# An ensemble code in medial prefrontal cortex links prior events to outcomes during learning

Silvia Maggi[1], Adrien Peyrache[2], and Mark D. Humphries[1*]

1. Faculty of Biology, Medicine, and Health, University of Manchester, Manchester, UK.

2. Montreal Neurological Institute, McGill University, Montreal, Canada.

* Contact: mark.humphries@manchester.ac.uk

**Abstract**

The prefrontal cortex is implicated in learning the rules of an environment through trial and error. But it is unclear how such learning is related to the prefrontal cortex's role in short-term memory. Here we asked if the encoding of short-term memory in prefrontal cortex was used by rats learning decision rules in a Y-maze task. We found that a similar pattern of neural ensemble activity was selectively recalled after reinforcement for a correct decision. This reinforcement-selective recall only reliably occurred immediately before the abrupt behavioural transitions indicating successful learning of the current rule, and faded quickly thereafter. We could simultaneously decode multiple, retrospective task events from the ensemble activity, suggesting the recalled ensemble activity had multiplexed encoding of prior events. Our results suggest that successful trial-and-error learning is dependent on reinforcement tagging the relevant features of the environment to maintain in prefrontal cortex short-term memory.

## Introduction

Learning the statistical regularities of an environment requires trial and error. But how do we know what is relevant in the environment in order to learn its statistics? In other words: how do we know what to remember? It seems likely that medial prefrontal cortex plays a role here (Euston et al., 2012; Laubach et al., 2015): it is needed for trial and error learning of correct behavioural strategies (Ragozzino et al., 1999; Ragozzino, 2007; Rich and Shapiro, 2007), neuron and neural ensemble activity represents abstract and context-dependent information related to the current strategies (Jung et al., 1998; Rich and Shapiro, 2009; Hyman et al., 2012; Horst and Laubach, 2012), and changes to ensemble activity correlate with shifts in learnt behavioural strategies (Durstewitz et al., 2010; Karlsson et al., 2012; Powell and Redish, 2016). Moreover, medial prefrontal cortex receives a direct projection from the CA1 field of the hippocampus that may allow the integration of spatial information about the environment (Jones and Wilson, 2005; Hoover and Vertes, 2007; Burton et al., 2009; Benchenane et al., 2010; Spellman et al., 2015). But medial prefrontal cortex also plays a role in short-term and working memory for objects, sequences, and other task features (Miller, 2000; Miller and Cohen, 2001; Baeg et al., 2003; Averbeck et al., 2006; Averbeck and Lee, 2007; Fujisawa et al., 2008; Jun

et al., 2010; Machens et al., 2010; Spellman et al., 2015), upon which successful learning of statistical regularities may depend. It is unknown how relevant information about the statistics of the environment is tagged for memory in the medial prefrontal cortex.

An hypothesis we consider here is that reinforcement tags relevant choices and features to remember in order to learn the rules of the environment. This hypothesis predicts that the reliable appearance of reinforcement-driven short-term memory activity in medial prefrontal cortex should precede successful learning. While previous studies have shown that prefrontal cortex activity patterns shift between rule changes (Durstewitz et al., 2010; Karlsson et al., 2012), including immediately before a shift in behaviour (Powell and Redish, 2016), none have looked at learning-driven changes to prefrontal cortex activity in the naive animal, nor what the activity encodes about the task. We thus sought to test the hypothesis that ensemble activity in medial prefrontal cortex represents a short-term memory of task features and choices that are potentially necessary for learning the current rule.

To test this hypothesis, we analysed neural and behavioural data from rats learning new rules on a Y-maze. We took advantage of a task design in which there was a self-paced return to the start position of the maze immediately after the delivery or absence of reinforcement, yet no explicit working memory component to any of the rules. Consequently we could examine ensemble activity in medial prefrontal cortex during this self-paced return and ask whether or not a short-term memory encoding of reinforcement-tagged task features existed in the absence of overt working memory demands.

Here we report that medial prefrontal cortex contains an ensemble code that links prior events to reinforcement. We show that a similar pattern of ensemble activity was specifically recalled after reinforcement and not after errors. This recall only reliably occurred in sessions with abrupt shifts in behavioural performance that indicated successful learning of a rule, and not during external shifts in reinforcement contingency, or in other task sessions. The recalled pattern appeared shortly before the abrupt shift in behaviour, and faded shortly thereafter, consistent with a causal role in learning. From the activity of the recalled ensemble, we could simultaneously decode retrospective task parameters and choices in a position-dependent manner, even from the naive animal. Together, these results show that learning was preceded by reinforcement-triggered ensemble activity that retrospectively and multiply encoded task parameters. They provide a link between the roles of medial prefrontal cortex in working memory and in rule learning, and suggest that reinforcement tags prefrontal cortex-based representations of choices and environment features that are relevant to trial and error learning of statistical regularities in the world.

# Results

In order to address whether and how medial prefrontal cortex neural activity encodes short-term memory during rule learning, we used medial prefrontal cortex population recording data previously obtained from a maze-based rule-learning task (Peyrache et al., 2009). Four rats learnt rules for the direction of the rewarded arm in a Y-shaped maze, comprising a departure arm and two goal arms with light cues placed next to the reward ports (Figure 1a). Each session was a single day with approximately 30 minutes of training, and 30 minutes of pre- and post-training sleep. During training, the rat initiated each trial from the start of the departure arm; the trial ended when the rat arrived at the end of its chosen goal arm. During the following inter-trial interval the rat consumed any delivered reward, if the correct arm was chosen according to the current rule, and then made a self-paced return to the start position after consuming the reward, taking on average 70

s ($67.8 \pm 5.4$ s, mean $\pm$ SEM) to complete the interval. Tetrode recordings from medial prefrontal cortex were obtained from the very first session in which each rat was exposed to the maze (Figure 1b). Thus, the combination of a self-paced post-decision period – without experimenter interference – and neural activity recordings from a naive state allowed us to test for medial prefrontal cortex ensemble activity correlating with short-term memory during rule learning.

After achieving stable performance of the current rule, indicated by 10 contiguous correct choices, the rule was changed, unsignalled, in sequence: go right; go to the cued arm; go left; go to the uncued arm. Notably, none explicitly required a working memory component (such as an alternation rule). In the original study (Peyrache et al., 2009), the session in which initial learning of each rule occurred was identified posthoc as the first with three consecutive correct trials followed by 80% performance until the end of the session; the first of the initial three correct trials was identified as the learning trial. Ten sessions met these criteria, and are dubbed here the "learning" sessions. We first confirmed that these ten learning sessions showed an abrupt transition in behavioural performance (Figure 1c), indicating the step-like change in behaviour commonly seen in successful learning of contingencies (Gallistel et al., 2004; Aziz-Zadeh et al., 2009; Durstewitz et al., 2010; Powell and Redish, 2016). In total, we examined 50 sessions, comprising 10 learning sessions, 8 rule change sessions, and 32 other training sessions (labelled "others" throughout).

## Reinforcement-driven recall of ensemble activity during learning

We sought to identify signatures of short-term memory encoding by examining ensemble activity in the inter-trial intervals within each session. One signature of similar memory encoding between inter-trial intervals would be a repeated pattern of co-active neurons across the intervals, indicating a similar encoding. To allow comparisons between intervals, we thus first identified the core population of neurons in each session by selecting those that were active in every inter-trial interval. The proportion of recorded neurons retained in the core population was on average $74 \pm 2$ % (SEM) across sessions (Supplementary Fig. S1, panel a). No clear difference in the size of this core population were observed between learning and any other session type (Supplementary Fig. S1, panel a), suggesting that any potential short-term memory encoding specific to learning was not then simply a change in the proportion of active neurons.

We then asked if this core population contained a repeated pattern of co-active neurons between inter-trial intervals. We characterised the co-activity for each inter-trial interval by computing the pairwise similarity between the Gaussian-convolved spike-trains of neurons in the core population (we use a Gaussian width of $\sigma = 100$ ms here, as in the example of Figure 1b; the effects of varying $\sigma$ are detailed below). The pattern of co-activity in interval $t$ was thus described by a matrix $S_t$ of pairwise similarities between neuron activity. To then quantify if a similar pattern of co-activity occurred between inter-trial intervals $t$ and $u$, we computed the similarity between their co-activation matrices $S_t$ and $S_u$. Repeating this for all pairs of inter-trial intervals in a session gave us a matrix $R$ of pairwise similarities between intervals, which showed where similar patterns of ensemble activity were recalled on different inter-trial intervals (Figure 2a) – we thus dubbed it the recall matrix.

We found that patterns of neuron co-activity were more similar between intervals after correct trials than after error trials (Figure 2a,b). We observed this preferential recall of ensemble activity following reinforcement in the majority of sessions (47/50 sessions; 37/50 had $p < 0.05$ for a Kolmogorov-Smirnov test between the distributions of recall values after correct and after error trials – example distributions in Fig. 2a). This result

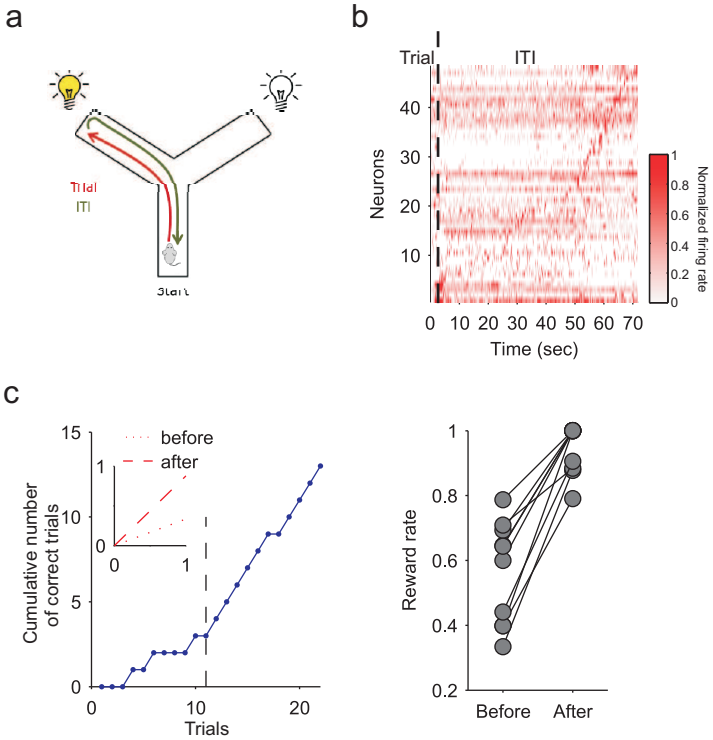



Figure 1: Task and learning sessions.
(**a**) Schematic representation of the Y maze. The trial starts with the animal at the start of the departure arm, and ends when it reaches the end of the chosen arm. The inter-trial interval (ITI) is a self-paced return back to the start position.
(**b**) Example medial prefrontal cortex population activity during a trip out and back to the start position. The heatmap shows the spike-trains for all recorded neurons, convolved with a Gaussian of width $\sigma = 100$ ms. The dashed line separates the trial and inter-trial interval periods. The firing rate of each neuron is a proportion of its peak rate, and neurons are sorted by the time of their peak firing rate.
(**c**) Learning sessions contain abrupt transitions in performance. Left panel: Learning curve for one example learning session. The cumulative number of correct trials shows a steep increase after the learning trial (black dashed line), indicating the rat had learnt the correct rule. Inset: fitted linear regressions for the cumulative reward before (dotted) and after (dashed) the learning trial, quantifying the large increase in the rate of reward accumulation after the learning trial. Right panel: the rate of reward accumulation before and after the learning trial for every learning session (one pair of symbols per learning session; one session's pair of symbols are obscured). The rate is given by the slopes of the fitted regression lines.

would suggest that reward triggered a specific pattern of ensemble co-activity during the inter-trial interval. However, we were mindful that the inter-trial intervals following a correct trial were generally much longer than those following error trials (correct inter-trial intervals: $79.1 \pm 6.4$ s; error inter-trial intervals: $48.4 \pm 3.7$ s), because the animal lingered at the reward location (Supplementary Fig. S1). This difference in duration could systematically bias estimates of co-activity, simply because many more spikes would be emitted during intervals after correct choices than after errors. Thus, greater similarity between ensemble activity patterns for correct intervals could simply be due to more reliable estimates of the similarity between each pair of neurons.

