on the fact that the best-fit parameters were consistent across at least 3/5 cross-validation folds for over 80% of subjects (83% in short trials and 87% in long trials) (S13 Fig). In the case of the learning models, the particle landscapes were updated using all training trials that occurred prior to the trial being simulated, whereas the model parameters were determined based on all of the training trials.

Trial-by-trial analysis was conducted by identifying subject that were best fit to the same learning model across both delays. Parameters from the cross-validation fold with lowest test-set MSE were used to perform trial-by-trial comparisons between the best-fit learning model and the fixed heterogeneous model. A PDE version of the model fitting procedure was used to extract a precise probability for the subject's responses, given a particular model and the log likelihood ratio was computed.

Implementing the neural field model

In the neural field model Eq (3), the firing rate nonlinearity $f(u(y, t))$ was taken to be a Heaviside function

$$H(u-\kappa) = \begin{cases} 1, & u \geq \kappa, \\ 0, & u < \kappa, \end{cases}$$

in which κ described the firing rate threshold.

Noise $\epsilon u(x, t)dW(x, t)$ was weak, multiplicative, and driven by a spatially-dependent, white-in-time, Wiener process with the spatial filter that decayed with distance $|x - y|$:

$$F(x-y) = \sqrt{\epsilon}\exp(-|x-y|),$$

and ϵ described the noise strength.

Input to the network corresponding to the true location of the stimulus target at location x_{targ} was given by

$$I(x,t) = I_0(1 - H(t - t_{\text{inp}}))\exp\left[-\frac{(x-x_{\text{targ}})^2}{2\sigma^2_{\text{inp}}}\right],$$

where I_0 was the strength of the input, t_{inp} was the length of time it lasted, and σ^2_{inp}

Table 3. Neural field parameter values.

Variable	Value
A_{inh}	0.35
σ_{inh}	3
A_n	0.4
n	4
κ	0.1
ϵ	0.5
I_0	1
t_{inp}	0.5
σ_{inp}	1
T_{delay}	10
A_{inp}	1
s_{learn}	0.8
dt	0.1
dx	0.036
γ_s	0.99

https://doi.org/10.1371/journal.pcbi.1011622.t003

parameterized the width of the input. Note, the location x_{targ} was sampled from the environmental distribution $P_{\text{env}}(x)$ as described above to comprise a long sequence $x_{1:N}$ across trials.

Neural activity evolved by applying Euler-Maruyama iterations to the timestep dt and Riemann integration with dx to the integral in the discretized version of Eq (3). The bump's centroid was then identified as the peak in neural activity at each time $\theta_{\text{cent}}(t) = \text{argmax}_{x\in[-\pi,\pi]}u(x,t)$. All model parameters are given in Table 3 and were selected to ensure bumps would not extinguish prior to the end of the delay period. Responses for each trial were reported as the location of centroid at the end of the delay period $\theta_{\text{cent}}(T)$.

Linking the neural field and particle models. The dynamics of bump solutions to Eq (3) can be reduced to first order to describe how their position $\tilde{\theta}(t)$ evolved over time, roughly approximating the centroid (peak location of neural activity). A reduced stochastic differential equation can be derived describing how this position evolves in time due to noise, inputs, and heterogeneity in the weight function. Technical details for such calculation can be found in [22, 45]. Here we give a brief sketch of such analysis, to demonstrate the tight mathematical link between our particle models and the stochastic dynamics of bump solutions to our neural field equations.

