PLOS | SUBMISSION

Task Learning Reveals Neural Signatures of Internal Models In Rodent Prefrontal Cortex

Abhinav Singh[1], Adrien Peyrache[2], Mark D. Humphries[1*],

1 Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
2 Montreal Neurological Institute, McGill University, Montreal, Canada

* mark.humphries@manchester.ac.uk

Abstract

The inherent uncertainty of the world suggests that brains should internally represent its structure using probabilities. This idea has provided a powerful explanation for a range of behavioural phenomena. But describing behaviour in probabilistic terms is not strong evidence that the brain itself explicitly uses probabilistic models. We sought to test whether populations of neurons represent such models in higher cortical regions, learn them, and use them in behaviour. Combining theories of probabilistic learning and sampling, we predicted that trial-evoked and sleeping population activity respectively represent the inferred and expected probabilities generated from an internal model of a behavioural task; and that these distributions would become more similar as the task was learnt. To test these predictions, we analysed population activity from rodent prefrontal cortex before, during, and after sessions of learning rules on a Y-maze. We found that population activity patterns on millisecond time-scales occurred far in excess of chance in both waking and sleep activity. The distributions of these patterns changed between sleep episodes before and after successful learning. Changes were greatest for patterns expressed at the maze's choice point and predicting correct choice of maze arm to obtain reward, consistent with the population activity representing an internal model of the task. As predicted, these changes consistently increased the similarity between the distributions in trials and in post-learning sleep, compared to pre-learning sleep, implying that the underlying probability distribution had stabilised over successful learning. Our results provide evidence that prefrontal cortex contains a probabilistic model of behaviour, which is updated by learning. They thus suggest sample-based internal models are a general computational principle of cortex.

Author Summary

The cerebral cortex contains billions of neurons. The activity of one neuron is lost in this morass, so it is thought that the co-ordinated activity of groups of neurons – "neural ensembles" – are the basic element of cortical computation, underpinning sensation, cognition, and action. But what do these ensembles represent? Here we show that ensemble activity in rodent prefrontal cortex represents samples from an internal model of the world - a probability distribution that the world is in a specific state. We find that this internal model is updated during learning about changes to the world, and is sampled during sleep. These results suggest that probability-based computation is a generic principle of cortex.

Introduction

How do we know what state the world is in? Behavioural evidence suggests brains solve this problem using probabilistic reasoning [1, 2]. Such reasoning implies that brains represent and learn internal models for the statistical structure of the external world [1, 3, 4]. With these models, neurons could represent uncertainty about the world with probability distributions, and update those distributions with new knowledge using the rules of probabilistic inference. Theoretical work has elucidated potential mechanisms for how cortical populations represent and compute with probabilities [5–9], and shown how computational models of inference

predict aspects of cortical activity in sensory and decision-making tasks [2,9,10]. Indeed, though most work on them has focussed on sensory or motor cortex [2,3,9,11], probabilistic internal models are a candidate general computational principle of cortex. But we lack experimental evidence that neural populations represent probabilistic internal models, and update those models through learning.

The medial prefrontal cortex (mPFC) is a natural candidate for addressing questions of internal models in higher cortices. It is necessary for learning new rules or strategies [12,13]. Changes in mPFC neuron firing times correlates with successful rule learning [14], suggesting that mPFC coding of task-related variables changes over learning. Further, mPFC population recording data from the outset of learning on a Y-maze task are available [15]. We thus use that data here to test the hypothesis that mPFC population activity encodes an internal model of a task, and that this model is updated by learning.

To address this hypothesis, we needed to specify two things: an algorithm plausibly used by the mPfC to learn and update a probabilistic internal model; and an implementation of those probability distributions by mPfC neurons. Together, these provide experimentally tractable predictions, by specifying when we expect probability distributions to change, and by specifying what form that change will take in terms of neural activity.

Learning of an uncertain rule from trial-by-trial feedback can be well captured by a probabilistic reinforcement learning algorithm. Such algorithms maintain a probability distribution $P_x(V)$ over the estimated value V of some task-related variable(s) x. Each probability distribution $P_x(V)$ is updated by whether reinforcement is received or not. Irrespective of the details of the algorithm, and as we illustrate below, a basic prediction of any probabilistic learning model is that the probability distributions stabilise after sufficient reinforcement. We thus sought signatures of stabilised probability distributions in the mPfC population activity.

To detect these probability distributions in neural activity we make use of the recent inference-by-sampling hypothesis [7,9,11,16,17]. In this theory, the probabilistic internal model is implemented by the synaptic weights between neurons. The moment-to-moment joint activity of these neurons thus represents samples from the encoded probability distribution. Neural activity evoked by external input represents samples from a "posterior" probability distribution for the world being in a particular state. A strong prediction of this theory is that if the model is encoded by synaptic weights, then spontaneous activity of the same neurons must still represent samples from the internal model. In the absence of external input, these are then samples from the "prior" probability distribution over the expected properties of the world. Such apparent sampling of posterior and prior distributions has been reported in V1 during observations of natural images and in darkness, respectively [16]. A series of models from Maass and colleagues have shown how generic cortical circuits can produce samples from an encoded probability distribution [7,17]. Thus, the sampling implementation is a reasonable candidate for testing the hypothesis of probabilistic internal models in mPfC.

The sampling implementation tells us we can experimentally access probability distributions by observing changes to the joint activity of a population. It also tells us how we might best isolate the encoded internal model, by observing spontaneous activity in the absence of task-related input. For our data, this type of spontaneous activity only occurs during sleep before and after sessions of trials on the maze. This reasoning led us to the hypothesis that we could observe sampling from the internal model during sleep. This hypothesis is consistent with the observations that waking activity in cortex, including mPfC, is coarsely recapitulated during subsequent slow-wave sleep [15,18–20]. We thus sought to isolate stable probability distributions by comparing population activity in sleep before and after training, and between sleep and training-evoked activity.

We show here that changes to moment-to-moment joint activity in mPFC populations during learning match these predictions. Patterns of joint activity during sleep occur above chance both before and after training, consistent with the sampling of an underlying probability distribution. A set of these patterns change their rate of occurrence after training sessions in which behavioural strategy changes. This set of patterns are predictive of task performance, consistent with them being samples from an internal model of the task. As predicted by the probabilistic reinforcement learning model, the direction of change brings the distribution of joint activity in sleep closer to the distribution in the trials after the behavioural strategy changes, indicating the underlying internal model has stabilised. These findings suggest mPFC represents and updates a sample-based internal model of the maze rules.

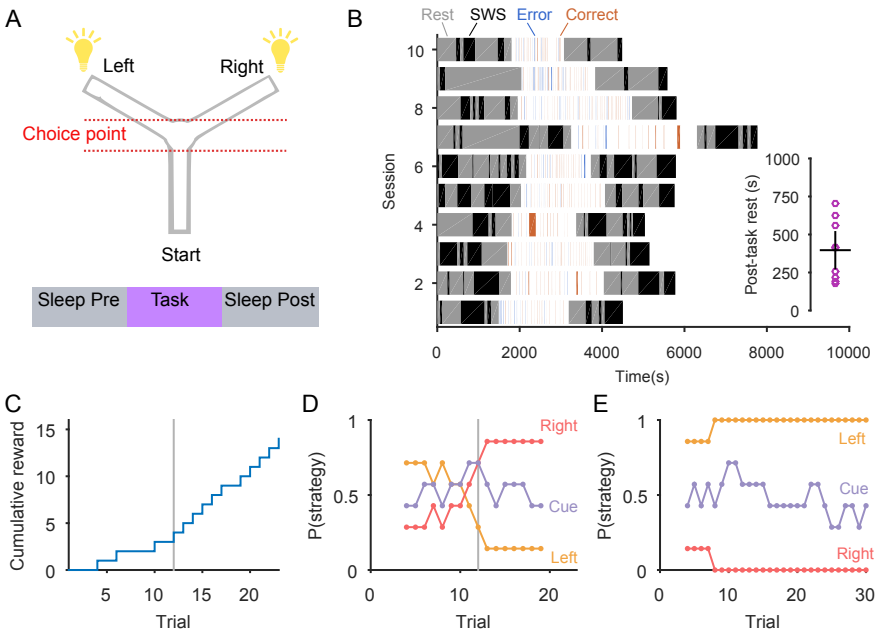

Fig 1. Task and behaviour. (*A*) Y-maze task set-up (top); each session included the epochs of pre-training sleep/rest, training trials, and post-training sleep/rest (bottom). One of four target rules for obtaining reward was enforced throughout a session: go right; go to the cued arm; go left; go to the uncued arm. No rat successfully learnt the uncued-arm rule. (*B*) Breakdown of each learning session into the duration of its state components. The training epoch is divided into correct (red) and error (blue) trials, and inter-trial intervals (white spaces). Trial durations were typically 2-4 seconds, so are thin lines on this scale. The pre- and post-training epochs contained quiet waking and light sleep states ("Rest" period) and identified bouts of slow-wave sleep ("SWS"). Inset: duration of the Rest period between the end of the last trial and the start of the first SWS bout (lines give mean ± 2 s.e.m.) (*C*) Cumulative reward curve from an example learning session. Grey line: learning trial. (*D*) Strategy selection in the same example learning session as panel C. The target rule was 'go right'. Strategy probability was computed in a 7-trial sliding window; we plot the mid-points of the windows. (*E*) Strategy selection in an example "stable" session of consistent behavioural choice. The target rule was 'go to the uncued arm'.

Results

Rats with implanted tetrodes in the mPfC learnt one of four rules on a Y-maze: go right, go to the randomly-cued arm, go left, or go to the uncued arm (Fig. 1*A*). Each rat experienced at least two of the rules. Each training session was a single day containing 3 epochs totalling typically 1.5 hours: pre-training sleep/rest, behavioural training on the task, and post-training sleep/rest. We primarily focussed on ten sessions where the animal reached the learning criteria for a rule mid-session (15-55 neurons per session, Fig. 1*B*). Each learning session had a marked increase in reward accumulation (Fig. 1*C*), correlating with a switch to a consistent, correct strategy (Fig. 1*D*). We also identified a separate set of stable-behaviour sessions, in which the rat consistently used the same strategy throughout irrespective of its accuracy (Fig. 1*E*). As we show below, we used these learning and stable sessions to seek changes to the hypothesised internal models of the task.

Probabilistic reinforcement learning model predicts stabilisation of probability distributions

We used a model of probabilistic reinforcement learning on a simulated Y-maze task to illustrate how a probabilistic internal model is updated during behaviour. Our model maintains a probability distribution over the expected reward obtained by choosing each strategy (Fig. 2A). As the actual distributions encoded by mPFC are unknown, we use this simplified representation as a proxy for more complex models with

distributions over the uncertain values of individual actions and the transitions they cause between states in the maze, which collectively make a strategy. On each simulated trial, the model stochastically chooses a strategy, takes the corresponding action, and observes the resultant feedback. The probability distribution of the selected strategy is then updated to increase or decrease the expected value and the variance around it, according to the feedback. The model is thus an example of general algorithms for updating probabilistic internal models from feedback.

Simulating the model shows how learning the correct strategy corresponds to the probability distributions stabilising (Fig. 2). Like the rat, the model shows a marked increase in reward accumulation (Fig. 2B), corresponding to the dominant selection of the correct strategy. Consistent reward accumulation will cause the distributions to stabilise (Fig. 2C), as their changes asymptotically decrease with continual successful outcomes (see SI Text for details). Thus the model shows the general prediction that successful learning corresponds to stability of the encoded probability distributions.

We use this model to demonstrate specific, testable predictions for the changes to the hypothesised internal models in mPfC. A key constraint here is that all neural recordings were spike-sorted within session only, so we can only seek predictions for within session changes - we outline further predictions for between-session changes, testable in future experiments, in the Discussion.

First, the model predicts that learning should change the probability distributions encoded in sleep. Learning-induced changes in the encoded probability distributions can be observed by comparing distributions taken before and after learning (Fig. 2B-C). If the spontaneous activity of sleep is sampling from the internal model, then we should observe these changes by comparing the distributions encoded in pre- and post-learning session sleep and finding that they are not the same (Fig. 2F). Calling the pre- and post-training distributions $P(Pre)$ and $P(Post)$, and the distance between those distributions $D(Pre|Post)$, then this prediction is that $D(Pre|Post) > 0$.

Second, the model predicts that learning should move the probability distributions in training closer to those sampled in post-training than pre-training sleep. The internal model will change within a learning session less after the learning trial than before it (Fig 2D-E), because of the increased stability of the probability distributions with learning (Fig. 2C). If the spontaneous activity of sleep is sampling from the internal model, then this means the distribution in post-learning trials will be closer to that in post-training sleep then pre-training sleep. Calling the post-learning distribution $P(Learn)$, then this prediction is that $D(Pre|Learn) > D(Post|Learn)$.

Third, the model predicts that stable behaviour correlates with stable probability distributions. While overt changes in behaviour must correlate with changes in neural activity guiding that behaviour, the converse need not be true: neural activity could change without behavioural change (as, for example, in working memory encoding of an object). Nonetheless, if the hypothesised internal model in mPfC is encoding the current behavioural strategy, then we expect that the probability distributions generated by the internal model will not change if behaviour is stable (Fig. 2B-E). Consequently, we expect the probability distributions in pre- and post-training sleep to be equidistant, on average, from the probability distribution in training. Calling the stable-trial distribution $P(Stable)$, then this prediction is that $D(Pre|Stable) \approx D(Post|Stable)$.

Firing rate distributions do not systematically change

Our implementation hypothesis is that these probability distributions are encoded by the patterns of joint activity of the population. But is this implementation hypothesis necessary - could these probabilities be simply encoded by the firing rates of neurons in the population?

We found that there were no systematic changes to the firing rates with learning. In each learning session, the distribution of firing rates changed more between the training epoch and each sleep epoch than between the sleep epochs (Fig 3A). This shift in distribution during the training epoch was accounted for by a sub-set of neurons whose change in rate compared to sleep was in excess of anything observed between sleep states (Fig. 3B). Of the ten learning sessions, only two showed a detectable change in firing rates across the population between the sleep epochs (Fig. 3C), and these two were in opposite directions. These data are consistent with the need to look at patterns of population activity, rather than individual neurons, to test our predictions.