To control for this, we used shuffled spike-trains to compute the expected pairwise similarities between neurons due to just the duration of each interval; from these shuffled-data similarity matrices per interval we then computed the expected recall matrix (Supple-

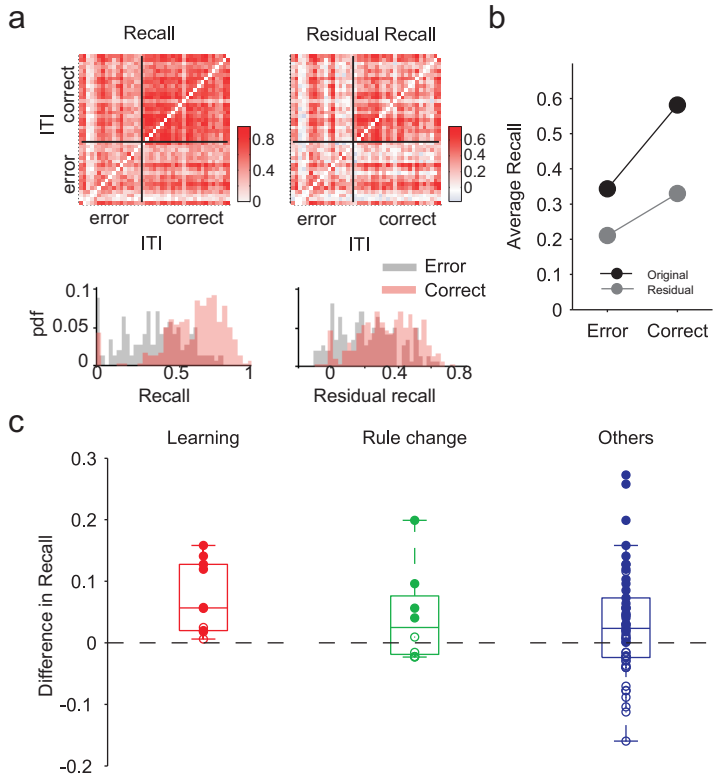



Figure 2: Outcome-selective recall of an ensemble activity pattern is learning-related.
(**a**) Left: the Recall matrix $R$ for one example session. The recall matrix is ordered by the outcome of the trial preceding the inter-trial interval (ITI). Each entry $R_{ij}$ is the recall value: the similarity between the core population's similarity matrix in intervals $i$ and $j$, a measure of how closely the same ensemble activity pattern was recalled in that pair of intervals. Below we plot the probability density functions for the distribution of recall values, separately for the post-error (red) pairs of intervals (bottom-left block diagonal in the recall matrix) and for the post-correct (black) pairs of intervals (top right block diagonal in the recall matrix). Right: the residual recall matrix for the same session, after correction for the effects of interval duration.
(**b**) The average recall values for post-error and post-correct intervals of the two matrices in panel **a**. The distribution of recall in the post-correct intervals was higher than in the post-error intervals (K-S test; recall: $P < 0.005$; residual recall: $P < 0.005$ ; N(correct) $= 24 \times 24 = 576$; N(error) $= 17 \times 17 = 238$.)
(**c**) The difference in average recall between the post-correct and post-error intervals, sorted by session type. Each dot is one session. Filled circles indicate a positive difference at $p < 0.05$ between the distributions of recall values in the post-error and post-correct intervals (Kolmogorov-Smirnov test).

mentary Fig. S2). Consequently, by subtracting this expected recall matrix from the data-derived recall matrix, we obtained a "residual" recall matrix describing just the similarity between ensemble activity patterns above those driven by common duration (Figure 2a). We used this residual recall matrix for all further analyses. With this correction, we still found that patterns of ensemble activity were more similar after correct trials than after error trials in the majority of sessions (34/50 sessions; 26/50 had $p < 0.05$ for a Kolmogorov-Smirnov test between the distributions of residual recall values after correct and after error trials). This result suggests that reward triggered a specific pattern of ensemble co-activity during the inter-trial interval.

We then examined how this reinforcement-driven recall of an ensemble activity pattern corresponded to the rats' behaviour (Figure 2c). We found that only learning sessions had

a systematically stronger recall of the same ensemble activity pattern after reinforcement (mean difference in recall: 0.072). Sessions in which the rule changed did not show a systematic recall after reinforcement (mean difference in recall: 0.042), ruling out external changes to contingency as the driver of the recall effect. Similarly, there was no systematic reinforcement-driven recall in the other sessions (mean difference in recall: 0.03), ruling out a general reinforcement-driven effect. When we further grouped these other sessions into those with evidence of incremental learning and those without, we still did not observe a systematic reinforcement-driven recall effect in either group (Supplementary Fig. S2). Finally, we tested the likelihood of obtaining ten systematically positive recall sessions by chance if we assumed recall was randomly distributed across the sessions. We repeatedly chose ten sessions at random from the 50; repeated 10,000 times, we found a probability of less than 0.01 of randomly obtaining 10 sessions which each had positive recall. Together, these data show that a similar pattern of ensemble activity was only reliably recalled following reinforcement during the self-driven step-change in behaviour indicative of learning a rule.

The reinforcement-driven recall of an ensemble activity pattern could potentially be specifically triggered only during reward consumption, and so not sustain across the self-paced return to the start position of the maze. To check this, we divided the neural activity into the activity occurring at the reward location and the activity during the self-paced return, and then separately computed the recall analysis on both. We found no systematic recall of a reinforcement-driven activity pattern at the reward location, and weak recall during the rest of the self-paced return (Supplementary Fig. S3), suggesting the reinforcement-driven recall of an ensemble activity pattern occurred throughout the inter-trial interval.

We asked how the recall of a pattern of ensemble activity was dependent on the temporal precision at which the similarity between neurons was computed. Here, this precision was determined by the width of the Gaussian convolved with the spike-trains. We found that the reinforcement-driven recall of an ensemble in learning sessions was consistent across a wide range of Gaussian widths from 20 ms up to around 140 ms (Supplementary Fig. S4). Moreover, across the same range of Gaussian widths, we also consistently found that the recall effect for the learning sessions was greater than for rule-change or other sessions (Supplementary Fig. S4). The reliable recall down to 20 ms, and the absence of a systematic recall effect for Gaussian widths above 200 ms, suggests the co-activity pattern was formed by relatively precise timing between spikes from different neurons, rather than just rate co-variation.

## Recall of ensemble activity patterns is specific to retrospective reinforcement

These results pointed to the hypothesis that, during successful rule learning, the reliable recall of a pattern of ensemble activity is triggered by prior reinforcement. To test this hypothesis, we asked whether the recalled ensemble activity pattern was specifically triggered by reinforcement, and whether it was specific to retrospective rather than prospective reinforcement.

To test if the recall was specifically triggered by reinforcement, we reorganised the residual recall matrix of each session by either the direction of the chosen goal arm (left/right) or by the cue position (left/right) on the previous trial. We found there was no systematic recall of ensemble activity patterns evoked by one direction over the other for either the chosen arm or the cue position (Figure 3a,b). The systematic recall effect during learning thus appeared to be specific to reinforcement.

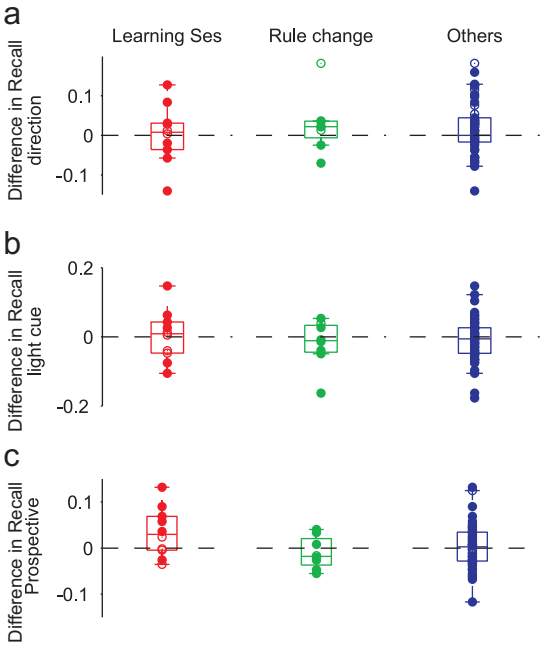



Figure 3: Recalled ensemble activity patterns are outcome-specific and encode retrospective outcome not future choice.
(**a**) The difference in average recall between intervals after choosing the left arm or after choosing the right arm, sorted by session type. Filled circles here and in other panels indicate a significant difference between the distributions of recall values in the two sets of intervals (Kolmogorov-Smirnov test, $p < 0.05$).
(**b**) As for panel **a**, but comparing intervals after the light cue appeared at the end of the left or the right arm.
(**c**) The difference in average recall between intervals before error or before correct trials, testing for the prospective encoding of upcoming choice.

Modulation of medial prefrontal cortex activity by expected outcome or anticipation of reinforcement has been repeatedly observed (Daw et al., 2006; Fellows, 2007; Sul et al., 2010; Kaplan et al., 2017), suggesting the recalled ensemble pattern could instead be a representation of the expected outcome on the next trial. To test if the recall effect was specific to retrospective reinforcement, we reordered the residual recall matrices according to the reinforcement received in the trial after the inter-trial interval. We found no systematic recall of an ensemble activity pattern in intervals preceding correct trials in any session type (Figure 3c). In particular, for the learning sessions the systematic recall we observed for retrospective outcomes was not observed for prospective outcomes (compare Figure 2c), and the magnitude of recall was larger for retrospective than prospective outcomes across all tested temporal precisions of similarity between spike-trains (Supplementary Fig. S4, panel d).