Ignoring noise ($\epsilon \to 0$), heterogeneity ($h(y) \to 0$), and input $I \to 0$, Eq (3) had bump solutions $\mathcal{U}(x)$ that satisfied the equation $\mathcal{U}(x) = \int_{-\pi}^{\pi} w(x-y)f(\mathcal{U}(y))dy$ [45]. This bump was marginally stable and lay on a continuous attractor, so it could be placed at any position $[-\pi, \pi]$ [46]. Without loss of generality, we assumed this position was initially $x = 0$, we could track dynamics of the bump's position $\tilde{\theta}(t)$ once noise, heterogeneity, and input were reintroduced by deriving a hierarchy of equations for the expansion $u = \mathcal{U}(x-\tilde{\theta}) + \epsilon\Phi(x-\tilde{\theta},t) + \epsilon^2\Phi_1(x-\tilde{\theta},t) + \cdots$. Enforcing solvability of this hierarchy introduced a condition requiring the sum of the noise, input, and heterogeneity to be orthogonal to the nullspace $\varphi(x)$ of the adjoint of the operator that comes from linearizing Eq (3) about the bump solution. The result was a drift-diffusion equation whose drift was determined by the energy landscape invoked by both the synaptic weight heterogeneity and input

$$d\tilde{\theta} = -U'(\tilde{\theta})dt + d\mathcal{W}(t),$$

precisely the form of Eq (2), where the drift had contributions from the weight heterogeneity and input

$$U'(\tilde{\theta}) = \underbrace{\frac{\int_{-\pi}^{\pi}\varphi(x)\int_{-\pi}^{\pi}w(x-y)h(y+\tilde{\theta})f(\mathcal{U}(y))dydx}{\int_{-\pi}^{\pi}\varphi(x)\mathcal{U}'(x)dx}}_{\text{heterogeneity}} + \underbrace{\frac{\int_{-\pi}^{\pi}\varphi(x)I(x+\tilde{\theta},t)dx}{\int_{-\pi}^{\pi}\varphi(x)\mathcal{U}'(x)dx}}_{\text{input}} \tag{8}$$

and the Wiener process noise $\mathcal{W}(t)$ had zero mean and variance

$$\langle \mathcal{W}(t)^2 \rangle = \epsilon^2 \frac{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\varphi(x)\mathcal{U}(x)\varphi(y)\mathcal{U}(y)C(x-y)dxdy}{\left[\int_{-\pi}^{\pi}\varphi(x)\mathcal{U}'(x)dx\right]^2}. \tag{9}$$

The heterogeneity and input introduced an energy landscape that steers the position $\tilde{\theta}(t)$ of the bump as it responds to noise fluctuations. As shown in [22, 45], by dropping the input term and considering a Heaviside nonlinearity $f(u) = H(u-\kappa)$, $f(\mathcal{U}(x)) = H(x+a) - H(x-a)$ and $\varphi(x) = \delta(x-a) - \delta(x+a)$ where a was the half-width of the bump such that $\mathcal{U}(x) > \kappa$ for $x \in [-a, a]$ and $\mathcal{U}(x) < \kappa$ otherwise and δ was a Dirac delta function. As such, we could simplify the energy landscape gradient formula to find

$$U'(\tilde{\theta}) = \alpha \int_{-a}^{a} [w(y-a) - w(a+y)]h(y+\tilde{\theta})dy.$$

Approximation with Fourier modes. Note that by decomposing the even weight function into its Fourier series, we have

$$w(x-y) = \sum_{k=0}^{\infty} w_k \cos(k(x-y)),$$

which allowed us to write

$$w(a-y) - w(a+y) = 2\sum_{k=1}^{\infty} w_k \sin(ka)\sin(ky).$$

In a similar way, we could decompose the function describing the heterogeneity in the weight

$$h(y) = \sum_{k=0}^{\infty} a_k \sin(ky) + b_k \cos(ky).$$

Approximating by the dominant Fourier mode (assume it is even, $m = \text{argmax}_k b_k$), we took $h(y) \approx b_m \cos(my)$. Integrating against the difference of the shifted homogeneous weight function, then we found $U'(\tilde{\theta}) \approx 2\alpha_m \sin(m\tilde{\theta})$ and thus $U(\tilde{\theta}) \approx -\frac{2\alpha_m}{m}\cos(m\tilde{\theta})$, where

$$\alpha_m = \frac{\alpha w_m}{2m}\sin(ma)(\sin(2{*}ma) - 2ma) + \sum_{k\neq m}\frac{2b_m w_k}{m^2-k^2}\sin(ka)[m\cos(ma)\sin(ka) - k\sin(ma)\cos(ka)].$$

Note also that as m and k differ more, the coefficient in the sum will decrease, suggesting the dominant terms from the series description of w will be those for the modes k indexed close to m. Thus, a scaling of the dominant Fourier mode of the weight heterogeneity well approximated the energy landscape associated with the bump's stochastic motion.