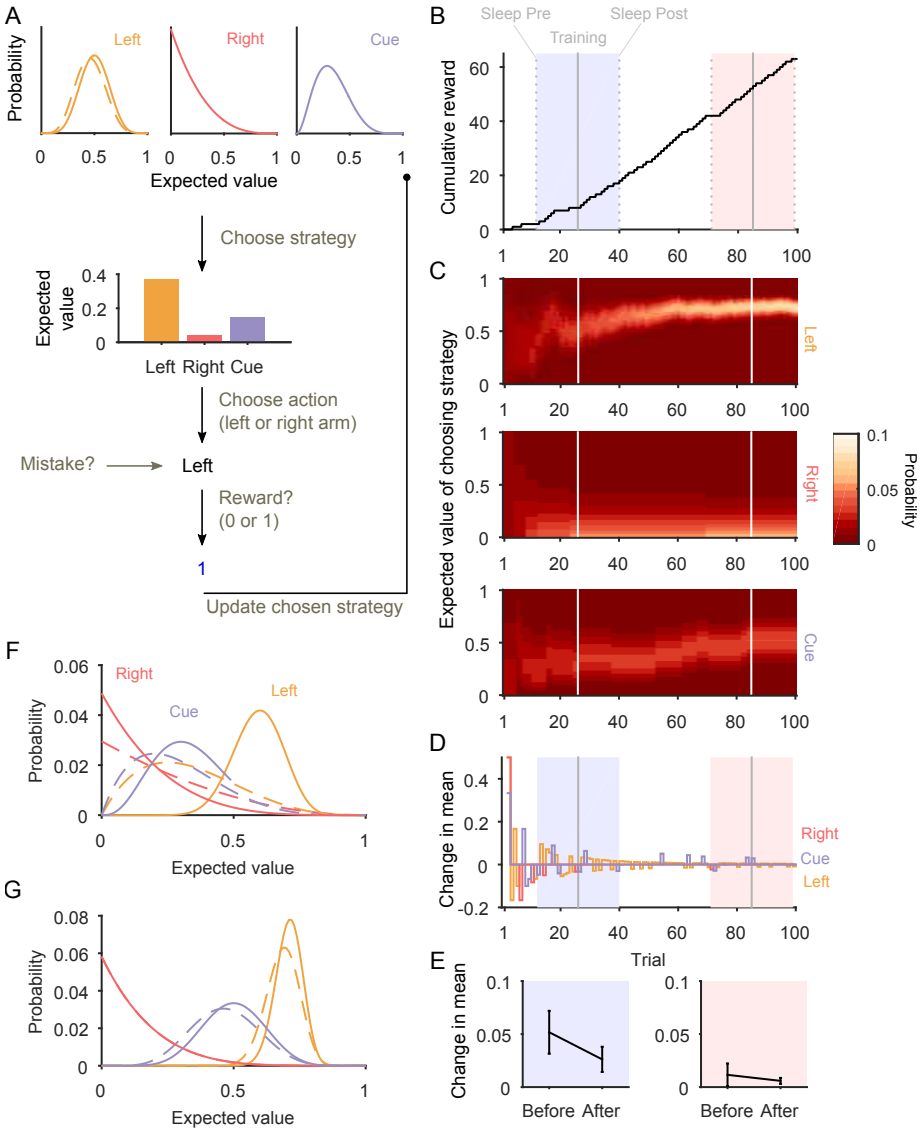

Fig 2. Probabilistic reinforcement learning model predicts stabilisation at learning. (*A*) Schematic of the model. The model maintains probability distributions over the expected value of choosing each strategy (dashed lines). On each trial, a strategy is selected according to the highest sample drawn from each distribution. The corresponding action, of selecting the left or right arm on the maze, is executed. Noise is introduced here as a small probability of executing the opposite action (labelled 'Mistake?'). Reward is obtained, and the probability distribution for the chosen strategy is correspondingly updated (solid lines). (*B*) Cumulative reward curve from an example simulation with reward for "go left". The blue shading identifies a virtual "learning" session, a group of trials around the identified learning trial (solid grey line; see Methods). The dotted grey lines identify the trials occurring immediately before pre and post-training sleep, whose distributions are then sampled in sleep. Red shading identifies an arbitrary later virtual session of stable behaviour, with consistent accumulation of rewards; the grey lines here identify the mid-session and putative pre- and post-session sleep. (*C*) Corresponding trial-by-trial probability distributions for the expected value of each strategy. Colour-scale gives probability; white lines indicate the learning and stable session mid-points. (*D*) Corresponding trial-by-trial change in the mean of the probability distribution updated on each trial. Shading conventions as per panel B. (*E*) The distribution of changes to the mean before and after the session mid-point, for the learning session (blue) and stable session (red). Error bars plot means and standard deviations. (*F*) Probability distributions for each strategy in pre- and post-training sleep for the learning session (dashed: pre; solid:post). (*G*) Probability distributions for each strategy in pre- and post-training sleep for the stable session (dashed: pre; solid:post).

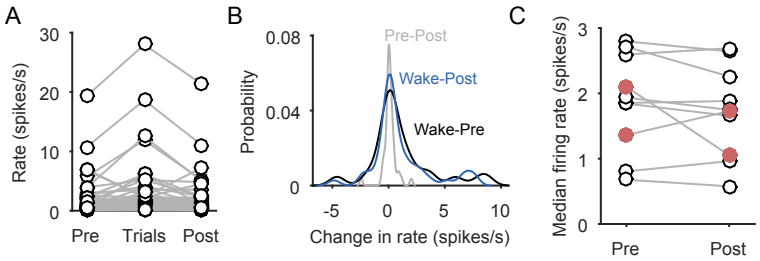

Fig 3. Firing rate distributions do not change between sleep epochs. (*A*) The distributions of firing rates in the three epochs of one learning session. Firing rates within epochs have a long-tailed distribution, with low firing rates dominating. (*B*) Histograms of each neuron's change in firing rate between all pairs of epochs in the same learning session. Changes between sleep epochs are small, and centred at zero. Changes between sleep and waking can be considerably larger, and in either direction. (*C*) The median firing rate in each sleep epoch, by session. The red symbols indicate the only two sessions with a detectable shift in firing rates between the sleep epochs at $\alpha = 0.05$ (Signed-rank test; see SI Table for numbers of neurons per session.

Millisecond precision spike correlation patterns consistent with sampling

We first tested that mPfC population joint activity patterns in the learning sessions were consistent with being samples from a probability distribution. Following previous work [7, 16, 17, 21], we defined the samples as population-wide activity patterns on millisecond time-scales. Activity patterns were characterised as a binary vector (or "word") of active and inactive neurons within some small time window (Fig. 4*A*). Statistical structure at millisecond time-scales has been characterised for populations in the retina [21–24] and primary visual cortex [16,25], but not for higher-order cortices. We thus first demonstrate that mPFC activity patterns on millisecond time-scales contain above-chance statistical structure.

We were primarily interested in co-activation patterns of more than one neuron firing together, as the occurrences of each pattern with a single active neuron (a single "1") can correlate strongly with that neuron's firing rate. We thus first determined the time-scales at which co-activation patterns appear. Figure 4*B* shows that at low millisecond time-scales the proportion of activity patterns containing co-active neurons increases by an order of magnitude when doubling the bin size. The smallest bin size with a non-negligible proportion of co-activation patterns was 2 ms, with $\sim 1\%$ (89731/7452300) of all patterns. This was also true for each epoch considered separately (Fig. 4*C*-*E*). We thus used a 2 ms bin size throughout, as this was the smallest time-scale with consistent co-activation patterns.

Such co-activation patterns could be due to persistent, precise correlations between spike-times in different neurons, or just due to coincident firing of otherwise independent neurons. We found that the proportion of co-activation patterns in the data exceeded those predicted for independent neurons by a factor of 3 (Fig. 4*B*) at low millisecond time-scales. This was also true for each separate epoch (Fig. 4*C*-*E*), extending up to a factor of at least 6 for the task trials (Fig. 4*D*). These data rule out the possibility that the excess of precise correlations was due to differences in brain state.

Our hypothesis that sleep and waking states represent distributions derived from the same internal model requires not just precise patterns, but largely the same patterns. If the set of patterns markedly differed between waking and sleep, then it would be implausible that they were drawn from the same underlying internal model. We found that each recorded population of N neurons had the same sub-set of all 2^N possible activity patterns in all epochs (Fig. 4*B*). Such a common set of patterns is consistent with their being samples generated from the same form of internal model across both behaviour and sleep.

Distributions of activity patterns change between sleep epochs during learning

With evidence that the joint population activity patterns in mPfC were both non-trivial and conserved between epochs, we could test our main predictions. If activity patterns are samples from a probability distribution, then two similar probability distributions will be revealed by the similar frequencies of sampling each pattern [16]. Our first prediction is that the probability distributions encoded in sleep will change due to learning during training (Fig. 5A). We thus test this prediction by comparing the distributions of patterns in pre- and post-training sleep for the ten learning sessions (Fig. 5B).

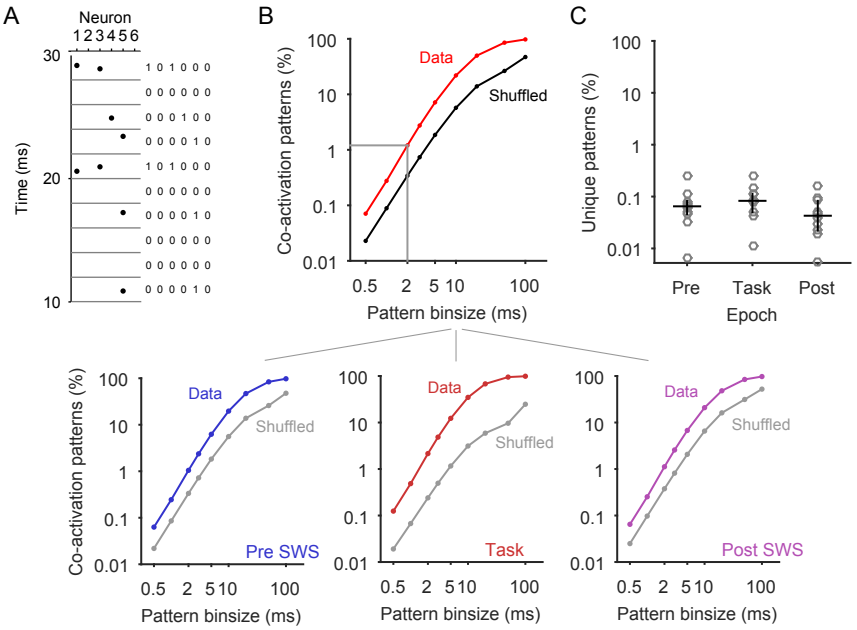

Fig 4. Activity pattern distributions during rule-learning. (A) The population activity of simultaneously recorded spike trains was represented as a binary activity pattern in some small time-bin (here 2 ms). (B) Proportion of co-activation patterns at each bin size (red line). Grey line indicates the proportion of $\sim 1\%$ at the bin size of 2 ms. In black we plot the corresponding proportion of co-activation patterns predicted if all neurons were firing independently; these are obtained by shuffling the inter-spike intervals of each neuron and recomputing the activity patterns. Error bars of ± 2 SEM are too small to see on this scale. Breakout plots: Proportion of co-activation patterns per epoch. Predicted proportions by independently-firing neurons are in grey. Error bars of ± 2 SEM are too small to see on this scale. (C) Consistent sampling of activity patterns across session epochs. Each circle is the proportion of all 2 ms activity patterns from the entire session that appeared only in that epoch. Black bar and line give the median and interquartile range across the 10 sessions. Note the log-scale, showing that the median proportion of unique patterns was less then 0.1% in all three epochs of the session.

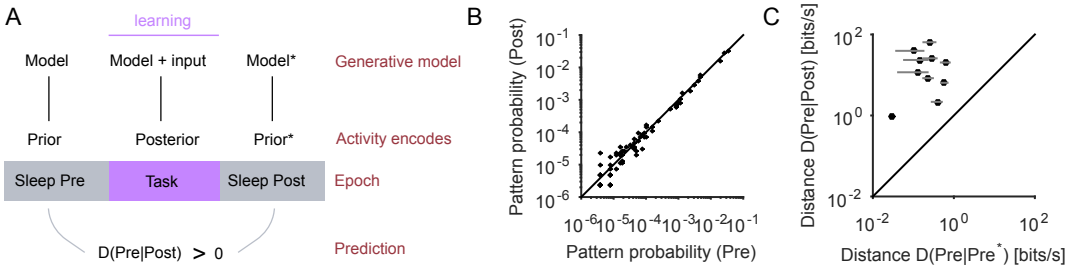

Fig 5. Distributions of joint activity patterns change between pre- and post-learning sleep. (A) Schematic of the theoretical prediction for the change in probability distributions between sleep epochs of a learning session. If learning during the task changes the encoded internal model, then spontaneous activity during sleep would sample from the updated model. Consequently, the distributions of joint population activity in pre- and post-training sleep should systematically differ. (B) The joint frequency of every occurring pattern in pre-training sleep (distribution $P(Pre)$) and post-training sleep (distribution $P(Post)$) for one session. (C) Distances between pre- and post-training sleep distributions (y-axis) for every learning session, compared to a per-session estimate of baseline differences (x-axis), obtained by bootstrap sampling of patterns within the pre-training sleep epoch. Error bars give the mean and 95% confidence interval on the bootstrapped within-epoch distance $D(Pre|Pre^*)$; identical results were obtained when using $D(Post|Post^*)$.

Our prediction is that the distance $D(Pre|Post)$ between these distributions should be greater than zero. Due to the finite duration of the two sleep epochs, and so the limited sampling of each activity pattern, identical underlying probability distributions will give rise to similar but not identical distributions of activity patterns. We thus estimate the expected distances for identical distributions by bootstrap sampling within each epoch, giving estimates of $D(Pre|Pre^*)$ and $D(Post|Post*)$ for the distances between sets of patterns drawn from identical underlying distributions.

In every learning session, we found the distance between sleep-epoch distributions $D(Pre|Post)$ was greater than within those epochs $D(Pre|Pre^*)$ (Fig. 5C). We found similar results when we estimated $D(Pre|Pre^*)$ by randomly dividing the sleep epochs into two sets of samples and computing the distance between the two (S1 Fig). In both cases, identical results were found when using post-training rather then pre-training sleep as the control epoch (results not shown). Consistent with the hypothesis of updated internal models sampled in post-training sleep, there is a systematic change in the population distribution of activity patterns between sleep epochs.

Distributions in sleep are consistent with an internal model of the task

If activity pattern changes between sleep epochs either side of learning are caused by an updated internal model in mPFC, then the changes to the distributions should be related to learning the task. Patterns that changed their occurrence between sleep epochs should indicate the parts of the model that was updated. To test this, we sought whether these updated patterns were encoding task variables. If not, then this would be evidence against encoding of an internal model.

Distribution changes correlate with trial outcomes

If the hypothesised internal model is updated by trial outcome, so trial outcome should be correlated with the consequent change in sampling of activity patterns. To test this, for each co-activation pattern, we found its ability to predict a trial's outcome by its rate of occurrence on that trial (Fig 6A). We then compared this outcome prediction to the change in sampling between pre- and post-training sleep (Fig. 6B). As all learning sessions were dominated by patterns that did not change between pre- and post-training sleep, precluding a straightforward correlation analysis (S2 Fig), we discretised the distribution of changes as a function of outcome prediction (Fig. 6C). We found a strong correlation between the outcome prediction and the likelihood of a pattern changing its sampling between the pre- and post-training sleep (Fig. 6D). This correlation was highly robust to how we constructed the distributions of change between sleep epochs (Fig. 6E-G). The learnt internal model, as evidenced by the updated patterns sampled from it, was seemingly encoding the task.