We were surprised that we could observe such a consistent difference between the retrospective and prospective recall in the learning sessions. By their nature, the learning sessions tend to be split into a sequence of error trials followed by a sequence of correct trials (cf Figure 1c), so each trial outcome is frequently preceded and followed by the same type of outcome. Consequently, we create similar groups of "correct" and "error" intervals whether we choose to split intervals into groups by their following correct trials or by their preceding correct trials. Nonetheless, the systematically stronger retrospective recall across a wide range of timescales, despite the few error trials interspersed with correct trials, suggests that the recall of ensemble activity is dependent on prior, not future,

reinforcement. (And as we show below, this conclusion is consistent with the complete absence of prospective coding of task elements by the ensemble's activity). Together, these results support the hypothesis that a specific pattern of ensemble activity triggered by just-received reinforcement appeared during successful rule learning.

## Appearance of the recalled ensemble activity anticipates the behavioural transition

This leaves opens the question of whether the appearance of this recalled ensemble pattern is a pre-condition of successful learning, or a read-out of already learnt information. If a pre-condition, then the recalled ensemble pattern should have appeared before the transition in behaviour indicating rule acquisition.

We thus sought to identify when the recalled ensemble activity pattern first appeared in each learning session. To look for this onset of the co-activity pattern, we put the recall matrix of each learning session in trial order (Figure 4a). For each inter-trial interval in turn, we compared the strength of recall in the inter-trial intervals before and after that interval (Figure 4b). We used the inter-trial interval corresponding to the largest difference in recall to identify the onset trial – the trial after which the ensemble activity pattern first appeared – as this indicated a step-increase in the similarity of activity patterns between inter-trial intervals.

We found that the recalled ensemble pattern appeared before or approximately simultaneously with the behavioural transition in all sessions (Figure 4c,d). This was true whether we used the original behavioural criterion from Peyrache et al. (2009), or our more stringent definition of "abrupt" change in the cumulative reward curve (the trial corresponding to the greatest change in slope of the reward accumulation curve; see Methods). The timing of the appearance of the recalled ensemble pattern was thus consistent with it being necessary for successful rule learning.

As the change to the ensemble activity was often abrupt and so close to the behavioural change, this raised the question of what change to the underlying neural circuit drove this change in activity. One possibility would be a physical alteration of connectivity, forming a true "structural" cell assembly (Harris, 2005). Alternatively, it could be a temporary effect, as might arise from a sustained change in neuromodulation (Durstewitz and Seamans, 2002; Benchenane et al., 2011), forming a transient "functional" cell assembly.

To decide between these alternatives, we tested for the presence of a long-lasting physical change by assessing the longevity of the recalled ensemble activity pattern. Specifically, we tested whether the recall of the ensemble was sustained until the end of the learning session by performing the onset analysis in reverse (Figure 4e): for each inter-trial interval in turn, we checked whether the recall after that interval was significantly smaller than before it (Kolmogorov-Smirnov test; see Methods). We indeed found a statistically robust fall in the recall of the ensemble activity pattern in every learning session. A strict ordering was always present: the decay of the recalled ensemble was after the identified onset of recall, but before the end of the session (Figure 4e), even though we did not constrain our analysis to this ordering. For the original set of identified learning trials, the decay trial was always after the learning trial (Figure 4e). (If we used our alternative learning-trial definition – the trial with the greatest change in reward accumulation – then 7 of the 10 sessions had decay after the learning trial, with 3 sessions showing decay before it). This analysis indicates the recalled ensemble activity pattern formed transiently during learning, and decayed quickly after learning was established.

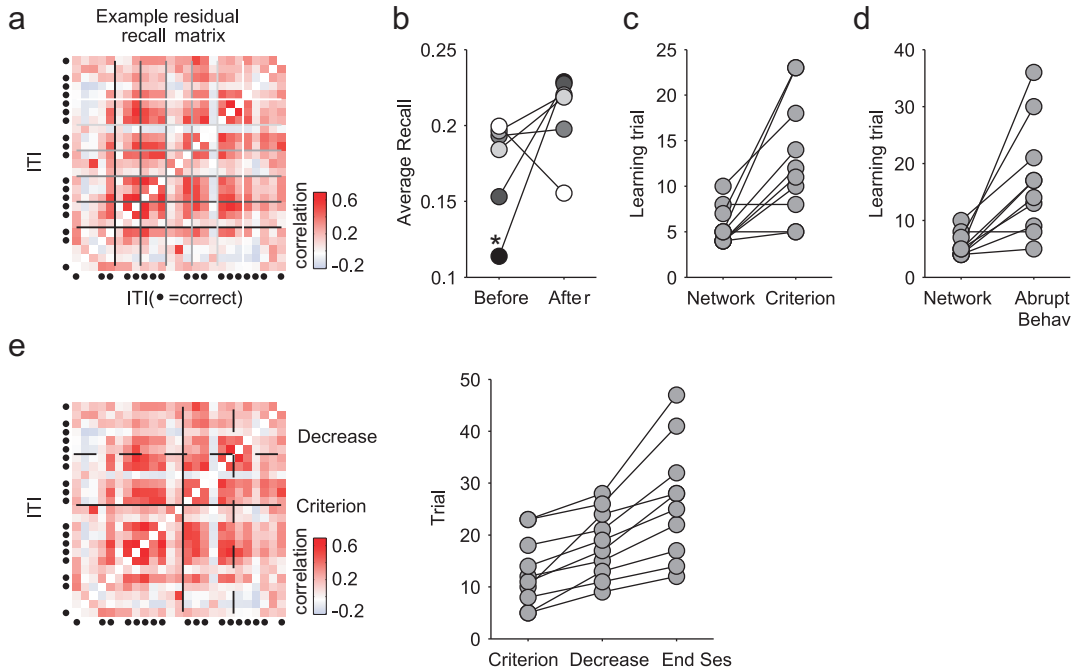



Figure 4: The recalled ensemble activity pattern anticipates behavioural learning.
(**a**) A residual recall matrix in its temporal order for one example learning session. Columns are ordered from left to right as the first to last inter-trial interval (rows ordered bottom to top). For each inter-trial interval, the distributions of the recall values before and after the selected inter-trial interval were compared (Kolmogorov-Smirnov statistic: see Methods). Each greyscale line corresponds to the selected subset of the dividing inter-trial intervals plotted in panel **b**.
(**b**) For each greyscale line in panel **a**, the corresponding average recall value before and after the dividing inter-trial interval. The asterisk indicates the inter-trial interval with the largest increase in recall after it, signalling the abrupt appearance of the recalled ensemble pattern.
(**c**) Comparison of the learning trial identified by the original behavioural criterion with the identified onset trial for the recalled ensemble activity pattern ('Network').
(**d**) As panel **c**, but with the behavioural learning trial identified as the trial with the steepest change in the cumulative reward (see Methods).
(**e**) Testing for decay of the ensemble activity pattern. Left panel: example residual recall matrix in trial order for one learning session. The black solid line is the learning trial, while the dashed line is the identified offset of the recalled ensemble activity pattern. Right panel: For each learning session the learning trial (original criterion) is compared to the identified offset of the ensemble recall, and to the last trial of the session.

## Medial prefrontal cortex ensembles had mixed, position-dependent, and retrospective encoding of task information

What did the recalled activity pattern encode? Its transient appearance, immediately before behavioural change but fading before the end of a session, suggests a temporary representation, akin to short-term memory. That the recalled pattern was triggered only by prior reinforcement suggests the hypothesis that the recalled ensemble was a working memory encoding of task features that were potentially relevant for learning. If it was a working memory for task features, then we should be able to decode prior task information from ensemble activity.

To address this, we assessed our ability to decode prior outcome, choice of goal arm, and light cue position from the core population's activity. As prefrontal cortex activity encoding often shows broad position dependence (Baeg et al., 2003; Hok et al., 2005;

Spellman et al., 2015), we divided the linearised maze into five equally-spaced sections (Figure 5a), and represented the core population's activity in each as the vector of its neurons' firing rates in that section. We used these firing rate vectors as inputs to a cross-validated linear decoder (Figure 5b), and compared their predictive performance to shuffled data (Methods).

We could decode prior outcome, choice of goal arm direction, and cue position well above chance performance, and often in multiple contiguous maze positions. We plot the absolute decoding performance for the "other" sessions in Figure 5c to illustrate that decoding at some maze positions was near-perfect, with some sessions decoded at 100% accuracy. The learning and rule-change sessions also had maze positions with near-perfect decoding across all sessions (Supplementary Fig. S5). Ensemble activity in medial prefrontal cortex thus robustly encoded multiple task events from the previous trial.

We then compared decoding performance between session and rule types. Chance decoding performance differed between task features (as the randomised light-cue was counter-balanced across trials, but each rat's choice and hence outcomes were not), and between session types and rule types (as rat performance differed between them). Thus we normalised each decoder's performance to its own control, and compared this relative decoding accuracy across sessions and rules (Figure 5d).

These comparisons revealed we could decode the prior choice of direction (left or right) in all types of session and regardless of whether the rule was direction- or cue-based (Figure 5d). Decoding of direction choice was robustly above chance while the rats moved from the end of the goal arm back to the maze's choice point (highlighted yellow); on cued-rule sessions, this decoding extended almost all the way back to the start position of the departure arm. Accurate decoding of direction choice could be observed from the very first session of each rat, and consistently across sessions before the first rule change (Figure 5e). These results indicated that medial prefrontal cortex always maintained a memory of prior choice, and did not need to learn to encode this task feature.

Similarly, we could decode the prior outcome (correct or error) in all types of session and regardless of whether the rule was direction- or cue-based (Figure 5d). Decoding of outcome was notably stronger at the end of the goal arm, where the reward was delivered, but could also be decoded above chance while the rats traversed the maze back to the start position (highlighted yellow). Nonetheless, decoding of outcome was again present from the very first session (Figure 5e). These results indicated that medial prefrontal cortex always encoded the trial's outcome, and did not need to learn to encode this task feature.

By contrast to the encoding of prior direction and outcome, we could only reliably decode the prior cue position in two specific locations (Figure 5d). The prior cue position was consistently encoded at the end of the goal arm for both cue and direction rules, likely corresponding to whether or not the light was on at the rat's position. But the only sustained encoding of prior cue position while the rat traversed the maze was during learning sessions for cue-based rules (yellow highlighted position and red open circles in Figure 5d). There was no sustained encoding of the cue during learning sessions of direction rules (red filled circles in Figure 5d). And this sustained encoding of the cue did not appear in the first session, nor in any session before the first change to the cue-based rule (Figure 5e). Consequently, these results suggest that only in learning sessions did the core population encode the memory of the prior cue position, and only when relevant to the learnt rule.