Narrow bump approximation. Assuming the bump width was narrow compared to the length scale of the heterogeneity, we could estimate the integral using the trapezoidal rule

$$U'(\tilde{\theta}) = \alpha a[(w(2a) - w(0))h(-a + \tilde{\theta}) + (w(0) - w(2a))h(a + \tilde{\theta})],$$

so by expanding the even weight function $w(2a) \approx w(0) + 2a^2 w''(0)$ as well as linearizing the heterogeneity $h(\pm a + \tilde{\theta}) \approx h(\tilde{\theta}) \pm ah'(\tilde{\theta})$, we obtained

$$U'(\tilde{\theta}) \approx -4\alpha a^3 w''(0)h'(\tilde{\theta}),$$

and thus

$$U(\tilde{\theta}) \approx -4\alpha a^3 w''(0)h(\tilde{\theta}),$$

so the energy landscape generated for the bump position from weight heterogeneity $h(y)$ was approximately proportional to the negative shape of the heterogeneity. As such, any neurons whose emanating synapses were potentiated/depressed then attracted/repulsed the bump.

Plasticity rules in neural field model. Experience-dependent learning was invoked in the neural field model with an evolving pre-synaptic neural modulation so that $w(x, y) = (1 + s(y, t)) \cdot w_{hom}(x - y)$, updated each trial N based on presynaptic neural activation, $u(y, T^N_{inp})$, at the time T^N_{inp} when network was stimulated on the Nth trial in response to the cue at x_N. Changes to the pre-synaptic modulation term follow the rule

$$\Delta s(y, T^N_{inp}) = \beta_s \cdot u(y, T^N_{inp}) - \gamma_s \cdot s(y, T^{N-1}_{inp}),$$

so the first term in the modulation change was potentiation with the profile of neural activity at the time of stimulation on the Nth trial T^N_{inp}, with strength β_s, and the second term represented the effects of depression during intertrial intervals of inactivity with strength $\gamma_s \in (0, 1)$. Such a rule implemented an activity-dependent form of presynaptic potentiation, which depends only presynaptic activity and affects only synapses emanating from those neurons, deemed transmitter-induced long-term plasticity by [59] and for which multiple mechanisms have been proposed [60–62]. Equivalently, this is a slower form of short term plasticity used in previous neural field models of working memory [22, 38]. As in the case of the energy landscape, we could determine the long-term limiting heterogeneity $s_\infty(y)$ resulting from the learning rule combined with an environmental prior $P_{\text{env}}(\theta)$. Approximating the shape of the instantiated bump by a von Mises distribution centered at the location of the stimulus value on each trial and assuming weak modifications to the heterogeneity, the long time limit gave

$$s_\infty(y) \approx \beta_s \int_{-\pi}^{\pi} P_{\text{env}}(y - y')\exp[\beta_u \cos y']dy',$$

and, by making the approximation
$P_{\text{env}}(y - y') \approx \frac{1}{2\pi} + A\cos(my)\cos(my') + A\sin(my)\sin(my')$, then

$$s_\infty(y) \approx \tilde{\beta}\cos(my),$$

consistent with the expected form of synaptic heterogeneity and resulting energy landscape.