Distributions are sampled around the task decision point

As the rats were performing a navigation task, we could also verify that the hypothesised internal model was sampled at a relevant location for performing the task. If good outcome prediction indicates patterns that are indeed sampled from the relevant internal model for behavioural strategy, then we would expect these patterns to occur at or close to the maze decision point, where the strategy is relevant. We thus checked the locations of the co-activation patterns as a function of outcome prediction. We found that the outcome-predictive activity patterns preferentially occurred around the choice point of the maze (Fig. 7). Particularly striking was that patterns strongly predictive of outcome rarely occurred in the starting arm (Fig. 7A). Together, the selective changes over learning to outcome-specific (Fig. 6) and location-specific (Fig. 7) activity patterns are consistent with learning updating a behaviourally-relevant internal model, which is sampled in sleep.

Activity distributions during learning converge between training and post-training sleep

Our key prediction is that learning should cause not just a change but a stabilisation of the probability distributions derived from the internal model. The above evidence shows change between sleep epochs, but not the direction of change. To examine the direction of change, we consider the activity pattern distribution

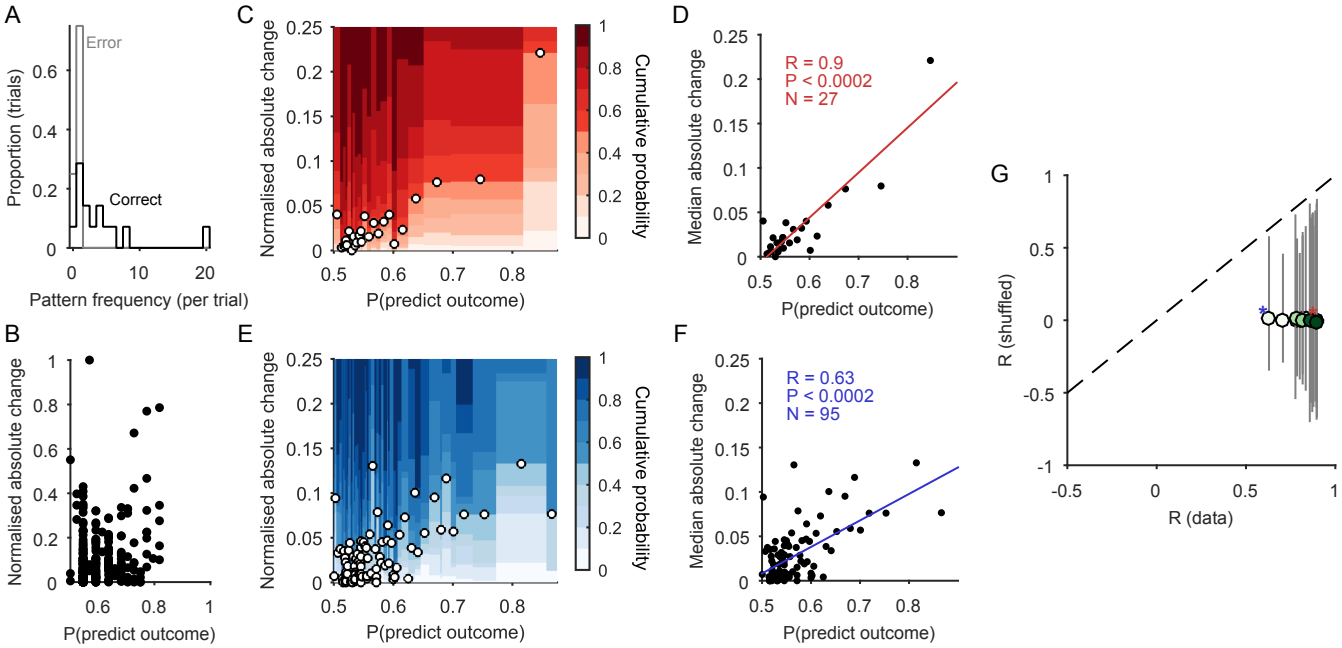

Fig 6. Coding of trial outcome by sampled activity patterns. (A) Example distributions of a pattern's frequency conditioned on trial outcome from one learning session. (B) For all co-activation patterns in one learning session, a scatter plot of outcome prediction and (absolute) change in pattern frequency between pre- and post-training sleep. Change is normalised to the maximum change in the session. (C) Distributions of the change in pattern frequency as a function of the patterns' outcome prediction probability. Co-activation patterns from all ten learning sessions were binned by outcome prediction into variable size bins containing the same number of patterns. Each column is the cumulative probability density for the change in pattern frequency between pre- and post-training sleep, over all patterns in that bin. Circles give the median absolute change for each distribution. In this example, distributions were built using bins with 90 data-points each. Unbinned data are analysed in S2 Fig. (D) Correlation of outcome prediction and median change in pattern occurrence between pre- and post-training sleep from C. Red line is a linear regression ($P < 0.0002$, permutation test). (E)-(F) As C-D, for the worst-case correlation observed, using 25 data-points per bin. (G) Robustness of correlation results. Solid dots plot the correlation coefficient R between outcome prediction and median change in pattern frequency obtained for different binnings of the data; green colour-scale is proportional to the number of patterns per bin (light to dark: few to many patterns per bin, range 20-100). Asterisks indicate data points correspond to panels C-D and E-F. Lines each give the entire range of R obtained from a 5000-repeat permutation test; none reach the equivalent data point (dashed line shows equality), indicating all data correlations had $P < 0.0002$.

$P(Learn)$ in the post-learning trials, and test whether $D(Pre|Learn) > D(Post|Learn)$; that is, whether the post-learning distributions and post-training sleep distributions converge (Fig. 8A).

We indeed found that the two distributions converged. In 9 of the 10 learning sessions the post-learning distribution $P(Learn)$ was closer to the distribution in post-training sleep $[P(Post)]$ than in pre-training sleep $[P(Pre)]$ (Fig. 8B). On average the post-learning distribution of patterns was 20.5% (95% CI=[7.4,33.7])%]) closer to the post-training than the pre-training sleep distribution (Fig. 8E). Together, these results are consistent with the hypothesis that learning updates an internal model in mPfC, causing an increased stability of the probability distributions encoded in joint population activity.

Convergence of the probability distributions in sleeping and waking is a key prediction of our theory, as it supports both the prediction of increased stability of distributions over learning, and the hypothesis that sleep is sampling from a prior distribution generated by the internal model. Consequently, we sought to thoroughly check the robustness of this result. We used the Kullback-Liebler divergence to measure the distance $D(X|Y)$ between two distributions (X, Y) as it provides the most complete characterisation of that distance, but estimating it accurately from limited sample data has known issues [26]. These issues are relevant here as we

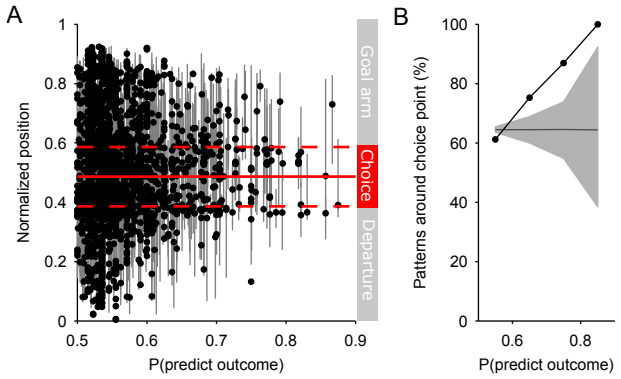

Fig 7. Outcome predicting activity patterns are sampled in the choice area. (*A*) Scatter plot of each pattern's outcome prediction and the positions of its occurrence in the maze (dot is median position; grey line is interquartile range); all positions are given as a proportion of the linearised maze from the start of the departure arm. Red lines indicate the approximate centre (solid) and boundaries (dashed) of the maze's choice area (cf Fig 1*A*). (*B*) Proportion of activity patterns whose interquartile range of positions enters the choice area (black dots and line). The grey region shows the mean (line) and 95% range (shading) of proportions from a permutation test. The data exceed the upper limit of expected proportions for all outcome-predictive patterns.

had a relatively small number of activity patterns in $P(Learn)$ (SI Table) due to the shortness of each trial (Fig. 1B), and some sessions had activity patterns up to 35 neurons in length.

We checked that our results were robust to different choices for measuring distance and the size of patterns. We re-computed all distances using the Hellinger distance, a non-parametric measure that provides a lower bound for the Kullback-Liebler divergence. Reassuringly, we found the same results: the post-learning distribution $P(Learn)$ of activity patterns was consistently closer to the distribution in post-training $[P(Post)]$ than in pre-training sleep $[P(Pre)]$ (S3 Fig; mean convergence 22.7%, 95% CI [15.4,29.9]%). Moreover, the Kullback-Liebler and Hellinger distance measures strongly correlated across sessions (S3 Fig). Similarly, when we re-constructed all distributions using a maximum of 15 neurons per pattern in each session, we found the same results (S4 Fig; mean convergence 36.8%, 95% CI [24.6,49]% using Kullback-Liebler divergence). Together, these checks suggest that the convergence was not an artifact of the issues in reliably estimating the Kullback-Liebler divergence.

Another possible source of issues was the choice of a 2 ms bin size for the activity patterns. We found that the convergence between the task $P(Learn)$ and post-training sleep $P(Post)$ distributions was robust to the choice of activity pattern bin size across an order of magnitude from 2 to 20 ms (S5 Fig). Our results thus do not depend on some arbitrary choice of bin size. Above a bin size of 50 ms, convergence was statistically indistinguishable from zero, meaning that the pre- and post-training sleep distributions are equidistant, on average, from the post-learning distribution. This suggests that the behaviourally relevant time-scales for activity patterns are indeed on the order of a few milliseconds.

Activity distributions do not converge during stable behaviour

In contrast to the learning sessions, our theory predicts that stable behaviour throughout a session likely represents at best minor changes to the underlying probability distributions in mPfC. Consequently, the stable-trial distribution $P(Stable)$ should be on average equidistant from those in pre- and post-training sleep, such that there is no convergence: $D(Pre|Stable) \approx D(Post|Stable)$ (Fig. 8C). In our data, there were 13 sessions with at least 90% of trials containing the same behavioural choice (left, right, or cued arm; Fig. 1D). In only 7 of the 13 stable sessions was the trial distribution $P(Stable)$ closer to the distribution in post-training sleep $[P(Post)]$ than in pre-training sleep $[P(Pre)]$ (Fig. 8*E*). On average the trial distribution of patterns was not closer to the post-training than the pre-training sleep distribution (mean convergence: 11.7%, 95% CI: [-11.7,35.2]%) (Fig. 8*E*). Lowering the threshold for identifying stable sessions to 85% trials with the same choice, giving 17 sessions, did not change the results (mean convergence: 5.8%, 95% CI:

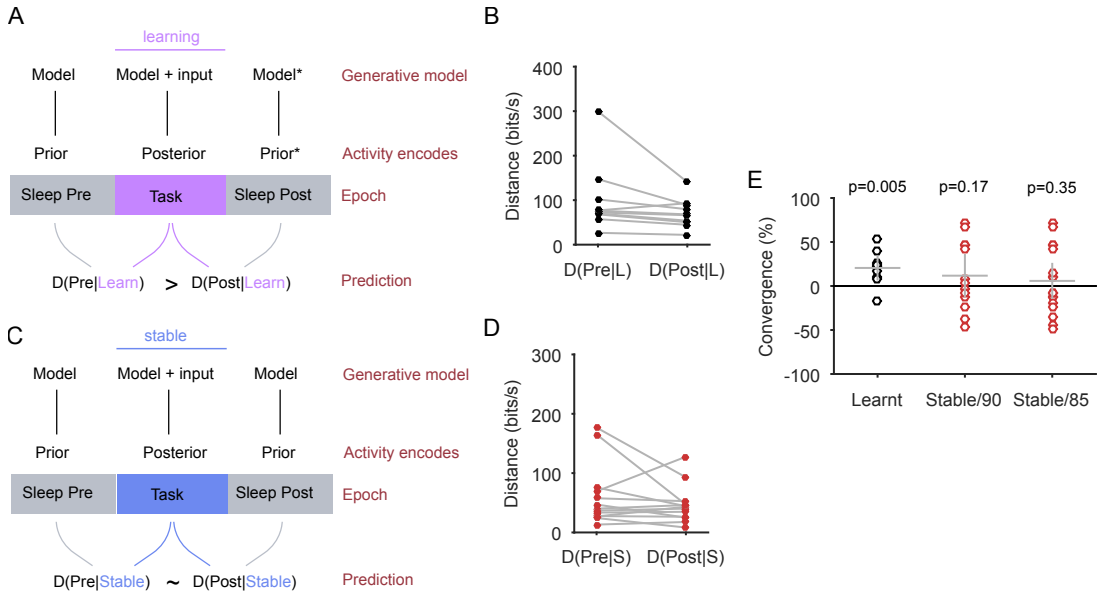

Fig 8. Differing convergence of activity pattern distributions between training and post-training sleep for learning and stable sessions. (A) Schematic of the theoretical prediction for the change in probability distributions between training and sleep epochs of a learning session. Reinforcement during training will change the internal model of the task, and these changes will be smaller after the correct strategy is acquired by the animal. Consequently, the distributions of joint population activity in post-training sleep and post-learning should be systematically closer than between those distributions in pre-training sleep and post-learning. (B) Distances between the distributions of pattern frequencies in sleep and training epochs; one dot per learning session. $D(X|Y)$: distance between pattern distributions in epochs X and Y: Pre: pre-training SWS; Post: post-training SWS; L: post-learning trials. (C) Schematic of the theoretical prediction for the change in probability distributions between training and sleep epochs of a stable session. Stable behavioural strategy implies a stable internal model in mPfC. Consequently, the distance between distributions of joint population activity in post-training sleep and training should be similar to the distance between the distributions in pre-training sleep and training. (D) Distances between the distributions of pattern frequencies in sleep and training epochs in all stable sessions (here with at least 90% of trials with the same choice). S: training trials. (E) Scatter of convergence between post-training sleep and post-learning trials across all learning and stable sessions (circles). Convergence is $[D(Pre|X) - D(Post|X)]/\max\{D(Pre|X), D(Post|X\}$, expressed as a percentage. A value greater than zero means that the training-epoch distribution X of activity patterns is closer to the distribution in post-training sleep than the distribution in pre-training sleep. Stable session results are plotted for both thresholds of 90% (13 sessions) and 85% (17 sessions). Grey lines give means and 95% confidence intervals.

[-13.6,25.2]%). Again, we found these results were robust to using the Hellinger distance (S3 Fig) and to using smaller activity patterns (S4 Fig). Thus, joint population activity during stable behaviour was consistent with the predicted lack of change to the internal model in mPfC.

Ruling out other causes of convergence

It seems remarkable that the sampling of temporally precise population activity patterns in prefrontal cortex could systematically change during learning. While these changes are consistent with our theory, they could also have a number of alternative explanations. Here we check three main alternatives: could they be explained as a "reverberation" of recent activity? By some form of selective replay of neural activity due to reward? Or by the change in brain state between waking and sleeping?