Strikingly, we found that decoding of prospective choice or outcome on the next trial was at chance levels throughout the inter-trial interval (Figure 6). These results were consistent with our finding that the ensemble activity pattern preceding correct trials

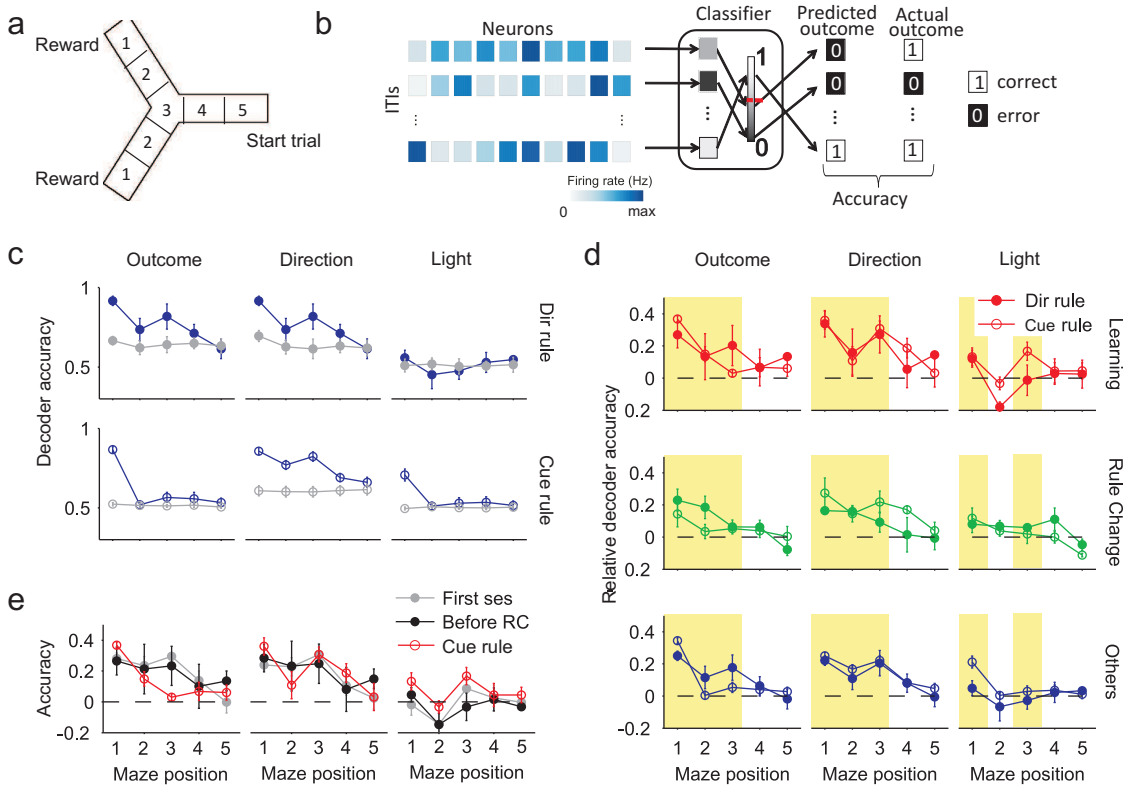



Figure 5: Position-dependent encoding of recent task relevant information.
(**a**) Graphical representation of the five equally size section of the maze. Position 1 is the goal arm end, with the reward delivery port. Position 3 is the choice point during the trial.
(**b**) Schematic of decoding task events from the core population's firing rate vector. For each inter-trial interval, and for each of five positions in the maze, the population's firing rate vector is given as input to a linear decoder. The decoder attempts to classify the population vector as belonging to one of two possible labels (error or correct, for prior outcome; left or right, for prior direction choice and prior cue position), given a threshold (red dashed line) on the decoder's output. The accuracy of the decoder is given as the proportion of correctly predicted labels. For robustness, we use cross-validation to create the predictive model for each inter-trial interval; and we compare predictive performance to that of a randomised control: cross-validated classifiers applied to data with permuted labels.
(**c**) Example decoder accuracy as a function of maze position for the "other" sessions. For these sessions, we show here the absolute decoder performance for each of the three classified features (prior outcome, prior direction choice, and prior light positions), separated by the rule type (direction or light cue-based rules). Each data point is the mean ± SEM accuracy at that maze position. Chance levels of performance are plotted in grey, and were defined separately for each session type and each rule type (see Methods). We plot here and in panel **d** the results for a logistic regression decoder; other decoders are plotted in the Supplementary Fig. S6.
(**d**) Relative decoder accuracy over all session types and rule types. Each data point is the mean ± SEM accuracy in excess of chance (0; dashed line) over the indicated combination of session (learning, rule-change, other) and decoded feature. Each panel separately plots the decoding for cued-rules (open symbols) and direction-rules (filled symbols). Highlighted groups of positions indicate consistent departures from chance performance in at least one session type. We replot these results grouped by rule-type in the Supplementary Fig. S5.
(**e**) Decoding of task features at the start of learning. Similar to panel **d**, each data point is the mean ± SEM accuracy in excess of chance. In each panel, the grey line gives the accuracy over the first session of each animal, the black line the accuracy over all the sessions before the first rule change, and the red line gives the accuracy over the first light cue session for each animal.

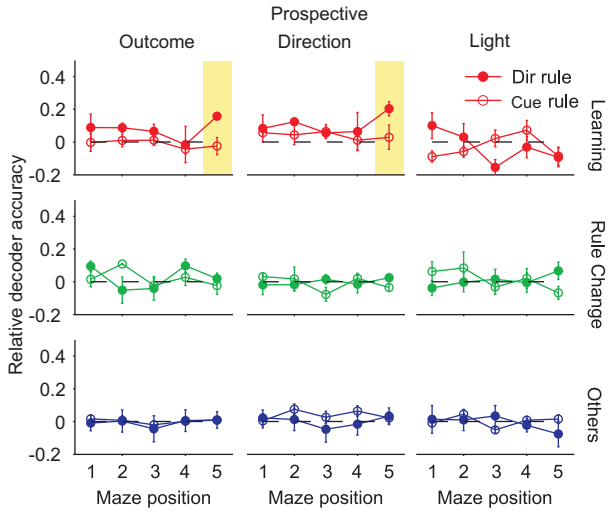



Figure 6: Prospective encoding of outcome and choice for predictable rule. We plot here the relative decoder accuracy over all session types and rule types for features on the immediately following trial. Compare to Figure 5d. Each data point is the mean ± SEM accuracy in excess of chance (dashed line) over the indicated combination of session (learning, rule-change, other) and decoded feature. Each panel separately plots the decoding for cued-rules (open symbols) and direction-rules (filled symbols). As a sanity check that our cross-validation of the decoder and shuffled controls were working, we also decoded the prospective light position: as this was randomised, the ensemble activity could not predict its position and so should only have been decoded at chance levels – which it was. We replot these results grouped by rule-type in Supplementary Fig. S5.

was not systematically recalled (Figure 3c). They also show that the decoding of prior task features from the core populations' activity was non-trivial. The only above-chance decoding of prospective information was observed for direction-based rules, where we found that decoding of future choice and outcome was above chance level only for learning sessions and only at position 5, where the animal makes a U-turn before starting the new trial. This suggests that medial prefrontal cortex activity around the start of the trial could also be related to the upcoming decision when the task rule is successfully learnt; future work will explore this idea.

Collectively, the decoder analysis showed we could decode multiple task features from the immediate past from population activity in the medial prefrontal cortex – in some cases, perfectly (Figure 5c, and Supplementary Fig. S5) – but not the immediate future. Moreover, our ability to decode these prior task features was consistent across a range of tested decoders, as was the sustained encoding of prior cue position only during the learning of cue-based rules (Supplementary Fig. S6). Thus, we suggest that the specific pattern of recalled ensemble activity triggered by reinforcement is the repeated synchronisation of the multiplexed encoding of prior choice, outcome, and cue position relevant to learning the current rule.

# Discussion

We sought to understand how short-term memory in medial prefrontal cortex may support the trial-and-error learning of rules from a naive state. To do so, we analysed population activity in medial prefrontal cortex from rats learning rules on a maze, and asked if the activity during the inter-trial interval carried signatures of short-term memory for rule-relevant features of the task. We found that a specific pattern of ensemble activity was

recalled only after reinforced trials, and only reliably during sessions in which the rats learnt the current rule for the first time. This dependence on prior outcome, and the transient appearance of the ensemble activity pattern, was consistent with a short-term memory encoding, rather than a persistent change to the underlying neural circuit.

We could robustly decode prior outcome and direction choice from ensemble activity across all sessions, but found that encoding of the prior cue position was specific to learning sessions for the cue-based rules. This suggests that the recalled ensemble is a repeated synchronisation of multiple encodings across the neural population, with rule-appropriate suppression or enhancement of cue encoding. We thus propose that reinforcement tags features to sustain in medial prefrontal cortex working memory, and does this by reliably triggering a specific pattern of ensemble activity that jointly encodes relevant task features.

## Ensemble recall precedes behavioural learning

We only reliably observed the recall of the same pattern of co-activity between neurons following correct trials during learning. This pattern of co-activity reliably appeared before or simultaneously with the step-like change in behavioural performance indicating rule acquisition. The timing thus suggests a causal link between the appearance of the recalled ensemble activity pattern and successful learning of the correct strategy for the current rule.

Our results support recent studies of prefrontal cortex population activity that reported how the pattern of population activity in rodent prefrontal cortex changes with or immediately prior to an internally-driven shift in behavioural strategy (Durstewitz et al., 2010; Powell and Redish, 2016). We extend these prior results in three ways. First, prior work has studied scenarios where animals well-trained on one contingency experienced a change in that contingency. Here, we have shown that such abrupt shifts in population activity patterns can occur from the naive state. Consequently, they encode initial acquisition as well as uncertainty (Karlsson et al., 2012). Second we have shown that such an abrupt shift in population activity happens for a putative working memory representation. Third, we have shown that this shift is selectively triggered by prior reinforcement. Nonetheless, our results add to the growing evidence that an abrupt shift in prefrontal cortex population activity is a necessary condition for the successful expression of a new behavioural strategy.

The observed shifts in prefrontal cortex population activity while both acquiring new strategies and switching between well-learnt strategies are potentially driven by distinct processes. In our task, we observed that the shift in population activity correlated with correct performance. Recent studies of strategy-switching (Narayanan and Laubach, 2008; Karlsson et al., 2012; Laubach et al., 2015; Powell and Redish, 2016) have reported that errors are the main trigger for the shift in population activity. We propose these findings are respectively consistent with separate roles of the prefrontal cortex in initial acquisition driven by unexpected reward, and performance monitoring driven by unexpected errors (Laubach et al., 2015).