Supporting information

S1 Fig. Total mean distortion as amplitude and number of wells was varied for two example delay periods. Optimal particle model identified based on minimum mean distortion (magenta dots). (a) Low diffusion ($\sigma = 0.01$) leads to a optimal models with higher number of

wells. (b) Moderate diffusion ($\sigma = 0.05$) leads to optimal models with a variable number of wells based on amplitude. (c) High diffusion ($\sigma = 0.1$) leads to a optimal models with lower number of wells.
(TIF)

S2 Fig. Comparing energy landscapes ($U(\theta)$) and heterogeneous feature value distribution ($P_{env}(\theta)$), we find the conditional probability of response $P_{resp,env}(\theta'|\theta)$ and the marginal probability of response $P_{resp}(\theta)$ for particle models with homogeneous (a) and heterogeneous (b) (four wells at environmental distribution peaks) landscapes. Parameters as listed in Methods Table 1.
(TIF)

S3 Fig. Total mean distortion as amplitude and number of wells was varied for three example delay periods. Optimal particle model identified based on minimum mean distortion (magenta dots). (a) Low diffusion ($\sigma = 0.01$) leads to a optimal models with higher number of wells. (b) Moderate diffusion ($\sigma = 0.05$) leads to optimal models with a variable number of wells based on amplitude, often harmonics of the number of environmental peaks. (c) High diffusion ($\sigma = 0.1$) leads to a optimal models with lower number of wells.
(TIF)

S4 Fig. (a) The conditional and marginal probabilities when the heterogeneous particle model has more wells than the environment and offset from the peak locations. This offset leads memoranda to drift to offset locations and shows moderate distortion for all values of θ. (b) Total mean distortion in offset heterogeneous particle models as compared to non-offset models for moderate diffusion ($\sigma = 0.05$). Positive values corresponds with higher levels of distortion in offset models. Parameters: $T_{Delay} = 1$, $n = 8$ offset = 45, all others as listed in Methods Table 1.
(TIF)

S5 Fig. Ordering of observations in the learning particle model does not change the shape of the learned landscape or overall distortion. (a) 10 iterations of the learning model with the same observations but randomized permutations produce potential landscapes with the same shape but differing amplitudes. (b) 10 iterations of the learning model with no diffusion (drift only) and the same observations but randomized permutations show the same overall mean distortion after many trials with minor variations in the learning rate. Parameters used: $\sigma = 0$, all others as listed in Methods Table 1.
(TIF)

S6 Fig. (a) All fixed heterogeneous models. Static Heterogeneous model includes three free parameters: amplitude, number of wells, and diffusion. Offset Heterogeneous includes amplitude and number of wells, diffusion, and one additional parameter for offset. Dual heterogeneous considers five parameters: amplitude and number of wells for the first component, amplitude and number of wells for the second component, and diffusion. (b) Learning particle models. Each updates iteratively based on three parameters: width of the bump, depth of the bump, and diffusion. Target-Only learning incorporated only the target prompted for response, and Target+ Distractor incorporated both items. Priors refer to initial landscape, beginning either with a homogeneous (flat) landscape or a heterogeneous landscape that matched the human population biases. Parameter ranges as listed in Methods Table 2.
(TIF)

S7 Fig. Subjects' best matched models for short and long delays.
(TIF)

S8 Fig. Subject best-fit learning model trial-by-trial analysis. (a) Example of trial-by-trial energy landscape changes (top) and log-likelihood ratio between best learning model compared to the best fixed heterogeneous model (bottom) in subject with the same best-fit learning model for both short and long delays. Red dashed line shows where learning qualitatively overcomes the initial biases in the landscape and corresponds to an increase in LLR trials in favor of the learning model. (b) Total LLR summed across trials for all subjects that were consistently best matched to the same learning model (top) and the fraction of subjects with a total LLR that is positive across trials (bottom). We see that most subjects increase their total LLR over time, suggesting that learning models are becoming more aligned with subjects' responses.
(TIF)

S9 Fig. Histograms of subjects' offsets for: (a) Fixed heterogeneous subjects for short and long delays. (b) All subjects that were consistently best matched to the same model class or different model classes. (c) Subjects consistently matched to the learning model class or to the fixed heterogeneous class. Yellow denotes fixed heterogeneous class, purple denotes learning class, and blue denotes unclassified.
(TIF)