Convergence is not a recency effect

We examined periods of slow-wave sleep in order to most likely observe the sampling of a putative internal model in a static condition, with no external inputs and minimal learning. But as the post-learning trials by definition occur towards the end of a learning session, this raises the possibility that the closer match between training and post-training sleep distributions is a recency effect, due to some trace or reverberation in sleep of the most recent task activity.

There are two bits of evidence against this explanation. First, the time-scales involved make this unlikely. Bouts of slow-wave sleep did not start until typically 8 minutes after the end of the task (mean 397 s, S.D. 188 s; Fig. 1B). Any reverberation would thus have to last at least that long to appear in the majority of post-training slow-wave sleep distributions.

Second, we find no evidence of convergence between the activity in training and the intervening period before the first bout of slow-wave sleep. This "rest" epoch contains quiet wakefulness and early sleep stages. If convergence was just a recency effect, then we would expect that distributions $[P(Rest)]$ of activity patterns in this more-immediate "rest" epoch would also converge with the post-learning distributions. We did not find this: across sessions, there was no evidence that the distribution in post-training rest $[P(Rest)]$ consistently converged with the post-learning distribution $[P(Learn)]$ (Fig. 9; mean convergence: -4.6%, 95% CI: [-24.6,33.8]%). Thus the observed convergence is inconsistent with a recency effect.

Convergence is not a consequence of long runs of reward

A notable property of the learning sessions is that they contain long runs of successful trials. One alternative explanation for the convergence is that the post-training sleep replays activity that correlated with successful outcomes. If it did then the post-training sleep activity would be closer to the post-learning activity in the training epoch as this was when most of the successful outcomes occurred.

To answer test this explanation for convergence, we made use of the 8 sessions in which the rats experienced a rule change. As rule changes occurred only after 10 consecutive correct trials [15], these sessions contain long, sequential runs of rewards at the start, rather than the end, of the session. Consequently, if the post-training sleep is preferentially replaying activity that correlated with successful outcomes, then the activity pattern distributions of pre-change correct trials $[P(C)]$ and of post-task sleep $[P(Post)]$ should also converge: $D(Pre|C) > D(Post|C)$. However, we found no convergence (mean: 4.4%, 95% CI: [-16.2,25.1]%); Fig. 9). For the effect sizes observed for the learning sessions, there was sufficient power to recover the same effect size at $\alpha = 0.05$ with $N = 8$ sessions (KLD: learning session effect size $d = 0.96$, rule-change session power = 0.7; Hellinger: $d = 2.36$, power ≈ 1), which argues against low power causing the lack of convergence for the rule-change sessions.

Convergence is a consequence of changes to correlations, not just firing rates

Our convergence of distributions was measured across a change in brain state between waking and sleeping. While within each state the occurrence of co-activation patterns exceeds chance by an order of magnitude (Fig. 4C-E), this still leaves open the possibility that the change in population firing rates between states (Fig. 3) could artificially cause their activity pattern distributions to increase in similarity [20, 27]. To control for this, we used the "raster" model [20] to generate surrogate sets of spike-trains that matched both the mean

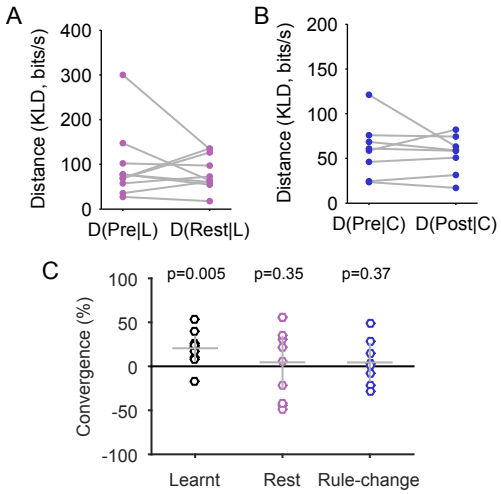

Fig 9. Convergence is not a consequence of recency or rewards. (A) Distances between the distributions of pattern frequencies in pre-training sleep and training epochs compared to the distances between post-training rest and training epochs; one dot per learning session. R: rest epoch. L: post-learning trials. (B) Distances between the distributions of pattern frequencies in pre-training sleep and pre-rule change epochs compared to the distances between pre-training sleep and pre-rule change epochs. C: pre-rule change epoch. (C) Results from panels A-B expressed as convergence, and compared to the convergence of post-learning trials and post-training sleep from panel C in Fig. 8. Grey lines give mean and 95% confidence intervals; P-values from 1-tailed Wilcoxon signrank test, with N=10 (Rest) and $N = 8$ (rule-change) sessions).

firing rates of each neuron, and the distribution of total population activity in each time-bin ($K = 0, 1, \ldots, N$ spikes per bin). Consequently, the occurrence rates of particular activity patterns in the raster model are those predicted to arise from neuron and population firing rates alone.

We fitted the raster model to the post-training sleep neuron and population firing rates. If the change in population firing rate during SWS caused the convergence, then the raster model should exactly capture the statistics of the SWS firing. This would predict that the distribution of activity patterns in the model and in the data are approximately equivalent $D(Model|Data) \approx 0$; and, consequently, that the convergence would be explained if $D(Post|Learn) \approx D(Post - model|Learn)$. We found that the distance between data and model-derived distributions in post-training sleep was always greater than baseline (Fig. 10A). Thus rate changes alone cannot account for the convergence between the training and post-training sleep distributions.

Our activity patterns were built from single units, unlike previous work using multi-unit activity [16, 20, 21, 23, 28], so we expected our patterns to be sparse with rare synchronous activity. Indeed our data are dominated by activity patterns with no spikes or one spike (Fig. 4B-E; we breakdown the distributions at 2 ms in S6 Fig). If all patterns had only no spikes or one spike, then the raster model spike trains would be exactly equivalent to the data. Given the relative sparsity ($\sim 1\%$) of co-activation patterns in our data, it is all the more surprising then that we found such a consistent difference between the model and data-derived distributions.

It follows that the true difference between data and model is in the relative occurrence of co-activation patterns. To check this, we applied the same analysis to distributions built only from these co-activation patterns, drawn from data and from the raster model fitted to the complete data. With the co-activation patterns, we found that the distance between data and model-derived distributions in post-training sleep was always greater than baseline (Fig. 10B). Consequently, we found that the data-derived distance $D(Post|Learn)$ was always smaller than the distance $D(Post - model|Learn)$ predicted by the raster model (Fig. 10C). These results indicate that much of the convergence between training and post-training sleep distributions could not be accounted for by firing rates alone; rather, the convergence is due to the selective changes of specific co-activation patterns.

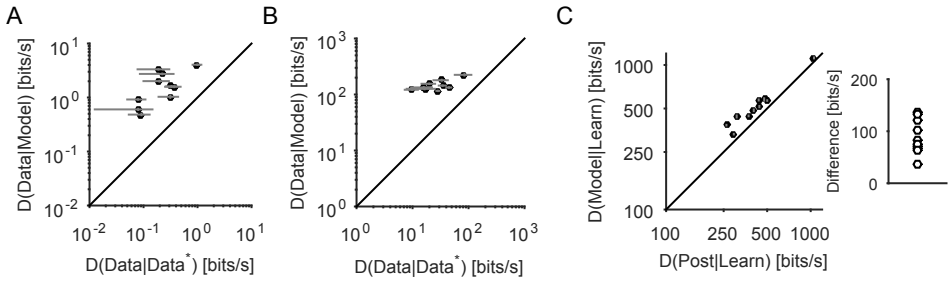

Fig 10. Convergence is caused by changes in correlation, not population firing rate. (*A*) Distances between model and data distributions for post-training sleep epochs (y-axis) for every learning session, compared to a per-session estimate of baseline differences (x-axis), obtained by bootstrap sampling of patterns within the post-training sleep epoch. Error bars give the mean and 95% confidence intervals on the bootstrapped within-epoch distance $D(Post|Post^*)$ [x-axis], and the 100 repeats of the raster model (y-axis). (*B*) As in *A*, using only activity patterns with $K \geq 2$ spikes from data and model. (*C*) The distance between the task and post-task sleep distributions $D(Post|Learn)$ is always smaller than predicted by population firing rate changes during sleep alone $D(Model|Learn)$, as given by the raster model. Error bars give the mean and 95% confidence intervals over the 100 repeats of the raster model, too small to see on this scale. Inset: plot of the difference between model mean and the data for each session: $D(Model|Learn)$ - $D(Post|Learn)$

Discussion

We have found converging evidence that mPfC contains a probabilistic internal model of behaviour. Our evidence rests on the hypothesis that neural populations represent probability distributions as samples, encoded by the moment-to-moment patterns of joint activity. Such precise patterns appeared far above those predicted by rates alone across sleeping and waking. Select patterns changed their frequency of occurrence in sleep epochs that occurred either side of behavioural learning. This select sub-set predicted trial outcomes, and appeared at the maze's decision-point. And their change correlated with learning, such that the distribution of patterns converged between sleep and post-learning trials. Our results thus match the predictions for how learning by reinforcement should update a probabilistic internal model. Consequently, they are evidence that mPfC population activity encodes an internal model of a task, and that this model is updated by learning.

Prefrontal cortex has been implicated in both planning and working memory during spatial navigation [29–32], and executive control in general [33, 34]. Our results suggests a probabilistic basis for these functions. In particular, prefrontal cortex has been implicated in both the representation of current goals [35, 36] and strategies [37]. Both these functions are consistent with an internal model that relates sensory information to the statistical structure of the world, and the use of that model to plan behaviour.

Further probing the hypothesis of internal models in mPfC

Our theoretical account makes further testable predictions that should reveal the extent to which it is useful. The simplest and strongest prediction is that the probability distribution within a training epoch should become more similar between consecutive sessions as the task is learnt (The SI Text demonstrates these predictions in the probabilistic reinforcement learning model). Testing this prediction would require tracking an identical population of neurons across multiple days of behavioural training. Confidently isolating an identical group of neurons is just out of reach of current electrophysiological tools; but new technological advances, such as ultra-high density probes [38], could make this prediction testable soon.

Another prediction that requires identical populations is that the distance between sleep distributions in stable sessions should be smaller than the corresponding distance in learning sessions (Fig. 2F). This would directly test the hypothesis that sleep samples from the prior distribution generated from the internal model: as the internal model should not markedly change once the correct behaviour is acquired. We cannot currently test this prediction, as again we would need the same set of neurons tracked across multiple days of behavioural training. Indeed, the measured distances between distributions change as a function of the number of neurons even when the neurons are taken from a single recording (S7 Fig), let alone between different sets of neurons in consecutive sessions - we await the advent of stable electrophysiological recordings

across multiple days.

Testing this stable-session prediction would also require a set of sessions in which the stable strategy was also the correct strategy. Only then could one be confident that the stable behaviour should correspond to a stable internal model, as the model and feedback match. Obtaining a match of stable behaviour and rule would mean adjusting the task to leave the animals at asymptotic behaviour for multiple sessions, before changing the rewarded rule.

What we may still learn if we are wrong

We are not unaware that there is an extensive theoretical apparatus underpinning the predictions we test here. But this seems inevitable if systems neuroscience is to move towards an hypothesis-driven era, simply due to the need to simultaneously account for behaviour, corresponding neural activity, and their coincident (or not) changes upon learning. Compounding this complexity is the requirement that a testable theory must posit a computational problem, an algorithm, and its neural implementation. Here we examine the implications of those requirements, and what we may still learn from these results if the specific choices are wrong in detail, but right in substance, or wrong in toto.

At the most abstract level, our hypothesis is that the mPfC encodes probabilistic internal models of the world, that are updated by reinforcement. Our specific simulated model used abstract representations of strategies as an illustration of probabilistic coding, but our hypotheses do not rest on knowing the specific representations of behaviour in mPfC; they only require that the internal model is represented using probabilities. The immediate prediction is that the encoded probability distributions will stabilise upon successfully learning the relevant internal model. There are other theoretical ways in which neural populations can encode probability distributions [2, 6, 39]. But few are amenable to direct testing by experiment without numerous additional assumptions. Consequently, our hypothesis of internal models may still be true, even if our specific implementation hypothesis is false.

But this is not very satisfactory. By choosing the specific sampling implementation for the probability distributions, we have made our internal model hypothesis falsifiable in principle. Indeed, we have shown here that population activity in other sets of sessions do not show the changes predicted by our model for learning sessions (Figs. 8, 9), nor did the same learning sessions systematically converge for periods of awake, resting activity (Fig. 9). Our data could thus falsify the sampling hypothesis.

We chose the sampling hypothesis for two reasons. First, there are good models for how a generic cortical circuit can sample from an underlying probability distribution [7, 17], making it a candidate computational principle for cortex. Second, because the sampling hypothesis makes it easy to check for changes to the hypothesised sampled probability distributions, by computing the distances between distributions of population joint activity patterns. But with these strengths comes the technical issue that the theoretically best distance estimator - the Kullback-Liebler divergence - is also the most difficult to measure accurately with finite samples. To counteract this, we have extensively checked its behaviour (see also S7 Fig), and re-checked our key results with a different, non-parametric distance measure.

Even if the general hypothesis of probabilistic internal models in prefrontal cortex turns out to be wrong, our data provide constraints on the dynamics of cortex. Studies of prefrontal cortex coding generally assume that information is conveyed by firing rates [31, 32, 40, 41] or rate correlations [29, 42, 43]. By contrast, here we show evidence of ensemble coding at highly precise time scales, of both outcome and position dependence. We found it remarkable that we could extract anything of interest at this resolution, and checked these results extensively, including the use of large-repeat permutation tests. Previously, such fine-scale structure of stimulus-evoked population activity patterns has only been observed in the retina and V1 during passive observation [16, 21, 23, 24]. We extend these results to show that such fine time-scale correlation structure can be observed in cortical regions for executive control, and be evoked by tasks.

Previous studies have observed strong similarities between spontaneous and evoked firing rates [44–47] or firing sequences [19] in cortex. These findings imply that the underlying cortical circuit has similarly constrained dynamics in both spontaneous and evoked states [48]. Extending these results, we found a highly similar set of precisely-timed activity patterns across sleeping and task performance, which suggests that cortical population activity is underpinned by similar dynamics in both states, and those dynamics can reproduce patterns with high temporal precision. Maass and colleagues [7, 17] have shown that a range of cortical network models can produce specific distributions of such precise activity patterns, provided they have

a source of noise (such as synaptic release failure) to produce stochastic wandering of the global activity level. Our data support these models, and suggest that global activity oscillations during slow-wave sleep [49, 50] do not prevent the stochastic sampling of activity patterns, providing a target for future modelling studies.

Replay and resampling

Our proposal that the spontaneous activity of sleep is sampling from an internal model suggests an alternative interpretation of "replay" phenomena [15,18]. Replay of neural activity during waking in a subsequent episode of sleep has been inferred by searching for matches of patterns of awake activity in sleep activity. The better match of waking activity with subsequent sleep than preceding sleep has been taken as evidence that replay is encoding recent experience, perhaps to enable memory consolidation. However, our observation that the distributions of patterns in stable sessions' trials are not specifically sampled in post-training sleep (Fig. 8) is incompatible with the simple replay of experience-related activity in sleep.