This hypothesis fits well with theories of reinforcement learning based on the firing of midbrain dopamine neurons (Schultz et al., 1997; Bayer and Glimcher, 2005; Bromberg-Martin et al., 2010; Eshel et al., 2016). In the naive state, unexpected reward elicits phasic dopamine activity; in the well-learnt state, reward is expected so elicits no dopamine response, but unexpected errors elicit phasic dips in dopamine activity. These bursts and dips can act as teaching signals for long-term plasticity (Gurney et al., 2015) or modulatory signals for short-term changes in neural dynamics (Durstewitz and Seamans, 2002; Benchenane et al., 2011). As medial prefrontal cortex connects directly and via

striatum to midbrain dopaminergic neurons (Watabe-Uchida et al., 2012) and receives direct input from them (Lammel et al., 2008; Beier et al., 2015), there is a potential direct link between reward-feedback mechanisms and adaptive behaviour (Laubach et al., 2015).

## Functional cell assemblies are potentially necessary for learning but not performance

That the recalled ensembles only appeared around clear episodes of behavioural learning means they are thus candidate cell assemblies (Harris, 2005): an ensemble that appeared during the course of learning. We distinguished here between structural and functional cell assemblies. In a structural assembly, the ensemble's activity pattern is formed by some underlying physical change, such as synaptic plasticity of the connections between and into the neurons of the ensemble (Harris, 2005; Holtmaat and Caroni, 2016), and is thus a permanent change. In a functional assembly, the ensemble's activity pattern is formed by some temporary modulation of existing connections - e.g. by new input or neuromodulation (Benchenane et al., 2011), and is thus a temporary change. Our analysis suggested that the recalled ensembles were a functional assembly, as they decayed before the end of the session in which they appeared, often decaying soon after the learning trial itself. We thus propose that this short-term memory ensemble is necessary only for the successful trial-and-error learning of a new rule, and not for the ongoing successful performance of that rule.

## Encoding in prefrontal cortex from the naive state

Consistent with prior reports of mixed selectivity in prefrontal cortex (Jung et al., 1998; Jun et al., 2010; Horst and Laubach, 2012; Rigotti et al., 2013), we could decode multiple task features from the joint activity of a small population of neurons. Extending these reports, we showed here that these encodings were position dependent, and that this encoding was exclusively retrospective during the inter-trial interval - despite there being no explicit working memory component to the rules. Our data thus show a short-term memory for multiplexed task features even in the absence of overt working memory demands.

One of our more unexpected findings was that we could reliably decode both the prior choice of direction and the prior trial's outcome across all sessions, regardless of whether they contained clear learning, externally-imposed rule changes, or neither these events. Our decoder used the vector of firing rates at a given maze position as input. Consequently, our ability to decode binary labels of prior events (correct/error trials or left/right locations) implies that there were well separated firing rate vectors for each of these labels. But this does not mean the neurons' firing rates were consistently related for a given label (such as a prior choice of the left arm of the maze). Indeed, it could imply anything from the two labels being encoded by the only two vectors of firing rates that ever appeared, to the two labels being encoded by two distinct groups of neurons whose firing rates within each group were never correlated. The reliable appearance of the same pattern of neuron co-activity only during learning thus implies that only during these sessions was the firing rate vector reliably similar. This suggests that learning to synchronise the encoded features, and not the learning of the encoding itself, is necessary for acquiring a new rule.

An interesting detail with potentially broad implications is that we could decode both the choice of prior direction and prior outcome from the very first session that each rat experienced the Y-maze. Either this implies that medial prefrontal cortex learnt rep-

resentations of direction and outcome so fast that they were able to make a significant contribution to decoding by population activity within the very first session. Or it implies that medial prefrontal cortex does not need to learn representations of direction and outcome, meaning that such encoding is always present. Future work is needed to distinguish which of the broad spectrum of features encoded by the prefrontal cortex are either consistently present or learnt according to task demands. Demarcating the classes of features that the prefrontal cortex innately or learns to remember would further advance our understanding of its contribution to adaptive behaviour.

# Acknowledgments

We thank Matt Jones for comments on the manuscript, and the Humphries' lab (Abhinav Singh, Javier Caballero, Mat Evans) for discussion. M.D.H and S.M. were supported by a Medical Research Council Senior non-Clinical Fellowship award (MR/J008648/1) to M.D.H.

# Methods

## Task description and electrophysiological data

For full details on training, spike-sorting, and histology see (Peyrache et al., 2009). Four Long-Evans male rats with implanted tetrodes in prelimbic cortex were trained on a Y-maze task (Figure 1a). Each recording session consisted of a 20-30 minute sleep or rest 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. Periods of slow-wave sleep were detected offline automatically from local field potential recordings (details in Peyrache et al., 2009).

During training, every trial started when the rat left the beginning of the start arm and finished when the rat reached the reward port at the end of its chosen goal arm. A correct choice of arm according to the current rule was rewarded with drops of flavoured milk. Each inter-trial interval lasted from the end-point of the trial, though any reward consumption, until the rat completed its self-paced return to the beginning of the start arm.

Each rat had to learn the current rule by trial-and-error. The rules were sequenced to ensure cross-modal shifts: 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 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 53 experimental sessions taken from the study of Peyrache et al. (2009), representing a set of training sessions from naive until either the final training session, or until choice became habitual across multiple consecutive sessions (consistent selection of one arm that was not the correct arm). In this data-set, each rat learnt at least two rules, and the four rats respectively contributed 14, 14, 11, and 14 sessions. We used 50 sessions here, omitting one session for missing position data, one in which the rat always chose the right arm (in a dark arm rule) preventing further decoding analyses (see below), and one for missing spike data in a few trials. Tetrode recordings were obtained from the first session for each rat. They 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. Simultaneous tracking of the rat's position was recorded at 30 Hz.

In order to identify ensembles and track them over each session, we first selected the $N$ neurons that were active in all the inter-trial intervals. The $N$ spike-trains of this core population were convolved with a Gaussian ($\sigma = 100$ ms) to obtain a spike-density function $f_k$ for the $k$th spike-train. All the recall analysis was repeated for different Gaussian widths ranging from 20 ms to 240 ms (Supplementary Fig. S4). Each spike-train was then Z-scored to obtain a normalised spike-density function $f^*$ of unit variance: $f_k^* = (f_k - \langle f_k \rangle)/\sigma_k$, where $\langle f_k \rangle$ is the mean of $f_k$, and $\sigma_k$ its standard deviation.

## Testing for reinforcement-driven ensembles

We sought to compare the co-activity of neurons within the core population across the inter-trial intervals of a session, in order to determine if the same pattern of co-activity recurred. To do so, for each inter-trial interval $t$ we first computed a pairwise similarity matrix $S_t$ between the spike-density functions for all $N$ neurons. Similarity here was the rectified correlation coefficient, retaining all positive values, and setting all negative values to zero, such that each entry $S_t(i, j)$ for the pair of neurons $(i, j)$ was in the range $[0, 1]$, from 0 meaning never co-active to 1 meaning identically co-active.

We then compared the core population's co-activity patterns between inter-trial intervals $t$ and $u$ by computing the pairwise similarity between $S_t$ and $S_u$. We do this by computing the rectified correlation coefficient between vectors of the unique values in $S_t$ and $S_u$, giving a scalar value $R(t, u) \in [0, 1]$. Note that this is why negative correlations between neurons were omitted: if we had not, then positive $R(t, u)$ could correspond to either a set of neurons that were similarly correlated in both intervals, or a set of neurons that were similarly anti-correlated in both intervals. Thus we used only pairwise similarity between neurons to disambiguate these two cases, and specifically identify co-activity between neurons.

By computing $R(t, u)$ for all pairs of inter-trial intervals in a session, we thus formed the recall matrix $R_I$ for session $I$, capturing the similarity of activity patterns between all inter-trial intervals in that session. We grouped the entries of $R$ into two groups according to the same type of inter-trial interval – predominantly whether they were intervals following correct or following error trials. These created the block diagonals $R_1$ and $R_2$ (such as $R_{error}$ and $R_{correct}$, as illustrated in Figure 2a). We summarised the recall between groups by computing the mean of each block. We detected statistically meaningful differences by computing the Kolmogorov-Smirnov test for a difference between the distributions of values in the two blocks.

In the main text, we report that there is higher average similarity in $R_{correct}$ than $R_{error}$ in many sessions. However, there was a strong tendency for inter-trial intervals following correct trials to be longer in duration than inter-trial intervals following error trials (Supplementary Fig. S1), and so the estimates of pairwise similarity may be biased. In order to dissect the contribution of the different durations we defined a null model. For each session we defined a shuffled recall matrix $\hat{R}$ obtained from the average of 1000 recall matrices computed on shuffled spike trains, with each shuffled spike-train keeping its inter-spike interval distribution fixed. In this way we destroyed any task-specific temporal pattern of the spike train and we quantify the contribution to pairwise similarity due solely to the length of the inter-trial interval. Our final residual recall matrix $\tilde{R} = R - \hat{R}$ is obtained as the difference between the Recall matrix and the average shuffled recall matrix (Figure 2a; Supplementary Fig. S1).

This dissection of the contribution of duration was why we used similarity ($[0, 1]$) rather

than correlation ($[-1, 1]$) between $S_t$ and $S_u$ to compute $R(t, u)$, and consequently $\hat{R}(t, u)$. First because it then allowed the residual recall value $\tilde{R}(t, u)$ to fall in the range $[-1, 1]$. Secondly because, if we had used correlation between $S_t$ and $S_u$, then if both $R(t, u)$, and $\hat{R}(t, u)$ were negative and $R(t, u) > \hat{R}(t, u)$, then the residual recall value would be positive ($\tilde{R}(t, u) > 0$); thus indicating a similarity between co-activity patterns in intervals $t$ and $u$ despite both data and shuffled recall matrices indicating a dissimilarity.

For the residual recall matrix, we summarised and tested the differences between the two groups (such as post-error and post-correct inter-trial intervals) in the same way as detailed above, given the new block diagonals $\tilde{R}_1$ and $\tilde{R}_2$. When grouping by session type, we plotted the difference between the block diagonals' means as $\Delta$ = mean($\tilde{R}_1$) - mean($\tilde{R}_2$).

## Behavioural analysis

A learning trial was defined following the criteria of the original study (Peyrache et al., 2009) as the first of three correct trials after which the performance was at least 80% correct for the remainder of a session. Only ten sessions contained a trial which met these criteria, and so were labelled "learning" sessions. We checked that these identified trials corresponded to an abrupt change in behaviour by computing the cumulative reward curve, then fitting a piecewise linear regression model: a robust regression line fitted to the reward curve before the learning trial, and another fitted to the reward curve after the learning trial. The slopes of the two lines thus gave us the rate of reward accumulation before ($r_{before}$) and after ($r_{after}$) the learning trial.