S10 Fig. Total mean distortion in the homogeneous and fixed environment-matched heterogeneous neural field models. Bootstrapped averages ($N_{Boot} = 1e3$) show a significant decrease in distortion for the heterogeneous synaptic connectivity. All model parameters as listed in Methods Table 3.
(TIF)

S11 Fig. Neural field model with initial heterogeneous connectivity that does not match the current environment updates to match the current environmental distribution and decrease distortion. (a) initial connectivity. Black lines denote current environment's attractor locations. (b) final connectivity scheme. (c) Average total distortion across trials. All model parameters as listed in Methods Table 3.
(TIF)

S12 Fig. Average total distortion across all trials for subjects consistently best fit by learning models across both delays (black traces). Plots were created using the parameters with the lowest MSE across all cross-validation folds. Average total distortion from the neural field model (red trace) using the parameters listed in Methods Table 3. The rate that distortion is reduced appears qualitatively similar across the subject-fit particle models and neural field model.
(TIF)

S13 Fig. Histograms of the number of consistent parameter matches for a subject's best-fit model across all 5 cross-validation folds for short and long trials. In short (long) trials, 83% (87%) of subjects were matched to the same parameters 3 or more times.
(TIF)

Acknowledgments

We thank Matthew Panichello and Timothy Buschman for supplying the previously published human data we use here to fit our models and for their discussions with us concerning their

protocols [16]. We thank Matthew Panichello and Krešimir Josić for the helpful feedback and conversations about preliminary drafts of the manuscript.

Author Contributions

Conceptualization: Tahra L. Eissa, Zachary P. Kilpatrick.

Investigation: Tahra L. Eissa.

Methodology: Tahra L. Eissa, Zachary P. Kilpatrick.

Supervision: Zachary P. Kilpatrick.

Visualization: Tahra L. Eissa.

Writing – original draft: Tahra L. Eissa.

Writing – review & editing: Tahra L. Eissa, Zachary P. Kilpatrick.

References

1. Postle BR. Working memory as an emergent property of the mind and brain. Neuroscience. 2006; 139 (1):23–38. https://doi.org/10.1016/j.neuroscience.2005.06.005 PMID: 16324795
2. Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nature neuroscience. 2014; 17 (3):347–356. https://doi.org/10.1038/nn.3655 PMID: 24569831
3. Luck SJ, Vogel EK. Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in cognitive sciences. 2013; 17(8):391–400. https://doi.org/10.1016/j.tics.2013.06.006 PMID: 23850263
4. Fougnie D, Suchow JW, Alvarez GA. Variability in the quality of visual working memory. Nature communications. 2012; 3(1):1–8. https://doi.org/10.1038/ncomms2237 PMID: 23187629
5. Bays PM. Noise in neural populations accounts for errors in working memory. Journal of Neuroscience. 2014; 34(10):3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014 PMID: 24599462
6. Taylor R, Bays PM. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. Journal of Neuroscience. 2018; 38(32):7132–7142. https://doi.org/10.1523/JNEUROSCI.1018-18.2018 PMID: 30006363
7. Bliss DP, Sun JJ, D'Esposito M. Serial dependence is absent at the time of perception but increases in visual working memory. Scientific reports. 2017; 7(1):1–13. https://doi.org/10.1038/s41598-017-15199-7 PMID: 29116132
8. Barbosa J, Stein H, Martinez RL, Galan-Gadea A, Li S, Dalmau J, et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature neuroscience. 2020; 23(8):1016–1024. https://doi.org/10.1038/s41593-020-0644-4 PMID: 32572236
9. Ploner CJ, Gaymard B, Rivaud S, Agid Y, Pierrot-Deseilligny C. Temporal limits of spatial working memory in humans. European Journal of Neuroscience. 1998; 10(2):794–797. https://doi.org/10.1046/j.1460-9568.1998.00101.x PMID: 9749746
10. Schneegans S, Bays PM. Drift in neural population activity causes working memory to deteriorate over time. Journal of Neuroscience. 2018; 38(21):4859–4869. https://doi.org/10.1523/JNEUROSCI.3440-17.2018 PMID: 29703786
11. Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cerebral cortex. 2000; 10(9):910–923. https://doi.org/10.1093/cercor/10.9.910 PMID: 10982751
12. Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995; 14(3):477–485. https://doi.org/10.1016/0896-6273(95)90304-6 PMID: 7695894
13. Wang XJ. Synaptic reverberation underlying mnemonic persistent activity. Trends in neurosciences. 2001; 24(8):455–463. https://doi.org/10.1016/S0166-2236(00)01868-3 PMID: 11476885
14. Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nature neuroscience. 2014; 17(3):431–439. https://doi.org/10.1038/nn.3645 PMID: 24487232
15. Constantinidis C, Klingberg T. The neuroscience of working memory capacity and training. Nature Reviews Neuroscience. 2016; 17(7):438–449. https://doi.org/10.1038/nrn.2016.43 PMID: 27225070