By contrast, our proposal suggests that the similarity between waking and sleep activity is due to the stabilisation of the internal model, not recent experience per se. The similarity of the patterns in waking and subsequent sleep is then caused by sampling from a similar model, not by explicitly recalling patterns that occurred in waking activity. Indeed, if our proposal is true, then it suggests there may be situations where we observe "pre-play" of waking activity in preceding sleep activity. Our observation that sessions with stable behaviour show no convergence of waking and post-training sleep distributions is compatible with this: in those sessions, pre- and post-training sleep distributions were equidistant on average from the waking distribution, and so potentially both pre-play and replay could be observed. Our theory is thus suggesting that replay may be a signature of resampling.

Implications for the probabilistic brains hypothesis

How a cortical region encodes an internal model is an open question. A strong candidate, assumed by the sampling hypothesis, is the relative strengths of the synaptic connections both into and within the encoding cortical circuit [7,8,11,17]. The activity of a cortical circuit is strongly dependent on the pattern and strength of the connections between its neurons [51, 52]. Consequently, defining the underlying model as the circuit's synaptic network allows both model-based inference through synaptically-driven activity and model learning through synaptic plasticity [11].

While the hypothesis that brains compute using probabilities is widely-discussed, most evidence for it has been from observations of behaviour that is consistent with probabilistic inference. Strong evidence for probabilistic brains requires detecting the representation and use of probability distributions in circuit-level neural activity [39]. We have presented here initial experimental evidence that neural populations represent probabilistic internal models, and update those models through learning. Our results advance the case that probabilistic internal models are a candidate general computational principle of cortex.

Materials and Methods

Task and electrophysiological recordings

Four Long-Evans male rats with implanted tetrodes in prelimbic cortex were trained on the Y-maze task (Fig. 1*A*). Each recording session consisted of a 20-30 minute sleep or rest epoch (pre-training epoch), in which the rat remained undisturbed in a padded flowerpot placed on the central platform of the maze, followed by a training epoch, in which the rat performed for 20-40 minutes, and then by a second 20-30 minute sleep or rest epoch (post-training epoch); see (Fig. 1*B*). Every trial started when the rat reached the departure arm and finished when the rat reached the end of one of the choice arms. Correct choice was rewarded with drops of flavoured milk. Each rat had to learn the current rule by trial-and-error, either: go to the right arm; go to the cued arm; go to the left arm; go to the uncued arm. To maintain consistent context across all sessions, the extra-maze light cues were lit in a pseudo-random sequence across trials, whether they were relevant to the rule or not.

The data analysed here were from a total set of 50 experimental sessions taken from the study of [15], representing a set of training sessions from naive until either the final training session, or until choice became

habitual (consistent selection of one arm that was not the correct arm). The four rats respectively had 13, 13, 10, and 14 sessions. From these we have used here ten learning sessions, eight rule change sessions, and up to 17 "stable" sessions (see below).

Tetrode recordings were spike-sorted only within each recording session for conservative identification of stable single units. In the sessions we analyse here, the populations ranged in size from 15-55 units. Spikes were recorded with a resolution of 0.1 ms. For full details on training, spike-sorting, and histology see [15].

Session selection and strategy analysis

We primarily analysed here data from the ten sessions in which the previously-defined learning criteria were met: the first trial of a block of at least three consecutive rewarded trials after which the performance until the end of the session was above 80%. In later sessions the rats reached the criterion for changing the rule: ten consecutive correct trials or one error out of 12 trials. Thus each rat learnt at least two rules, with eight rule-change sessions in total.

We also sought sessions in which the rats made stable choices of strategy. For each session, we computed the probability $P(rule)$ that the rat chose each of the three rules (left, right, cued arm) per trial. Whereas $P(left)$ and $P(right)$ are mutually exclusive, $P(cued-arm)$ is not, and has an expected value of 0.5 when it is not being explicitly chosen because of the random switching of the light cue. A session was deemed to be "stable" if $P(rule) > \theta$ for one of the rules. Here we tested both $\theta = 0.9$ and $\theta = 0.85$, giving $N = 13$ and $N = 17$ sessions respectively. These also respectively included 2 and 4 of the rule-change sessions. For the time-series in Fig. 1D,E, we estimated $P(rule)$ in 7-trial windows, starting from the first trial, and sliding by one trial.

Probabilistic reinforcement learning model

To illustrate the expected behaviour of a probabilistic internal model during learning, we constructed a Bayesian reinforcement learning model of the Y-maze task. We modelled the trial-by-trial behaviour as a Bayesian multi-arm bandit problem [53], where the agent's task on each trial was to chose which strategy to adopt, based on a probabilistic estimate of the value of each strategy. We use this simplified representation as a proxy for more complex models with probability distributions over the uncertain values of individual actions and the transitions they cause between states in the maze, which collectively make a strategy.

Here we report results from modelling three strategies: go to the left arm; go to the right arm; and go to the cued arm. For each strategy x, the agent maintained a posterior probability distribution over the value of choosing that strategy $V_x \in [0, 1]$, given by a Beta distribution $P(V_x)$ with parameters (α_x, β_x). On each trial t, the winning strategy was chosen using Thompson sampling: a random value ζ_x was sampled from the probability distribution $P(V_x)$ for each strategy, and the strategy s with the highest sampled value was chosen. The corresponding action was then chosen: left, right, or cued arm (where, as per the experiment, the cued arm was randomly chosen on each trial). There was a small probability η of a mistake in choosing the corresponding action: if a mistake was made, then the opposite action was chosen (being the uncued arm for the cued-arm strategy). We used $\eta = 0.2$ for the simulations reported here. This was implemented to include noise into the decision process, providing a better replication of the rats' behaviour (see SI Text). Having taken the action, the agent received reward according to the current rule (left, right, or cued arm), with $R = 1$ if the action corresponded to the rule, and $R = 0$ otherwise. The reward was then used to update the probability distribution $P(V_s)$ of the chosen strategy s.

The full Bayesian update of the posterior should be proportional to $P(V_s|R = r) \propto P(R = r|V_s)P(V_s)$, where $P(R = r|V_s)$ is the likelihood function for the outcome r given the probability distribution over the strategy's value, and $P(V_s)$ is the prior distribution over that value.

In simulation, we make use of the standard result that, assuming a binomial likelihood function $P(R = r|V_s)$ because each trial is a Bernoulli trial, then the Beta distribution $P(V_s)$ is the conjugate prior. Consequently, Bayesian updating is obtained by just updating the parameters of $P(R = r|V_s)$ by $(\alpha + r, \beta + (1 - r))$ [53,54]. Distributions $P(V_x)$ for trial 1 was set to the uniform distribution $(\alpha = 1, \beta = 1)$.

To make comparisons with the behavioural data, we made proxy estimates of learning trials, and then virtual "sessions" around those trials. For each simulation, the nominal "learning trial" was identified as the trial in the cumulative reward curve corresponding to the greatest inflection in reward rate. To do this, we

fitted a piecewise linear slope around each trial t, with one line fitted to eleven trials before and including t, and one line fitted to eleven trials after and including t. The trial t_l with the greatest increase in slope from the before to the after line was selected as the "learning" trial.

A virtual session was given by the 14 trials before and after the chosen learning trial, giving a session length of 29 trials. The trials corresponding to the beginning (t_{pre}) and end (t_{post}) of this virtual session were deemed the pre- and post-training "sleep" epochs for the model.

Activity pattern distributions

For a population of size N, we characterised population activity from time t to $t+\delta$ as an N-length binary vector with each element being 1 if at least one spike was fired by that neuron in that time-bin, and 0 otherwise. In the main text we predominantly use a bin size of $\delta = 2$ ms; Fig. 8 shows the robustness of the main results to the choice of bin size. We build patterns using the number of recorded neurons N, up to a maximum of 35 for computational tractability. The probability distribution for these activity patterns was compiled by counting the frequency of each pattern's occurrence and normalising by the total number of pattern occurrences. The number of neurons used in each analysis is listed in the SI Table; where we needed to use less than the total number of recorded neurons, we ranked them according to their coefficient of variation of their firing rate between the three epochs, and choose the M least variable; in practice this sampled neurons from across the full range of firing rates.

To test the predicted proportion of co-activation patterns by independently firing neurons, we shuffled inter-spike intervals for each neuron independently, then reconstruct the activity patterns at the chosen bin size. This procedure keeps the same inter-spike interval distribution for each neuron, but disrupts any correlation between neurons. As both the training and sleep epochs were broken up into chunks (of trials and SWS bouts, respectively), we only shuffled inter-spike intervals within each chunk. We repeated the shuffling 20 times, and in Fig. 4*C-E* we plot for the shuffled data the means and error bars of $\pm$ 2 s.e.m. (too small to see on the scales of the axes).

Comparing distributions

We quantified the distance $D(P|Q)$ between probability distributions P and Q using both the Kullback-Liebler divergence (KLD) and the Hellinger distance.

The KLD is an information theoretic measure to compare the similarity between two probability distributions. Let $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ be two discrete probability distributions, for n distinct possibilities – for us, these are all possible individual activity patterns. The KLD is then defined as $D_{\mathrm{KLD}}(P|Q) = \sum_{i=1}^{n} p_i \log_2(\frac{p_i}{q_i})$. We normalised this by unit time (2 ms bins except where noted) to obtain the information rate in bits/s.

There are 2^N distinct possible activity patterns in a recording with N neurons. Most of these activity patterns are never observed, so for computational tractability we exclude the activity patterns that are not observed in either of the epochs we compare. The empirical frequency of the remaining activity patterns is biased due to the limited length of the recordings [26]. To counteract this bias, we use the Bayesian estimator and quadratic bias correction exactly as described in [16]. The Berkes estimator assumes a Dirichlet prior and multinomial likelihood to calculate the posterior estimate of the KLD; we use their code (github.com/pberkes/neuro-kl) to compute the estimator. We then compute a KLD estimate using all S activity patterns, and using $S/2$ and $S/4$ patterns randomly sampled without replacement. By fitting a quadratic polynomial to these three KLD estimates, we can then use the intercept term of the quadratic fit as an estimate of the KLD if we had access to recordings of infinite length [26, 55]. This final estimate varies according to the patterns sub-sampled in order to fit the quadratic; however, in our data the variation introduced by the sub-sampling is negligible on the scale of the distances measured (S7 FigC).

We attempted here to characterise the population's joint activity as fully as possible, by making use of as many simultaneously recorded individual neurons as possible. We capped our activity patterns to a maximum of $N = 35$ neurons; but this still means that, for some populations, a full estimation of KLD using the above Bayesian estimator would mean enumerating all 2^{35} patterns every time we computed a KLD estimate. This is computationally intractable; moreover, in extensively checking the results and the raster model (see below) we produced thousands of KLD calculations for each population. So we sought a practical

solution, and set $P = 0$ for all activity patterns that were not in both distributions being compared. Our data shows only a tiny fraction of activity patterns that appear in one distribution and do not appear in the other, so we expected the disagreement between KLD computed using the full enumeration of all 2^N patterns and using $P = 0$ to be small, and not to qualitatively affect results. We tested this explicitly for a full enumeration using $N = 15$ for all learning-session populations, and found that setting $P = 0$ did not qualitatively affect the results, nor showed a systematic bias in the distances measured by either approach (S7 FigD). We note that this is not, in general, a safe assumption: we can only do this here because of the very low proportion of unique patterns in each compared distribution. Moreover, we checked the main results throughout with a different measure of inter-distribution distance - the Hellinger distance - that did not rely on any bias-correcting estimators or priors.

The Hellinger distance for two discrete distributions P and Q is $D_{\mathrm{H}}(P|Q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2$. To a first approximation, this measures for each pair of probabilities (p_i, q_i) the distance between their square-roots. In this form, $D_{\mathrm{H}}(P|Q) = 0$ means the distributions are identical, and $D_{\mathrm{H}}(P|Q) = 1$ means the distributions are mutually singular: all positive probabilities in P are zero in Q, and vice-versa. The Hellinger distance is a lower bound for the KLD: $2D_{\mathrm{H}}(P|Q) \leq D_{\mathrm{KLD}}$. We observed that, for our data, there was a consistently strong correlation between the Hellinger distance and the KLD (S3 Fig), further suggesting that the issues in estimating accurate KLD did not affect our main results.

To compare distances between sessions we computed a normalised measure of "convergence". The distance between a given pair of distributions could depend on many factors that differ between sessions, including that each recorded population was a different size (S7 FigA,B), and how much of the relevant population for encoding the internal model we recorded. Consequently, the key difference between the distances $D(Pre|X) - D(Post|X)$ also depends on these factors. To compare the difference in distances across sessions, we computed a "convergence" score by normalising the difference by the scale of the maximum distance between training and sleep epochs: $[D(Pre|X) - D(Post|X)] \,/\, \max\{D(Post|X), D(Pre|X)\}$. We express this as a percentage, giving a range of $[-100, 100]\%$. Convergence greater than 0% indicates that the distance between the training epoch $P(X)$ and post-training sleep ($P(Post)$) distributions is smaller than that between the training and pre-training sleep (P(Pre)) distributions.

Estimating equivalence between distributions with finite samples

Even if two underlying probability distributions are exactly the same, empirical measurements of samples taken from them will not show exact equivalence $[D(P|Q) = 0]$ due to finite sampling effects. We estimated a baseline measure of equivalence for the activity distributions in the sleep epochs by bootstrapping the activity patterns within each epoch. To do this, we drew two sets of patterns with replacement from the set of empirically recorded patterns, and computed the distance between the two bootstrapped sets. This emulates the finite-sampling problem within the empirical data. We also tested a more severe version where the set of recorded activity patterns was split randomly in half and the distance computed between each half. However, as this procedure is itself halving the number of patterns, it induces more variation by further finite sampling; we plot these results in S1 Fig.

Outcome prediction

We examined the correlates of activity pattern occurrence with behaviour. To rule out pure firing rate effects, we excluded all patterns with $K = 0$ and $K = 1$ spikes, considering only co-activation patterns $K \geq 2$; that is, those with two or more active neurons.

To check whether individual activity patterns coded for the outcome on each trial, we used standard receiver-operating characteristic (ROC) analysis. For each pattern, we computed the distribution of its occurrence frequencies separately for correct and error trials (as in the example of Fig. 6A). We then used a threshold T to classify trials as error or correct based on whether the frequency on that trial exceeded the threshold or not. We found the fraction of correctly classified correct trials (true positive rate) and the fraction of error trials incorrectly classified as correct trials (false positive rate). Plotting the false positive rates against the true positive rates for all values of T gives the ROC curve. The area under the ROC curve gives the probability that a randomly chosen pattern frequency will be correctly classified as from a correct trial; we report this as P(predict outcome).