To identify other possible learning trials within each learning session, we fitted this piecewise linear regression model to each trial in turn (allowing a minimum of 5 trials before and after each tested trial). We then found the trial at which the increase in slope ($r_{after} - r_{before}$) was maximised, indicating the point of steepest inflection in the cumulative reward curve. The two sets of learning trials largely agreed: we checked our results using this set too.

Amongst the other sessions, we searched for signs of incremental learning by again fitting the piecewise linear regression model to each trial in turn, and looking for any trial for which ($r_{after} - r_{before}$) was positive. We found 22 sessions falling in this category in addition to the 10 learning sessions. We called those new sessions Minor-learning (Supplementary Fig. S2).

## Testing the onset and offset of recall

In order to identify when the recalled ensemble activity pattern first appeared in a learning session, we arranged its residual recall matrix in trial order. For each trial in turn (with a minimum of 3 trials before and 5 after), we formed the block diagonals $R_{before}$ and $R_{after}$ (see Figure 4a), respectively giving all pairwise recall scores between inter-trial intervals before and after that trial. The distance between recall before and after was measured using the Kolmogorov-Smirnov statistic: the maximum distance between the empirical cumulative distributions of $R_{before}$ and $R_{after}$. The trial that had the maximum positive distance (an increase in recall from $R_{before}$ to $R_{after}$) and had $P < 0.05$ was identified as the onset of the recalled activity pattern. Similarly, the trial with the maximum negative distance that corresponded to a decrease in recall from $R_{before}$ to $R_{after}$ and had $P < 0.05$ was identified as the offset of the recalled activity pattern. In all learning sessions we observed a strict ordering of onset occurring before offset, and both occurring before the final tested trial of the session.

## Decoder analysis

To test whether it was possible to predict task-relevant information in a position-dependent manner from the core population's activity we trained and tested a range of linear decoders (Hastie et al., 2009). In the main text we report the results obtained using a logistic regression classifier, as this is perhaps the easiest classifier to interpret.

We first linearised the maze in five equally-sized sections, with the central section covering the choice point of the maze. During each inter-trial interval, we computed the $N$-length firing rate vector $R^p$, with each element $r_j^p$ being the firing rate of the $j$th core population neuron at position $p$. For each session of $T$ inter-trial intervals and each section of the maze $p$, the set of population firing rate vectors $R^p(1), \ldots, R^p(T)$ was then used to train a linear decoder to classify the relevant binary task information, either: the previous trial's outcome (labels: 0,1), the previously chosen arm (labels: left,right), or the previous position of the light cue (labels: left,right). (We also trained all decoders on the next outcome, arm choice, and light position to test for prospective encoding). To avoid overfitting, we used leave-one-out cross-validation, where each inter-trial interval was held out in turn as the test target and the decoder was trained on the $T-1$ remaining inter-trial intervals. The accuracy of the decoder for position $p$ in a given session was thus the proportion of correctly predicted labels over the $T$ held out test inter-trial intervals.

Because the frequency of outcomes and arm choices were due to the rat's behaviour, chance proportions of correctly decoding labels was not 50%. To establish chance performance for each decoding, we fitted the same cross-validated classifier on the same set of firing rate vectors at each position, but using shuffled labels across the inter-trial intervals (for example, we shuffled the outcomes of the previous trial randomly). We repeated the shuffling and fitting 50 times. For displaying the results in Figure 5, we subtracted the mean of the shuffled results from the true decoding performance. Separate results for the true and shuffled decoders are plotted in Supplementary Fig. S5, panel a.

We report in the main text the results of using a logistic regression classifier. To check the robustness of our results, we also tested three further linear decoders: linear discriminant analysis; (linear) support vector machines; and a nearest neighbours classifier. Each of these showed similar decoding performance to the logistic regression classifier (Supplementary Fig. S6).

# References


Averbeck, B. B. and Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *J Neurosci*, 27:2204–2211.

Averbeck, B. B., Sohn, J.-W., and Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nat Neurosci*, 9:276–282.

Aziz-Zadeh, L., Kaplan, J. T., and Iacoboni, M. (2009). aha!: The neural correlates of verbal insight solutions. *Human brain mapping*, 30(3):908–916.

Baeg, E. H., Kim, Y. B., Huh, K., Mook-Jung, I., Kim, H. T., and Jung, M. W. (2003). Dynamics of population code for working memory in the prefrontal cortex. *Neuron*, 40(1):177–188.

Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141.

Beier, K. T., Steinberg, E. E., DeLoach, K. E., Xie, S., Miyamichi, K., Schwarz, L., Gao, X. J., Kremer, E. J., Malenka, R. C., and Luo, L. (2015). Circuit architecture of VTA dopamine neurons revealed by systematic input-output mapping. *Cell*, 162(3):622–634.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P., and Wiener, S. I. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. *Neuron*, 66(6):921–936.

Benchenane, K., Tiesinga, P. H., and Battaglia, F. P. (2011). Oscillations in the prefrontal cortex: a gateway to memory and attention. *Curr Opin Neurobiol*, 21(3):475–485.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, 68(5):815–834.

Burton, B. G., Hok, V., Save, E., and Poucet, B. (2009). Lesion of the ventral and intermediate hippocampus abolishes anticipatory activity in the medial prefrontal cortex of the rat. *Behav Brain Res*, 199(2):222–234.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.

Durstewitz, D. and Seamans, J. K. (2002). The computational role of dopamine D1 receptors in working memory. *Neural Netw*, 15(4-6):561–572.

Durstewitz, D., Vittoz, N. M., Floresco, S. B., and Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66(3):438–448.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nat Neurosci*, 19(3):479.

Euston, D. R., Gruber, A. J., and McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, 76(6):1057–1070.

Fellows, L. K. (2007). Advances in understanding ventromedial prefrontal function the accountant joins the executive. *Neurology*, 68(13):991–995.

Fujisawa, S., Amarasingham, A., Harrison, M. T., and Buzsaki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat Neurosci*, 11(7):823–833.

Gallistel, C. R., Fairhurst, S., and Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the national academy of Sciences of the united States of america*, 101(36):13124–13131.

Gurney, K. N., Humphries, M. D., and Redgrave, P. (2015). A new framework for cortico-striatal plasticity: Behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS Biol*, 13(1):e1002034.

Harris, K. D. (2005). Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6(5):399–407.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, Berlin.

Hok, V., Save, E., Lenck-Santini, P. P., and Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proc Natl Acad Sci U S A*, 102(12):4602–4607.

Holtmaat, A. and Caroni, P. (2016). Functional and structural underpinnings of neuronal assembly formation in learning. *Nature Neuroscience*, 19:1553–1562.

Hoover, W. B. and Vertes, R. P. (2007). Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. *Brain Struct Funct*, 212:149–179.

Horst, N. K. and Laubach, M. (2012). Working with memory: evidence for a role for the medial prefrontal cortex in performance monitoring during spatial delayed alternation. *Journal of neurophysiology*, 108(12):3276–3288.

Hyman, J. M., Ma, L., Balaguer-Ballester, E., Durstewitz, D., and Seamans, J. K. (2012). Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proceedings of the National Academy of Sciences*, 109(13):5086–5091.

Jones, M. W. and Wilson, M. A. (2005). Theta rhythms coordinate hippocampal-prefrontal interactions in a spatial memory task. *PLoS Biol*, 3(12):e402.

Jun, J. K., Miller, P., Hernndez, A., Zainos, A., Lemus, L., Brody, C. D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. *J Neurosci*, 30(3):916–929.

Jung, M. W., Qin, Y., McNaughton, B. L., and Barnes, C. A. (1998). Firing characteristics of deep layer neurons in prefrontal cortex in rats performing spatial working memory tasks. *Cereb Cortex*, 8(5):437–450.

Kaplan, R., King, J., Koster, R., Penny, W. D., Burgess, N., and Friston, K. J. (2017). The neural representation of prospective choice during spatial planning and decisions. *PLOS Biology*, 15(1):e1002588.

Karlsson, M. P., Tervo, D. G., and Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338(6103):135–139.

Lammel, S., Hetzel, A., Hckel, O., Jones, I., Liss, B., and Roeper, J. (2008). Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron*, 57(5):760–773.

Laubach, M., Caetano, M. S., and Narayanan, N. S. (2015). Mistakes were made: neural mechanisms for the adaptive control of action initiation by the medial prefrontal cortex. *Journal of Physiology-Paris*, 109(1):104–117.

Machens, C. K., Romo, R., and Brody, C. D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J Neurosci*, 30:350–360.

Miller, E. K. (2000). The prefontral cortex and cognitive control. *Nature reviews neuroscience*, 1(1):59–65.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

Narayanan, N. S. and Laubach, M. (2008). Neuronal correlates of post-error slowing in the rat dorsomedial prefrontal cortex. *J Neurophysiol*, 100:520–525.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12:919–926.

Powell, N. J. and Redish, A. D. (2016). Representational changes of latent strategies in rat medial prefrontal cortex precede changes in behaviour. *Nature Communications*, 7.

Ragozzino, M. E. (2007). The contribution of the medial prefrontal cortex, orbitofrontal cortex, and dorsomedial striatum to behavioral flexibility. *Annals of the New York Academy of Sciences*, 1121(1):355–375.

Ragozzino, M. E., Wilcox, C., Raso, M., and Kesner, R. P. (1999). Involvement of rodent prefrontal cortex subregions in strategy switching. *Behav Neurosci*, 113(1):32–41.

Rich, E. L. and Shapiro, M. (2009). Rat prefrontal cortical neurons selectively code strategy switches. *Journal of Neuroscience*, 29(22):7208–7219.

Rich, E. L. and Shapiro, M. L. (2007). Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. *J Neurosci*, 27(17):4747–4755.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590.