16. Panichello MF, DePasquale B, Pillow JW, Buschman TJ. Error-correcting dynamics in visual working memory. Nature communications. 2019; 10(1):1–11. https://doi.org/10.1038/s41467-019-11298-3 PMID: 31358740

17. Papadimitriou C, Ferdoash A, Snyder LH. Ghosts in the machine: memory interference from the previous trial. Journal of neurophysiology. 2015; 113(2):567–577. https://doi.org/10.1152/jn.00402.2014 PMID: 25376781

18. Almeida R, Barbosa J, Compte A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. Journal of neurophysiology. 2015; 114(3):1806–1818. https://doi.org/10.1152/jn.00362.2015 PMID: 26180122

19. Bae GY, Olkkonen M, Allred SR, Flombaum JI. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. Journal of Experimental Psychology: General. 2015; 144(4):744–763. https://doi.org/10.1037/xge0000076 PMID: 25985259

20. Scocchia L, Cicchini GM, Triesch J. What's "up"? Working memory contents can bias orientation processing. Vision Research. 2013; 78:46–55. https://doi.org/10.1016/j.visres.2012.12.003 PMID: 23262055

21. Girshick AR, Landy MS, Simoncelli EP. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nature Neuroscience. 2011; 14(7):926–932. https://doi.org/10.1038/nn.2831 PMID: 21642976

22. Kilpatrick ZP. Synaptic mechanisms of interference in working memory. Scientific Reports. 2018; 8 (1):7879. https://doi.org/10.1038/s41598-018-25958-9 PMID: 29777113

23. Barbosa J, Stein H, Martinez RL, Galan-Gadea A, Li S, Dalmau J, et al. Interplay between persistent activity and activity-silent dynamics in prefrontal cortex underlies serial biases in working memory. Nature neuroscience. 2020; 23(8):1016–1024. https://doi.org/10.1038/s41593-020-0644-4 PMID: 32572236

24. Klingberg T. Training and plasticity of working memory. Trends in cognitive sciences. 2010; 14(7):317–324. https://doi.org/10.1016/j.tics.2010.05.002 PMID: 20630350

25. Litwin-Kumar A, Doiron B. Formation and maintenance of neuronal assemblies through synaptic plasticity. Nature communications. 2014; 5(1):1–12. https://doi.org/10.1038/ncomms6319 PMID: 25395015

26. Laughlin S. A simple coding procedure enhances a neuron's information capacity. Zeitschrift für Naturforschung c. 1981; 36(9-10):910–912. https://doi.org/10.1515/znc-1981-9-1040 PMID: 7303823

27. Ganguli D, Simoncelli EP. Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. Neural Computation. 2014; 26(10):2103–2134. https://doi.org/10.1162/NECO_a_00638 PMID: 25058702