Relationship of sampling change and outcome prediction

Within each session, we computed the absolute change $\delta_i = |p_i(pre) - p_i(post)|$ in each pattern's probability of occurrence between pre- and post-training SWS. To combine data across sessions, for each session we normalised all changes by the maximum change in that session: $\delta_i^* = \delta_i / \max_i\{\delta\}$. Normalised change scores were pooled over all learning sessions. Correlating these change scores against P(predict outcome) showed that the better a pattern predicted trial outcome, the more it tended to change probability between pre- and post-training SWS (S2 Fig). But as most patterns had little change and little prediction of outcome, this correlation was skewed.

Consequently, to better characterise the distributions of change between pre- and post-session sleep, we binned δ_i^* using variable-width bins of P(predict outcome): each consecutive bin-width was chosen in order to contain the same number of data-points in every bin. We computed the empirical cumulative distribution in each bin, to visualise the distribution of changes in pattern probability between sleep epochs, and the change in that distribution with P(predict outcome). To quantify this change, we regressed P(predict outcome) against the median change in each bin; we used the mid-point of each variable-width bin as the value for P(predict outcome). Our main claim is that prediction and change are dependent variables (Fig. 6*C*-*G*). To test this claim, we compared the data correlation against the null model of independent variables, by permuting the assignment of change scores to the activity patterns. For each permutation, we repeat the binning and regression. We permuted 5000 times to get the sampling distribution of the correlation coefficient R^* predicted by the null model of independent variables. To check robustness, all analyses were repeated for a range of fixed number of data-points per bin between 20 and 100.

Relationship of location and outcome prediction

The location of every occurrence of a co-activation pattern was expressed as a normalized position on the linearised maze (0: start of departure arm; 1: end of the chosen goal arm). Our main claim is that activity patterns strongly predictive of outcome occur predominantly around the choice point of the maze, and so prediction and overlap of the choice area are dependent variables (Fig. 7*B*). To test this claim, we compared this relationship against the null model of independent variables, by permuting the assignment of location centre-of-mass (median and interquartile range) to the activity patterns. For each permutation, we compute the proportion of patterns whose interquartile range overlaps the choice area, and bin as per the data. We permuted 5000 times to get the sampling distribution of the proportions predicted by the null model of independent variables: we plot the mean and 95% range of this sampling distribution as the grey region in Fig. 7*B*.

Raster model

To control for the possibility that changes in activity pattern occurrence were due solely to changes in the firing rates of individual neurons and the total population, we used the raster model exactly as described in [20]. For a given data-set of spike-trains N and bin size δ, the raster model constructs a synthetic set of spikes such that each synthetic spike-train has the same mean rate as its counterpart in the data, and the distribution of the total number of spikes per time-bin matches the data. In this way, it predicts the frequency of activity patterns that should occur given solely changes in individual and population rates.

For Fig. 10 we generated 1000 raster models per session using the spike-trains from the post-training SWS in that session. For each generated raster model, we computed the distance $D(Model|Data)$ between the distribution of patterns for that model $P(Model)$ and the corresponding data distribution $P(Data)$ of post-training SWS patterns. For each generated raster model, we then computed the distance between its distribution of activity patterns and the data distribution for post-learning trials $D(Post - model|Learn)$. This comparison gives the expected distance between the training and post-training SWS distributions due to firing rate changes alone. We plot the difference between the mean of $D(Post - model|Learn)$ over the 1000 raster models and the data $D(Post|Learn)$ in Fig. 10.

Statistics

Quoted measurement values are mean $\bar{x}$ and 95% confidence intervals for the mean $[\bar{x} - t_{\alpha/2,n}\mathrm{SE}, \bar{x} + t_{\alpha/2,n}\mathrm{SE}]$, where $t_{\alpha/2,n}$ is the value from the t-distribution at $\alpha = 0.05$ and given the number n of data-points used to obtain $\bar{x}$. All hypothesis tests used the non-parametric Wilcoxon signtest for a one-sample test that the sample median for the population of sessions is greater than zero. We used this one-tailed test throughout for the change in convergence, as the key prediction is that convergence is greater than 0% for the learning sessions. For learning sessions, we have $n = 10$ sessions; for rule-changes (Fig. 9) we have $n = 8$ sessions. For stable sessions we have $n = 13$ for $\theta = 0.9$ and $n = 17$ for $\theta = 0.85$.

Supporting Information

S1 Fig. Distribution changes between pre- and post-learning sleep tested against split data.

S2 Fig. Joint distribution of outcome prediction and change in sampling.

S3 Fig. Robustness of convergence between post-learning and post-training sleep distributions to the choice of distance measure.

S4 Fig. Robustness of convergence between post-learning and post-training sleep distributions to the number of neurons sampled per session.

S5 Fig. Robustness of learning-session convergence to the choice of bin size.

S6 Fig. Distributions of activity patterns at 2 ms bin size.

S7 Fig. Checking issues with the Kullback-Liebler divergence.

S1 Table. Numbers of neurons in each session and of activity patterns in each epoch.

S1 File. Behaviour and further predictions of the probabilistic reinforcement learning model.

Acknowledgments

We thank the Humphries lab (Javier Caballero, Mat Evans, Silvia Maggi) for discussions; Rasmus Petersen for comments on the manuscript; and P. Berkes and M. Okun for respectively making their KL divergence and raster model code publicly available.

References

1. Kording KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature. 2004;427:244–247.
2. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nat Neurosci. 2013;16:1170–1178.
3. Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. Science. 1995;269:1880–1882.
4. Dayan P, Abbot LF. Theoretical Neuroscience. anonymous, editor. Cambridge, MA: MIT Press; 2001.
5. Zemel RS, Dayan P, Pouget A. Probabilistic interpretation of population codes. Neural Comput. 1998;10:403–430.

6. Ma WJ, Beck JM, Latham PE, Pouget1 A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006;9:1432–1438.

7. Buesing L, Bill J, Nessler B, Maass W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput Biol. 2011;7:e1002211.

8. Kappel D, Habenschuss S, Legenstein R, Maass W. Network Plasticity as Bayesian Inference. PLoS Comput Biol. 2015;11:e1004485.

9. Haefner RM, Berkes P, Fiser J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron. 2016;90:649–660.

10. Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, et al. Probabilistic population codes for Bayesian decision making. Neuron. 2008;60:1142–1152.

11. Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci. 2010;14:119–130.

12. Ragozzino ME, Detrick S, Kesner RP. Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. J Neurosci. 1999;19:4585–4594.

13. Rich EL, Shapiro ML. Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. J Neurosci. 2007;27:4747–4755.

14. Benchenane K, Peyrache A, Khamassi M, Tierney PL, Gioanni Y, Battaglia FP, et al. Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. Neuron. 2010;66:921–936.

15. Peyrache A, Khamassi M, Benchenane K, Wiener SI, Battaglia FP. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat Neurosci. 2009;12:916–926.

16. Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science. 2011;331:83–87.

17. Habenschuss S, Jonke Z, Maass W. Stochastic computations in cortical microcircuit models. PLoS Comput Biol. 2013;9:e1003311.

18. Euston DR, Tatsuno M, McNaughton BL. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. Science. 2007;318:1147–1150.

19. Luczak A, Barthó P, Harris KD. Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron. 2009;62:413–425.

20. Okun M, Yger P, Marguet SL, Gerard-Mercier F, Benucci A, Katzner S, et al. Population rate dynamics and multineuron firing patterns in sensory cortex. J Neurosci. 2012;32:17108–17119.

21. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440:1007–1012.

22. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, et al. The structure of multi-neuron firing patterns in primate retina. J Neurosci. 2006;26:8254–8266.

23. Tkacik G, Marre O, Amodei D, Schneidman E, Bialek W, Berry MJ 2nd. Searching for collective behavior in a large network of sensory neurons. PLoS Comput Biol. 2014;10:e1003408.

24. Marre O, Botella-Soler V, Simmons KD, Mora T, Tkacik G, Berry MJ 2nd. High Accuracy Decoding of Dynamical Motion from a Large Retinal Population. PLoS Comput Biol. 2015;11:e1004304.

25. Ohiorhenuan IE, Mechler F, Purpura KP, Schmid AM, Hu Q, Victor JD. Sparse coding and high-order correlations in fine-scale cortical networks. Nature. 2010;466:617–621.

26. Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the sampling bias problem in spike train information measures. J Neurophysiol. 2007;98:1064–1072.

27. Fiser J, Lengyel M, Savin C, Orbán G, Berkes P. How (not) to assess the importance of correlations for the matching of spontaneous and evoked activity. 2013; p. arXiv:1301.6554.

28. Ganmor E, Segev R, Schneidman E. A thesaurus for a neural population code. Elife. 2015;4:e06134.

29. Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW. Dynamics of population code for working memory in the prefrontal cortex. Neuron. 2003;40:177–188.

30. Fujisawa S, Amarasingham A, Harrison MT, Buzsáki G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat Neurosci. 2008;11:823–833.

31. Ito HT, Zhang SJ, Witter MP, Moser EI, Moser MB. A prefrontal-thalamo-hippocampal circuit for goal-directed spatial navigation. Nature. 2015;522:50–55.

32. Spellman T, Rigotti M, Ahmari SE, Fusi S, Gogos JA, Gordon JA. Hippocampal-prefrontal input supports spatial encoding in working memory. Nature. 2015;522:309–314.

33. Miller EK. The prefrontal cortex and cognitive control. Nat Rev Neurosci. 2000;1(1):59–65.

34. Sul JH, Kim H, Huh N, Lee D, Jung MW. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron. 2010;66:449–460.

35. Matsumoto K, Suzuki W, Tanaka K. Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science. 2003;301:229–232.

36. Hok V, Save E, Lenck-Santini PP, Poucet B. Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. Proc Natl Acad Sci U S A. 2005;102:4602–4607.

37. Rich EL, Shapiro M. Rat prefrontal cortical neurons selectively code strategy switches. J Neurosci. 2009;29:7208–7219.

38. Jun JJ, Mitelut C, Lai C, Gratiy S, Anastassiou C, Harris TD. Real-time spike sorting platform for high-density extracellular probes with ground-truth validation and drift correction. bioRxiv. 2017;.

39. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 2004;27:712–719.

40. Pinto L, Dan Y. Cell-Type-Specific Activity in Prefrontal Cortex during Goal-Directed Behavior. Neuron. 2015;87:437–450.

41. Siegel M, Buschman TJ, Miller EK. Cortical information flow during flexible sensorimotor decisions. Science. 2015;348:1352–1355.

42. Averbeck BB, Sohn JW, Lee D. Activity in prefrontal cortex during dynamic selection of action sequences. Nat Neurosci. 2006;9:276–282.

43. Baeg EH, Kim YB, Kim J, Ghim JW, Kim JJ, Jung MW. Learning-induced enduring changes in functional connectivity among prefrontal cortical neurons. J Neurosci. 2007;27:909–918.

44. Tsodyks M, Kenet T, Grinvald A, Arieli A. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. Science. 1999;286:1943–1946.

45. Hromádka T, Deweese MR, Zador AM. Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol. 2008;6:e16.

46. O'Connor DH, Peron SP, Huber D, Svoboda K. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. Neuron. 2010;67:1048–1061.

47. Wohrer A, Humphries MD, Machens C. Population-wide distributions of neural activity during perceptual decision-making. Prog Neurobiol. 2013;103:156–193.

48. Marre O, Yger P, Davison AP, Frégnac Y. Reliable Recall of Spontaneous Activity Patterns in Cortical Networks. J Neurosci. 2009;29:14596–14606.

49. Destexhe A, Contreras D, Steriade M. Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states. J Neurosci. 1999;19:4595–4608.

50. Steriade M, Timofeev I, Grenier F. Natural waking and sleep states: a view from inside neocortical neurons. J Neurophysiol. 2001;85:1969–1985.

51. Cossell L, Iacaruso MF, Muir DR, Houlton R, Sader EN, Ko H, et al. Functional organization of excitatory synaptic strength in primary visual cortex. Nature. 2015;518:399–403.

52. Okun M, Steinmetz NA, Cossell L, Iacaruso MF, Ko H, Barthó P, et al. Diverse coupling of neurons to populations in sensory cortex. Nature. 2015;521:511–515.

53. Ghavamzadeh M, Mannor S, Pineau J, Tamar A. Bayesian Reinforcement Learning: A Survey. Foundations and Trends in Machine Learning. 2015;8:359–492.

54. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8(12):1704–1711.

55. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. Phys Rev Lett. 1998;80:197–200.

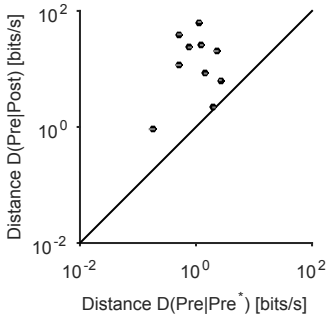

Figure S1: Distributions of joint activity patterns change between pre- and post-learning sleep. Distances between pre- and post-training sleep distributions (y-axis) for every learning session, compared to a per-session estimate of baseline differences (x-axis). Here the baseline difference was obtained by randomly dividing patterns within the pre-training sleep epoch into two equal groups, and computing the distance $D(Pre|Pre^*)$ between the two groups. Note that this further reduces the number of sampled patterns used to calculate the two distributions, and so further increases the variance in estimating the Kullback-Liebler divergence. All symbols lie above the diagonal. Error bars give the mean and its 95% confidence interval over 100 repeats of randomly choosing the two groups to compute $D(Pre|Pre^*)$; on this scale, the bars are the size of the symbols. Identical results were obtained when using $D(Post|Post^*)$.

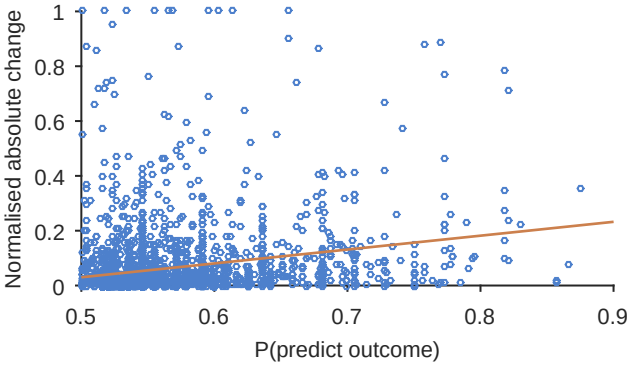

Figure S2: Joint distribution of outcome prediction and change in sampling. Here we plot every co-activation pattern's joint values of P(predict outcome) and the absolute normalised change in sampling between pre- and post-training slow-wave sleep ($N = 2353$ patterns with $K \geq 2$ spikes per pattern across all 10 sessions). The linear regression in red indicates a clear relationship between the two ($R = 0.22$, $P < 10^{-27}$). Nonetheless, the majority of patterns do not markedly change their sampling, nor are they predictive of outcome: 72% (1699/2353) have P(predict outcome) ≤ 0.6 and a change of less than 10%. Thus fitting a linear regression is not robust, as it is dominated by fitting to this majority that do not change. Rather, it is clear that there is a distribution of change for each P(predict outcome), which we analyse in the main text.