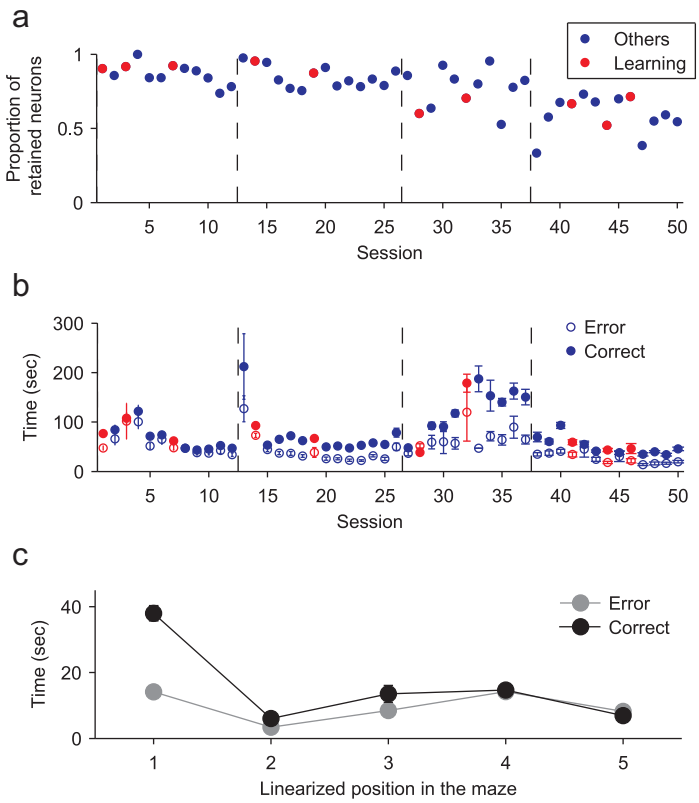
Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Spellman, T., Rigotti, M., Ahmari, S. E., Fusi, S., Gogos, J. A., and Gordon, J. A. (2015). Hippocampal-prefrontal input supports spatial encoding in working memory. *Nature*, 522(7556):309–314.

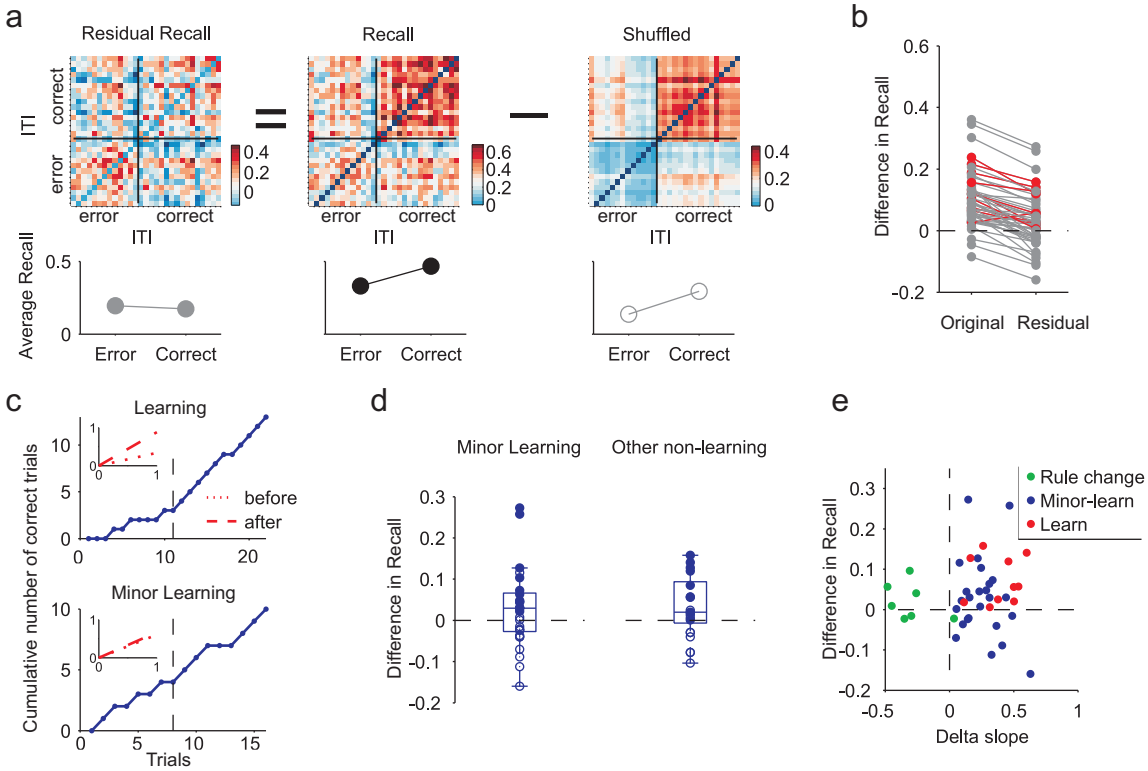
Sul, J. H., Kim, H., Huh, N., Lee, D., and Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, 66(3):449–460.

Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, 74:858–873.

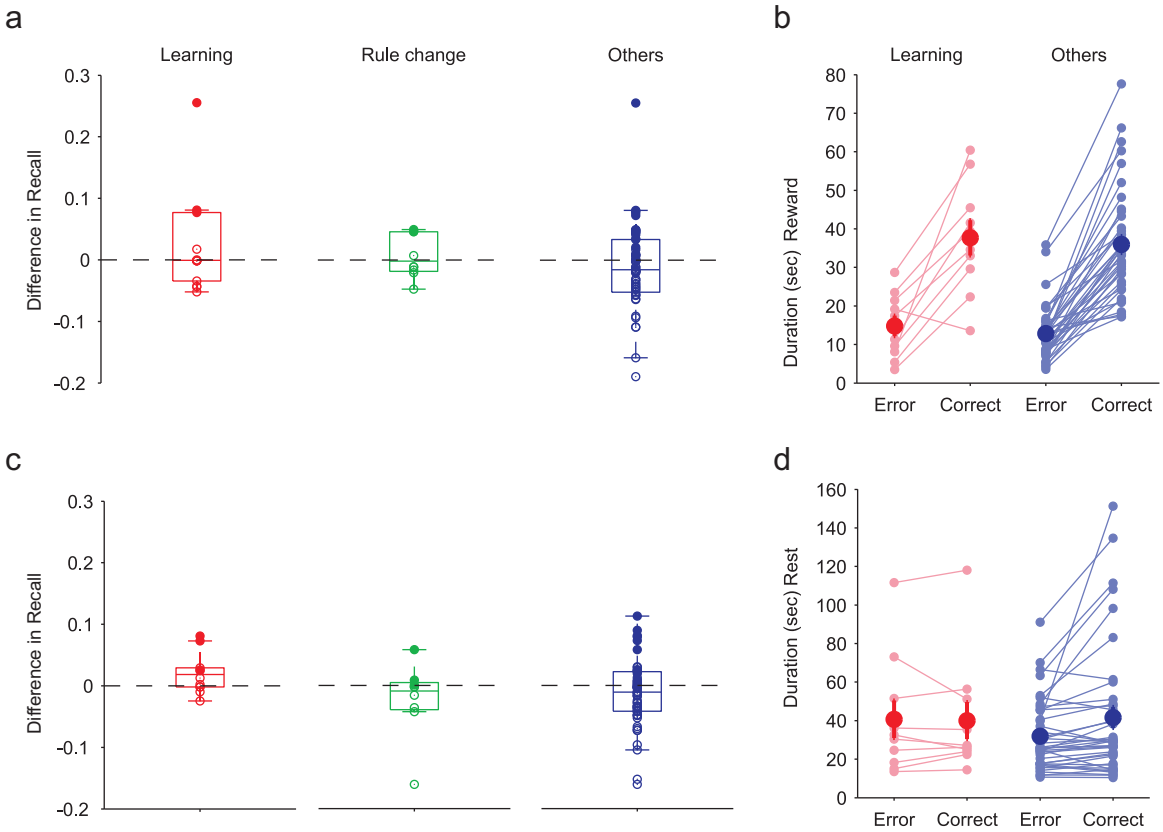
## Supplementary Figures



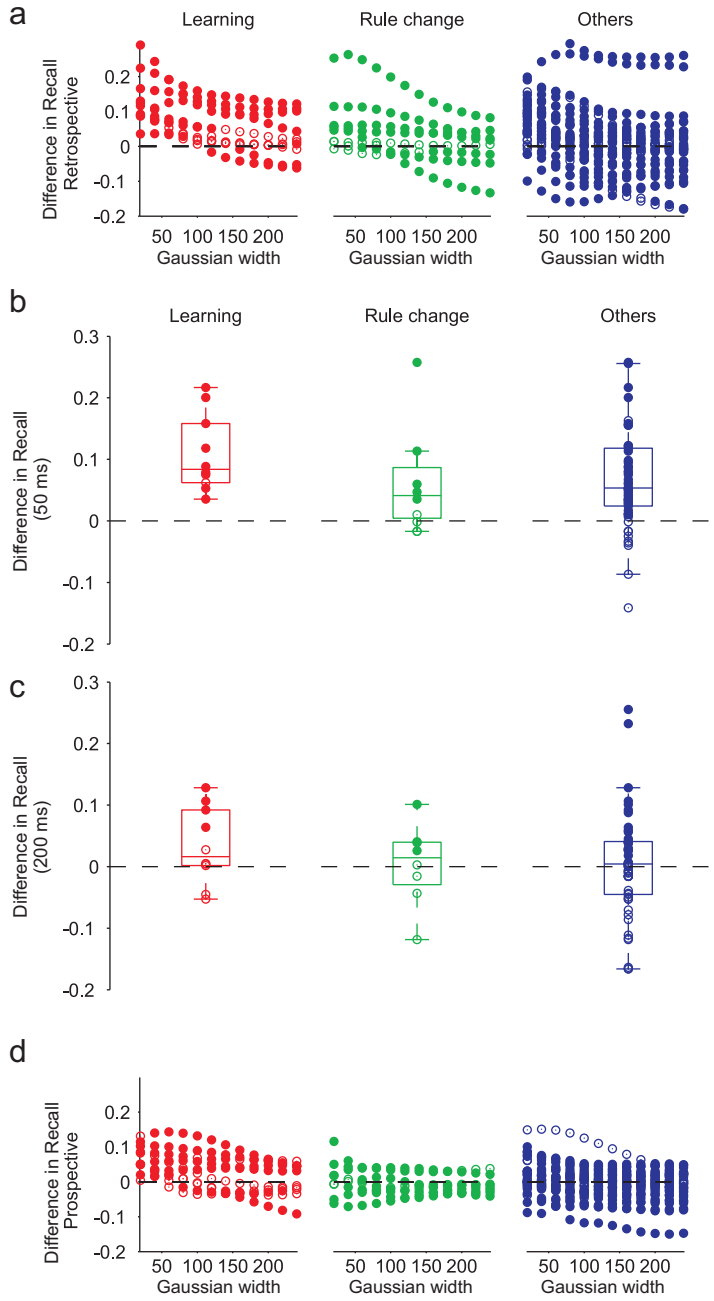

**Supplementary Figure S1**.
Statistics of neural populations and time periods during the inter-trial intervals for the 50 retained sessions.
(**a**) Proportion of neurons retained in the core ensemble were those that fired in every inter-trial interval. The black vertical dashed lines separate the sessions for each of the four rats.
(**b**) Durations of the inter-trial intervals within each session, given as the mean ± SEM duration in seconds, separated into post-correct (filled symbols) and post-error (open symbols) inter-trial intervals. Red symbols are the learning sessions.
(**c**) Time spent along the maze during the inter-trial intervals, given as the mean ± SEM seconds spent across all the animal and all the sessions for post-correct (black) and post-error (gray) inter-trial intervals. The maze has been linearised and divided in 5 equal sized sections, with position 1 being the reward location, position 3 the choice point of the Y-maze, and position 5 the end of the start arm – see Figure 5 (main text) for a schematic.