28. Renart A, Song P, Wang XJ. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron. 2003; 38(3):473–485. https://doi.org/10.1016/S0896-6273(03)00255-1 PMID: 12741993

29. Kilpatrick ZP, Ermentrout B, Doiron B. Optimizing Working Memory with Heterogeneity of Recurrent Cortical Excitation. The Journal of Neuroscience. 2013; 33:18999–19011. https://doi.org/10.1523/JNEUROSCI.1641-13.2013 PMID: 24285904

30. Pollock E, Jazayeri M. Engineering recurrent neural networks from task-relevant manifolds and dynamics. PLOS Computational Biology. 2020; 16. https://doi.org/10.1371/journal.pcbi.1008128 PMID: 32785228

31. Louie K, Glimcher PW. Efficient coding and the neural representation of value. Annals of the New York Academy of Sciences. 2012; 1251(1):13–32. https://doi.org/10.1111/j.1749-6632.2012.06496.x PMID: 22694213

32. Heider ER. Universals in color naming and memory. Journal of Experimental Psychology. 1972; 93:10–20. https://doi.org/10.1037/h0032606 PMID: 5013326

33. Hansen BC, Essock EA. A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. Journal of Vision. 2004; 4. https://doi.org/10.1167/4.12.5 PMID: 15669910

34. Goldman-Rakic PS. Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Progress in brain research. 1991; 85:325–336. https://doi.org/10.1016/S0079-6123(08)62688-6

35. Bays PM, Catalao RF, Husain M. The precision of visual working memory is set by allocation of a shared resource. Journal of vision. 2009; 9(10):7–7. https://doi.org/10.1167/9.10.7 PMID: 19810788

36. Burak Y, Fiete IR. Fundamental limits on persistent activity in networks of noisy neurons. Proceedings of the National Academy of Sciences. 2012; 109(43):17645–17650. https://doi.org/10.1073/pnas.1117386109 PMID: 23047704

37. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. Journal of Neuroscience. 1996; 16(6):2112–2126. https://doi.org/10.1523/JNEUROSCI.16-06-02112.1996 PMID: 8604055

38. Itskov V, Hansel D, Tsodyks M. Short-term facilitation may stabilize parametric working memory trace. Frontiers in computational neuroscience. 2011; 5:40. https://doi.org/10.3389/fncom.2011.00040 PMID: 22028690

39. Hansel D, Mato G. Short-term plasticity explains irregular persistent activity in working memory tasks. Journal of Neuroscience. 2013; 33(1):133–149. https://doi.org/10.1523/JNEUROSCI.3455-12.2013 PMID: 23283328

40. Seeholzer A, Deger M, Gerstner W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. PLoS computational biology. 2019; 15(4):e1006928. https://doi.org/10.1371/journal.pcbi.1006928 PMID: 31002672

41. Schapiro K, Josić K, Kilpatrick ZP, Gold JI. Strategy-dependent effects of working-memory limitations on human perceptual decision-making. Elife. 2022; 11:e73610. https://doi.org/10.7554/eLife.73610 PMID: 35289747

42. Brody CD, Romo R, Kepecs A. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. Current opinion in neurobiology. 2003; 13(2):204–211. https://doi.org/10.1016/S0959-4388(03)00050-3 PMID: 12744975

43. Lindner B, Kostur M, Schimansky-Geier L. Optimal diffusive transport in a tilted periodic potential. Fluct Noise Lett. 2002; 1:R25–R39. https://doi.org/10.1142/S0219477501000056

44. Koulakov AA, Raghavachari S, Kepecs A, Lisman JE. Model for a robust neural integrator. Nature neuroscience. 2002; 5(8):775–782. https://doi.org/10.1038/nn893 PMID: 12134153

45. Kilpatrick ZP, Ermentrout B. Wandering bumps in stochastic neural fields. SIAM Journal on Applied Dynamical Systems. 2013; 12(1):61–94. https://doi.org/10.1137/120877106