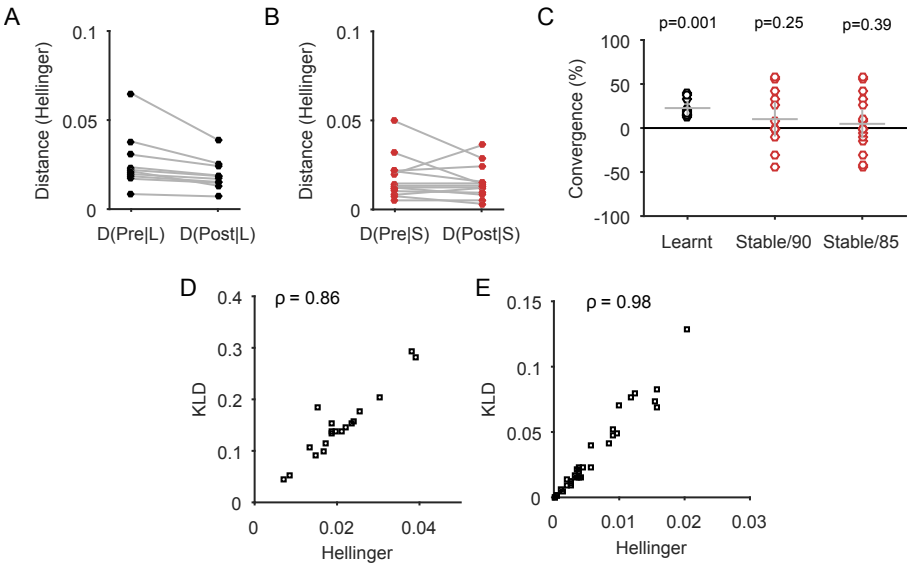

Figure S3: Robustness of learning and stable session predictions to distance measures. Here we re-test the main predictions from Figure 8 of the main text, using the non-parametric Hellinger distance instead of the Kullback-Liebler divergence. All results are qualitatively the same. (A) Distances between the distributions of pattern frequencies in sleep and training epochs; one dot per learning session. $D(X|Y)$: distance between pattern distributions in epochs X and Y: Pre: pre-training SWS; Post: post-training SWS; L: post-learning trials. (B) Distances between the distributions of pattern frequencies in sleep and training epochs in all stable sessions (here with at least 90% of trials with the same choice). S: training trials. (C) Scatter of convergence between post-training sleep and post-learning trials across all learning and stable sessions (circles). Convergence is $[D(Pre|X) - D(Post|X)] \,/\, \max\{D(Post|X), D(Pre|X)\}$, expressed as a percentage. Stable session results are plotted for both thresholds of 90% (13 sessions) and 85% (17 sessions). Grey lines give means and 95% confidence intervals. All P-values are from a 1-tailed Wilcoxon sign-rank test, with N=10 learning sessions. (D) The correlation between Hellinger distances and corresponding Kulback-Liebler divergences (KLD) between the sleep and training epochs, pooling the $D(Pre|L)$ and $D(Post|L)$ measurements (Hellinger distances from panel A; Kulback-Liebler divergences from Figure 8, panel B). Spearman's rank, $N = 20$; one outlier omitted from the plot for clarity but included in the correlation. Note the KLD is not converted into an information rate here, so that it can be directly compared to the Hellinger distance. (E) The correlation between Hellinger distances and corresponding Kulback-Liebler divergences between the pre- and post-training sleep epochs, pooled over learning and stable sessions (for the threshold of 85%). Spearman's rank, $N = 27$.

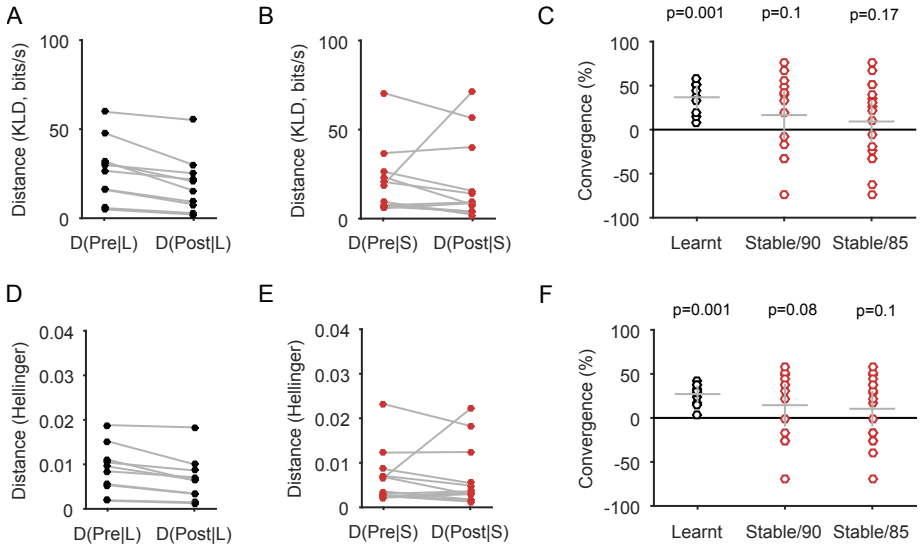

Figure S4: Robustness of learning and stable session predictions to number of neurons. Here we re-test the main predictions from Figure 8 of the main text, using $N = 15$ neurons per session to increase the reliability of the Kulback-Liebler divergence. All results are qualitatively the same. (A) Distances between the distributions of pattern frequencies in sleep and training epochs; one dot per learning session. $D(X|Y)$: distance between pattern distributions in epochs X and Y: Pre: pre-training SWS; Post: post-training SWS; L: post-learning trials. (B) Distances between the distributions of pattern frequencies in sleep and training epochs in all stable sessions (here with at least 90% of trials with the same choice). S: training trials. (C) Scatter of convergence between post-training sleep and post-learning trials across all learning and stable sessions (circles). Convergence is $D(Pre|X) - D(Post|X)/\max\{D(Pre|X), D(Post|X)\}$, expressed as a percentage. Stable session results are plotted for both the thresholds of 90% (13 sessions) and 85% (17 sessions). Grey lines give mean and 95% confidence intervals. All P-values are from a 1-tailed Wilcoxon signrank test, with N=10 learning sessions. (D) - (F) As panels A-C, using Hellinger distance.

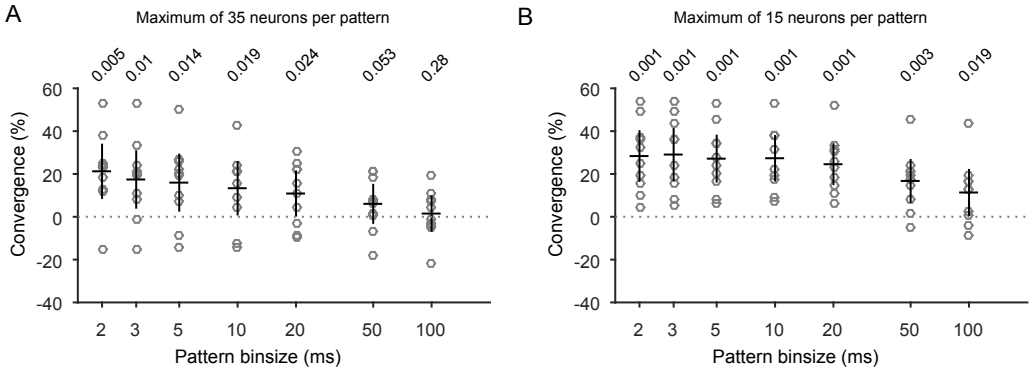

Figure S5: Robustness of learning-session convergence to the choice of bin size. The dependence of the learning session convergence of post-training sleep and post-learning distributions on the bin size used for constructing the activity patterns. Convergence computed using Kullback-Liebler divergence: $D(Pre|X) - D(Post|X)/\max\{D(Pre|X), D(Post|X)\}$, expressed as a percentage. (A) Bin size dependence of convergence for the full population, up to a maximum of 35 neurons per pattern. Circles are individual learning sessions ($N = 10$); black lines give means and 95% confidence intervals. All P-values are from a 1-tailed Wilcoxon signrank test. (B) As panel A, but using a maximum of 15 neurons per population.

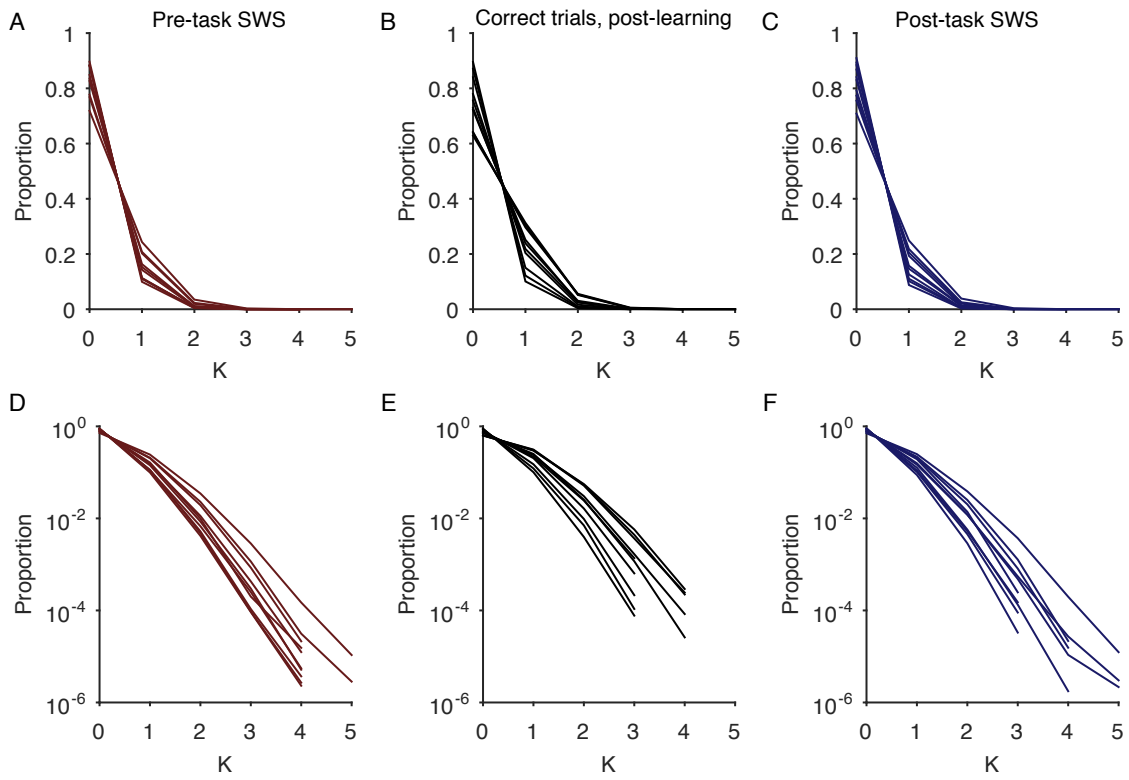

Figure S6: Distributions of synchronous spiking in all activity patterns at 2 ms bin size. (A)-(C) Distributions of the number of unique recorded activity patterns containing exactly K spikes, for pre-task SWS (A), correct task trials (B), and post-task SWS (C). Each line is the distribution for one session. (D)-(F) As A-C, plotted on a log-scale to visualise the tails of the distributions. Co-activation patterns ($K \geq 2$ synchronous spikes) form a small proportion of all patterns.

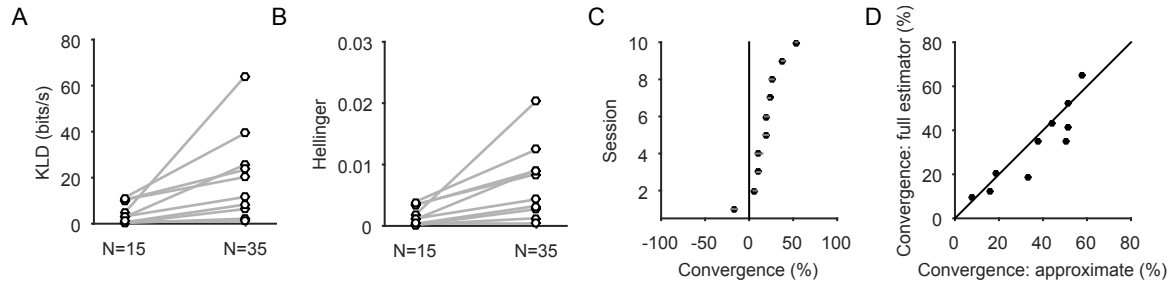

Figure S7: Checking issues in estimating distances between distributions. (A) The Kullback-Liebler distance between sleep epochs $D(Pre|Post)$ in the learning sessions, as a function of the maximum number of neurons per pattern (15 or 35). When using a maximum of 35, 8/10 sessions used their full recorded population. (B) As A, for the Hellinger distance. (C) Effects of variation in estimating the Kullback-Liebler distance. Here we plot the variation in the convergence score for each of the learning sessions over 100 repeated calculations of the Kullback-Liebler distance; symbols give mean distances; error bars plot two standard deviations - on this scale, they are approximately the width of the symbols. (D) Comparison of the convergence estimates for the learning sessions when using the full prior estimator of the unobserved portion of the activity pattern probability distribution (y-axis), and when using our approximation (x-axis). Here we use a maximum of 15 neurons per session, to allow tractable calculation of the full estimator.

S1 File: Behaviour and further predictions of the probabilistic reinforcement learning model

Abhinav Singh, Adrien Peyrache and Mark D. Humphries

Here we first demonstrate the general principle that posterior distributions stabilise over learning. We then discuss further predictions that arise from the Bayesian reinforcement learning model, which are testable in principle in future experiments. Finally, we further explore the behaviour of the Bayesian reinforcement learning model, to illustrate insights into the rats' behaviour on the Y-maze task.

Expected stabilisation of posterior distributions with learning

Estimating the probability distribution of some unknown value v_t (of, for example, a state or action) at time t, given all the rewards $(r_1, r_2, \ldots, r_t)$ up to time t, can be computed recursively using Bayes' theorem:

$$P(v_t|r_1, r_2, \ldots, r_t) \propto P(r_t|v_t)P(v_t|r_1, r_2, \ldots, r_{t-1}), \tag{1}$$

where the posterior distribution $P(v_t|r_1, r_2, \ldots, r_{t-1})$ for step $t-1$ becomes the prior distribution for step t. In general, given that r is stationary and given sufficient t, then the difference between the posterior and the prior $\delta = P(v_t|r_1, r_2, \ldots, r_t) - P(v_t|r_1, r_2, \ldots, r_{t-1})$ will become arbitrarily small. In other words, the posterior distribution will stabilise in any recursive Bayesian estimation.

We show now that this stabilisation of distributions is predicted to happen once our Bayesian reinforcement learning model has learnt the current rule. Once learnt, the agent will experience a long run of sustained rewards, with two consequences:

1. For the Beta distribution $P_x(v)$ modelling the correct strategy x this will mean a continuously increasing α_x, with β_x approximately fixed. As a result, we expect $\alpha_x \gg \beta_x$
2. The other Beta distributions, modelling the incorrect strategies, will be rarely updated (as they are only updated when selected). These distributions will thus be approximately stable.