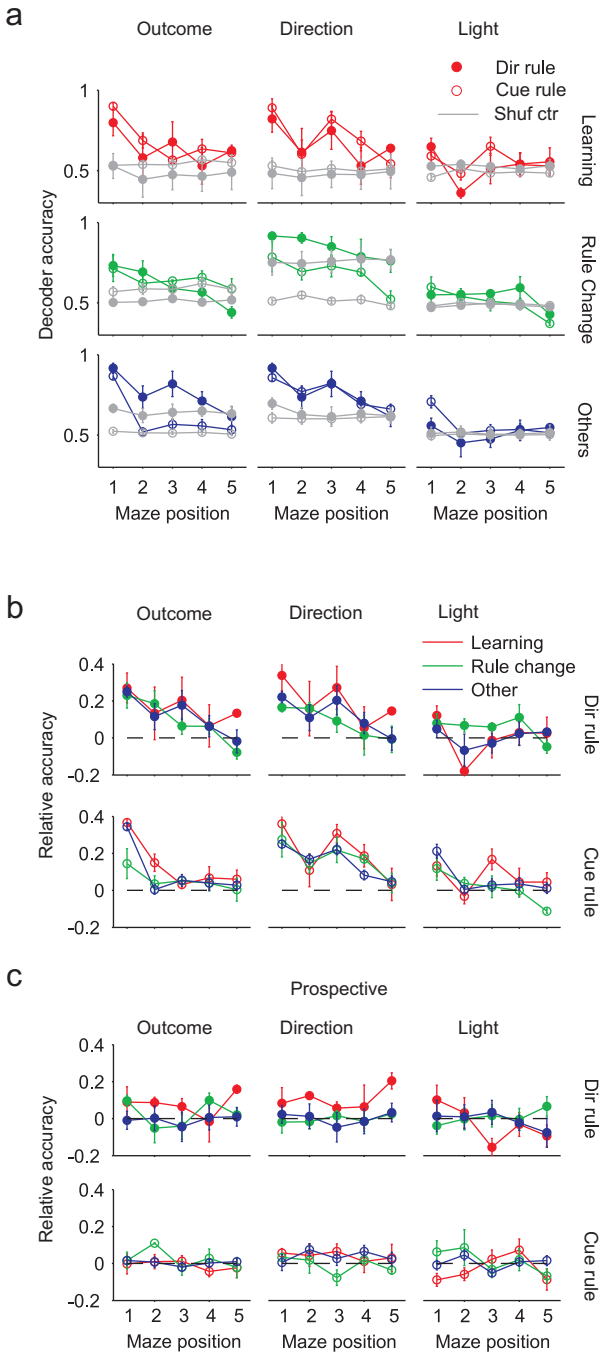

**Supplementary Figure S2**.
Recall of neural ensembles is learning-specific.
(**a**) Example correction of the recall matrix to remove the confounding effect of different durations of the post-correct and post-error inter-trial intervals. The residual recall matrix was obtained as the difference between the recall matrix and the mean matrix obtained from the shuffled inter-spike intervals (upper panels). For this example, the average recall values between error and correct intervals showed higher correlation among correct intervals in the shuffled model; this reversed the difference in recall between error and correct intervals (bottom panels) - in this case, ruling out a potentially higher recall during correct trials.
(**b**) Comparison of the difference between average correct recall and error recall (Delta recall) before and after correction by the shuffled control data. Red symbols are the learning sessions.
(**c**) To check whether the recall effect was specific to sessions showing abrupt learning (top panel; Figure 2, main text), we identified a subset of the other sessions with potential incremental or "minor" learning. These minor-learning sessions were any in which the curve of cumulative rewards contained a detectable upward inflection, as shown by the existence of any trial with a greater slope in a regression line after that trial than before it (insets, red lines). The vertical black dashed line is the identified learning trial.
(**d**) The difference between average correct recall and error recall (Delta recall) for the minor-learning and remaining other sessions. No systematic recall effect was observed for the minor-learning sessions, suggesting the recall effect was specific to abrupt transitions in behaviour.
(**e**) Relationship between behavioural change and the strength of recall. The difference between average correct recall and error recall (Delta recall) is plotted as a function of the difference between the slopes of the fitted lines before and after the learning trial (Delta slope). Sessions: learning (red), rule change (green), and minor-learning (blue). Delta slope for each rule change session was computed with respect to the rule change trial.



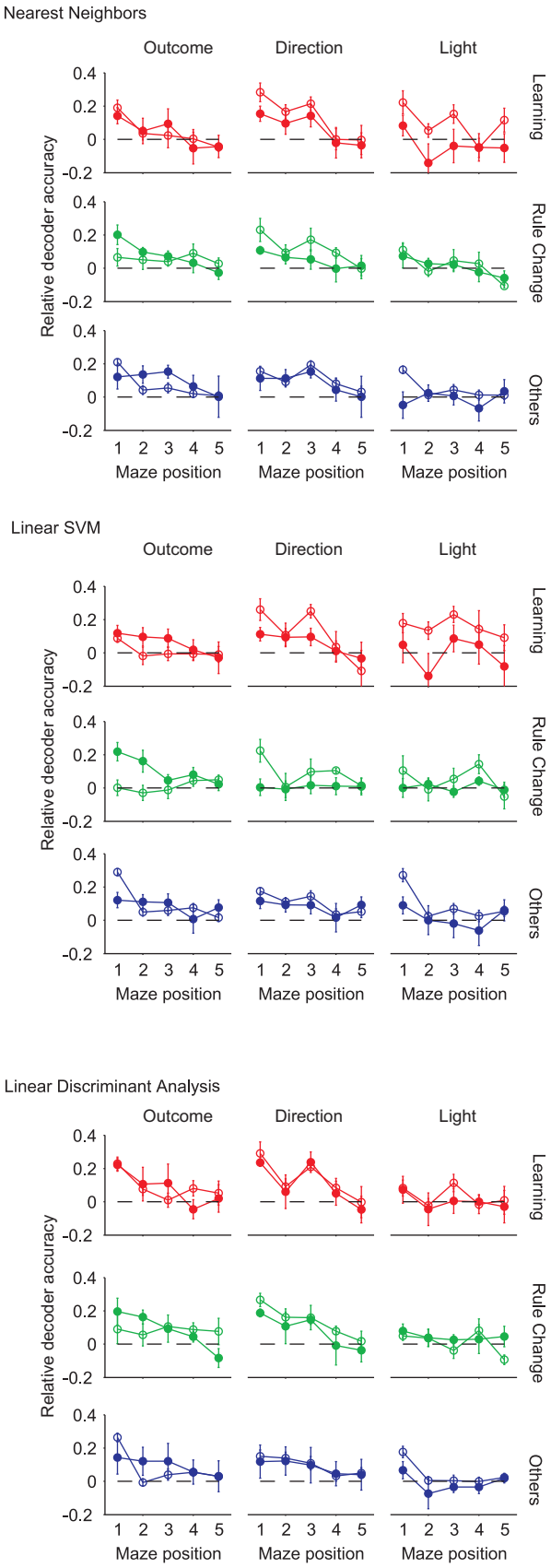

**Supplementary Figure S3**. Neural ensemble recall is not specific to reward consumption.
(**a**) The difference in average recall between post-error and post-correct intervals when using neural activity only at the reward location (position 1 in Figure 5a). No systematic post-reinforcement recall occurs at the reward location, for any session type.
(**b**) Duration of time spent by the animals in the reward location (position 1 in Figure 5a) for post-error and post-correct inter-trial intervals. Learning sessions (red) and all other sessions (blue) have similar average durations of time in the reward location for post-correct intervals and for post-error intervals (note that all others here include "rule change" and "others" of panel **a**). Each pair of dots is a session. The bold symbols are mean ± SEM of the entire distribution.
(**c**) As panel **a**, but using neural activity only during the self-paced return (positions 2-5 in Figure 5a) to the start position. The learning sessions show a weakly-positive reinforcement-specific recall.
(**d**) As panel **b**, for the duration of the self-paced return across positions 2-5 in the maze.




**Supplementary Figure S4**. Time-scale dependence of recall.
(**a**) Dependence of the recall of ensemble activity on the temporal precision of spike-train correlation. Here we plot the distribution of Delta recall across sessions as a function of the Gaussian width used to convolve the spike-trains. Retrospective recall, the difference in recall between intervals after correct and after error trials. Delta recall greater than zero indicates the interval similarity matrices were more correlated for correct than error intervals. Each symbol is one session. Filled circles indicate a difference at $p < 0.05$ between the distributions of recall values in the error and correct intervals (Kolmogorov-Smirnov test).
(**b**) Comparison of Delta recall across learning, rule change, and other sessions after spike-train convolution with a Gaussian 50 ms wide.
(**c**) As for panel **b**, but for a Gaussian 200 ms wide.
(**d**) Dependence of the prospective recall of ensemble activity on the temporal precision of spike-train correlation. As for panel A, but here we plot the prospective Delta recall, the difference in recall between intervals before correct and before error trials. We only see a systematic recall at the smaller tested Gaussian widths ($\leq 40$ ms), which is likely a reflection of the stronger retrospective recall effect at these widths.




**Supplementary Figure S5**.
Further decoding analysis.
(**a**) Decoding can be near perfect. Here we replot the breakdown of the decoding results in Figure 5**d** as the absolute accuracy of the decoders (where 1 is maximum, indicating correct prediction of every held-out inter-trial interval). Each data point is the mean ± SEM accuracy at that maze position The control results on shuffled inter-trial interval labels are shown as grey lines. The "other" sessions were plotted in Figure 5**c**.
(**b**) Comparison of above-chance decoding performance between the same rule types. Each data point is the mean ± SEM accuracy in excess of chance (dashed line) over the indicated combination of session type and rule type.
(**c**) Comparison of prospective decoding performance between the same rule types, confirming the absence of the prospective encoding of task-relevant information. Similar to panel **b**, here we plot the mean ± SEM above-chance accuracy of decoding prospective outcome, direction, or cue position, separately for sessions with direction or light rules. As in panel **b**, decoding accuracy is normalised by the corresponding shuffled control decoding performance (where 0 is identical to shuffled controls).




**Supplementary Figure S6**.
Robustness of the retrospective encoding of task-relevant information. Using the same layout as Figure 5**d**, here we summarise the decoding performance of three further classifiers we tested on the data to check the robustness of the decoding results. Top: Nearest Neighbors; middle: linear Support Vector Machine; bottom: Linear Discriminant Analysis.