46. Amari Si. Dynamics of pattern formation in lateral-inhibition type neural fields. Biological cybernetics. 1977; 27(2):77–87. https://doi.org/10.1007/BF00337259 PMID: 911931

47. Ben-Yishai R, Hansel D, Sompolinsky H. Traveling waves and the processing of weakly tuned inputs in a cortical network module. Journal of computational neuroscience. 1997; 4(1):57–77. https://doi.org/10.1023/A:1008816611284 PMID: 9046452

48. Bhalla US. Molecular computation in neurons: a modeling perspective. Current opinion in neurobiology. 2014; 25:31–37. https://doi.org/10.1016/j.conb.2013.11.006 PMID: 24709598

49. Benna MK, Fusi S. Computational principles of synaptic memory consolidation. Nature neuroscience. 2016; 19(12):1697–1706. https://doi.org/10.1038/nn.4401 PMID: 27694992

50. Wei XX, Stocker AA. A Bayesian observer model constrained by efficient coding can explain'anti-Bayesian' percepts. Nature Neuroscience. 2015; 18(10):1509–1517. https://doi.org/10.1038/nn.4105 PMID: 26343249

51. Bae GY, Olkkonen M, Allred SR, Wilson C, Flombaum JI. Stimulus-specific variability in color working memory with delayed estimation. Journal of Vision. 2014; 14(4):7. https://doi.org/10.1167/14.4.7 PMID: 24715329

52. Pratte MS, Park YE, Rademaker RL, Tong F. Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. Journal of Experimental Psychology: Human Perception and Performance. 2017; 43(1):6–17. https://doi.org/10.1037/xhp0000302 PMID: 28004957

53. de Gardelle V, Kouider S, Sackur J. An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. Journal of Vision. 2010; 10(10):6. https://doi.org/10.1167/10.10.6 PMID: 20884471

54. Bae GY, Luck SJ. Interactions between visual working memory representations. Attention, Perception, & Psychophysics. 2017; 79(8):2376–2395. https://doi.org/10.3758/s13414-017-1404-8 PMID: 28836145

55. Golomb JD. Divided spatial attention and feature-mixing errors. Attention, Perception, & Psychophysics. 2015; 77(8):2562–2569. https://doi.org/10.3758/s13414-015-0951-0 PMID: 26163064

56. Bae GY. Breaking the cardinal rule: The impact of interitem interaction and attentional priority on the cardinal biases in orientation working memory. Attention, Perception, & Psychophysics. 2021;. PMID: 34658001

57. Bays PM. Evaluating and excluding swap errors in analogue tests of working memory. Scientific Reports. 2016; 6(1):19203. https://doi.org/10.1038/srep19203 PMID: 26758902

58. Tang H, Riley MR, Singh B, Qi XL, Blake DT, Constantinidis C. Prefrontal cortical plasticity during learning of cognitive tasks. Nature Communications. 2022; 13(1):90. https://doi.org/10.1038/s41467-021-27695-6 PMID: 35013248

59. Zenke F, Gerstner W. Hebbian plasticity requires compensatory processes on multiple timescales. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017; 372(1715):20160259. https://doi.org/10.1098/rstb.2016.0259 PMID: 28093557

60. Salin PA, Malenka RC, Nicoll RA. Cyclic AMP mediates a presynaptic form of LTP at cerebellar parallel fiber synapses. Neuron. 1996; 16(4):797–803. https://doi.org/10.1016/S0896-6273(00)80099-9 PMID: 8607997

61. Lev-Ram V, Wong ST, Storm DR, Tsien RY. A new form of cerebellar long-term potentiation is postsynaptic and depends on nitric oxide but not cAMP. Proceedings of the National Academy of Sciences. 2002; 99(12):8389–8393. https://doi.org/10.1073/pnas.122206399

62. Kwon HB, Sabatini BL. Glutamate induces de novo growth of functional spines in developing cortex. Nature. 2011; 474(7349):100–104. https://doi.org/10.1038/nature09986 PMID: 21552280