So we can see the explicit stabilisation of $P_x(v)$ by calculating its change in mean and variance as a function of the number of rewards α. The mean of $P_x(v)$ is:

$$E(v) = \frac{\alpha}{\alpha + \beta}, \tag{2}$$

so the change in mean with increasing accumulated rewards is:

$$\frac{dE(v)}{d\alpha} = \frac{\beta}{(\alpha + \beta)^2}. \tag{3}$$

It is easy to see that as $\alpha \gg \beta$, so $dE(v) \to 0$ (Fig. P1A-B).

The variance of $P_x(v)$ is

$$Var(v) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}, \tag{4}$$

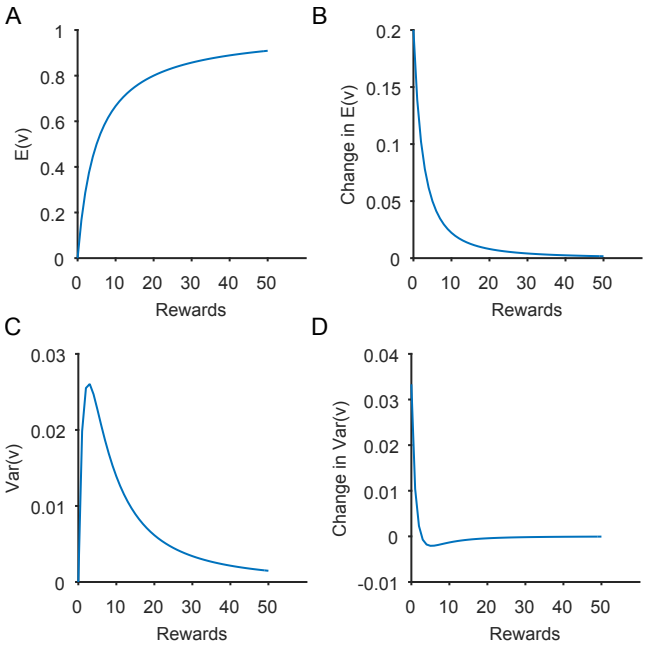

Figure P1: Stabilisation of probability distributions with reward. Here we illustrate the changes over cumulative reward in the scale and shape of the probability distribution $P_x(v)$ over the expected value of option x. All have $\beta = 5$. (A) Mean of $P(v)$ as a function of accumulated rewards. This shows how a linear increase in α - by the increase in rewards - gives an asymptotically stable estimate of the mean $E(v)$. (B) Change in mean of $P(v)$ as a function of accumulated rewards. (C) Variance of $P(v)$ as a function of accumulated rewards. (D) Change in variance of $P(v)$.

so the change in variance with increasing accumulated rewards is:

$$\frac{dVar(v)}{d\alpha} = \frac{\beta(-2\alpha^2 - \alpha(\beta+1) + \beta(\beta+1))}{(\alpha+\beta)^3(\alpha+\beta+1)^2} \tag{5}$$

Thus as $\alpha \gg \beta$, so $dVar(v) \approx (-2\alpha^2 - \alpha)/\alpha^5$; given the dominance of raising to the fifth power in the denominator, this also ensures $dVar(v) \to 0$ (Fig. P1C-D).

Further testable predictions of the Bayesian reinforcement learning model

Here we illustrate the two further predictions of the model, outlined in the Discussion. Testing these predictions requires future experiments.

The strongest prediction is that the probability distribution within a training epoch should become more similar between consecutive sessions as the task is learnt. Figure P2B shows how the distribution $P(v)$ for the correct strategy changes from session to session, becoming increasingly similar. If the joint activity of the population encodes this distribution, then the activity distributions should also become increasingly similar. An analytical challenge here is to estimate the distributions given data from all trials within a session.

Another, perhaps more experimentally amenable, prediction is that the distance between sleep distributions in stable sessions should be smaller than the corresponding distance in learning sessions. Figure P2C shows how the distribution $P(v)$ for the correct strategy changes between hypothetical pre- and post-training sleep, such that the difference between the two sleep epochs becomes negligible in later sessions of stable behaviour.

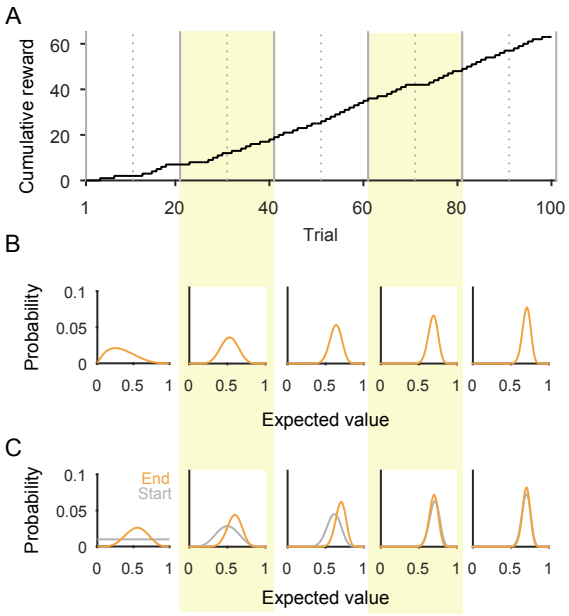

Figure P2: Further predictions of the Bayesian reinforcement learning model. Here we illustrate further predictions of a probabilistic internal model, amenable to testing in future experiments. (A) Cumulative reward curve for an agent learning the rule "go left". The curve is divided into arbitrary behavioural sessions (solid lines and colour shading); dashed lines indicate the mid-points of the behavioural sessions. (B) Mid-session distributions of $P(v)$ for the strategy "go left". Over learning, the within-session estimate of the $P(v)$ distribution is predicted to stabilise, and so becoming more similar between sessions. (C) Start (grey) and end (orange) session distributions of $P(v)$ for the strategy "go left", as a proxy for the distributions accessible in sleep before and after training. Over learning, the distributions in sleep should converge.

As noted in the Discussion, testing this prediction would require tracking an identical population of neurons across multiple days of behavioural training. In addition, both predictions would benefit from a different task design where the rule is not changed as soon as the animal has reached asymptotic performance. If instead the animal is left at asymptote for multiple behavioural sessions then reliable estimates of the activity distributions can be computed, and compared between sessions to test the hypothesised convergence to stability. One caveat with this approach is the increased risk of inducing habitual behaviour in the animals, making learning future rules more difficult.

Further behaviour of the Bayesian reinforcement learning model

Decoupling of strategy and action choice

Our Bayesian reinforcement learning model includes a noise term that selects the action (choosing the left or right arm) opposite to the chosen strategy with some small probability η. Without this term (i.e. $\eta = 0$), a noiseless agent learns rapidly, uniformly, and near-perfectly (Fig. P3A-B), in stark contrast to the observed rat behaviour.

The noise term thus simulates two things. First, a multi-armed bandit model cannot capture the complexity of learning the full task, so the missing complexity, and consequent "mistakes" from the perspective of the experimenter, are simulated by the noise term. Second, even if the hypothesised prefrontal cortex internal model was somehow learnt

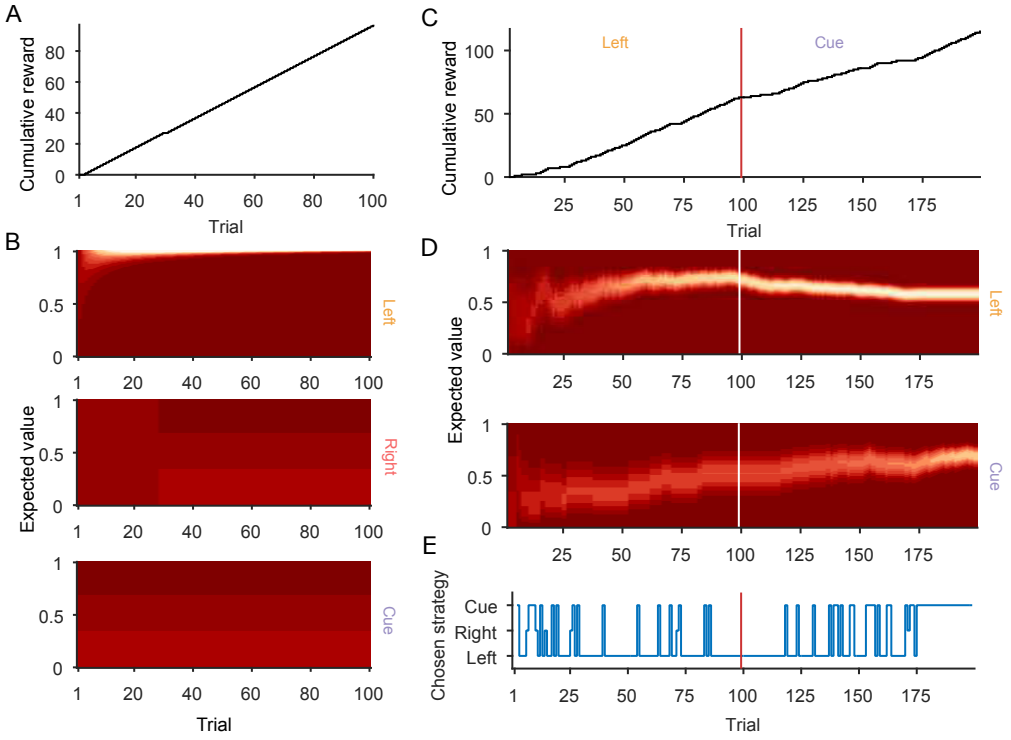

Figure P3: Further behaviour of the Bayesian reinforcement learning model. (A) Cumulative reward curve for a noiseless agent learning the rule "go left". (B) Corresponding changes to the probability distributions $P(v)$, over the value of each strategy. (C) Cumulative reward curve for an agent experiencing a rule change, from "go left" to "go to the cued arm". The curve plateaus after the rule change (red line), as the current strategy is incorrect. (D) Corresponding changes to the probability distributions $P(v)$ for the left and cued-arm strategies. Note that the new true value of the "go left" is not learnt, slowing the switch to the new strategy compared to initial learning. (E) Trial-by-trial strategy selection. Successful learning of the initially correct "go left" rule is evident by the dominance of selecting "go left" after about trial 27; similarly, successful learning of the new "cued arm" rule is evident by the emergence of selecting the "cued arm" strategy after about trial 175. Note though the persistence of selecting the wrong strategy ("go left") for many trials after the rule change.

perfectly, other neural systems also control behaviour [1–3]. Consequently, selection of an action need pay no heed to this particular system; from the perspective of any one action selection system, "noise" is inevitable.

There are interesting avenues here for further exploration. Our set-up of the Bayesian multi-armed bandit model treats the strategies as independent; consequently the chosen strategy only is updated, and the noise term means it can be updated by the wrong action, even if the strategy is correct. An alternative approach would be to treat the strategy selection as a sequential decision-making problem, with all strategies updated by whether there is evidence for them or not; in a Bayesian framework, this would require something like the multi-sequential probability ratio test [4–6]. Such an approach would not though change the basic prediction that the probability distributions stabilise over learning.

Learning a new rule

Our Bayesian reinforcement learning model also throws some light on what happens to the rat's behaviour after the rule has changed. Figure P3C shows that, as expected, the performance of the agent declines after the rule is changed: reward is obtained more

slowly until the new, correct strategy is acquired. The Bayesian model predicts that this acquisition of a new rule can take considerably longer than the acquisition of the first rule from a naive state. This is because the stable probability distribution originally acquired - $P(v)$ for "go left" in this example - can only change slowly with new evidence against it: the absence of a reward on each trial increments β for the Beta distribution, but as $\alpha \gg \beta$, so the mean changes very slowly (Eq. 2). Consequently it takes many more trials for the probability distribution over the expected value of the new, correct strategy to dominate (Fig. P3D).

This slow change means that the simulated agent shows stable selection of the previous, now wrong, strategy for many trials after the rule-change (Fig. P3D). We saw exactly this behaviour in the rat: 4 of the 8 rule change sessions showed the selection of the wrong, previous strategy for more than 85% of trials. The model explains this behaviour as the dominance of the prior strategy. That this dominance did not occur in all rule-change sessions suggests that the probability distributions for the pre-change rule were at different levels of stability. Thus it is not straightforward to make predictions for how the probability distributions should behave during the rule-change sessions.

References

[1] Swanson LW. Cerebral hemisphere regulation of motivated behavior. Brain Res. 2000;886:113–164.

[2] Humphries MD, Gurney K, Prescott TJ. Is there a brainstem substrate for action selection? Phil Trans R Soc B. 2007;362:1627–1639.

[3] Cisek P. Cortical mechanisms of action selection: the affordance competition hypothesis. Philos Trans R Soc Lond B Biol Sci. 2007;362:1585–1599.

[4] Bogacz R, Gurney K. The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Comput. 2007;19:442–477.

[5] Bogacz R, Larsen T. Integration of reinforcement learning and optimal decision-making theories of the Basal Ganglia. Neural Comput. 2011;23:817–851.

[6] Lepora NF, Gurney KN. The basal ganglia optimize decision making over general perceptual hypotheses. Neural Comput. 2012;24(11):2924–2945.

Neurons	Pre-training SWS	Post-learning trials	Post-training SWS	Rest
23	281001	57419	193500	315508
20	65007	49029	165519	350335
20	270012	34910	99488	282512
35	240992	20417	461972	92504
35	558510	43682	322499	131011
31	362007	26713	330485	206006
23	351996	50058	414982	205510
12	433009	29612	266493	204506
25	388006	50995	568512	105997
27	371013	64785	453993	90008

Table S1: Learning sessions: neurons and patterns. The Neurons column give the number of neurons used from each of the ten learning sessions to build the activity patterns; eight used all recorded neurons, two were capped at 35. The other columns give the total number of activity patterns in each epoch.

Neurons	Pre-training SWS	All trials	Post-training SWS
21	433009	42006	266493
19	377028	70435	468512
35	262999	76452	262511
35	341040	40062	250509
35	166511	70159	389510
35	104998	66319	16998
35	286491	66880	260521
35	109992	46539	209005
21	127997	71266	302997
19	346530	449624	448510
22	238523	30048	139999
17	521982	66071	330505
29	154498	144571	214992
12	107994	111723	204010
19	441977	108721	168996
21	90498	86011	112500
22	99508	97662	81003

Table S2: Stable sessions: neurons and patterns. The Neurons column give the number of neurons used from each of the 17 stable sessions (using the threshold of 85%) to build the activity patterns; nine used all recorded neurons, six were capped at 35. The other columns give the total number of activity patterns in each epoch.

Neurons	Pre-training SWS	Pre-rule change trials	Post-training SWS
19	273998	11163	170015
17	307511	49901	155486
23	261520	46241	203479
35	262999	18058	262511
35	166511	22933	389510
23	345519	43629	365490
29	154498	59996	214992
12	107994	26773	204010

Table S3: Rule-change sessions: neurons and patterns. The Neurons column give the number of neurons used from each of the eight rule-change sessions to build the activity patterns; six used all recorded neurons, two were capped at 35. The other columns give the total number of activity patterns in each epoch.