Learning of distant state predictions by the orbitofrontal cortex in humans

Abbreviated title: OFC in distant state predictions

G. Elliott Wimmer[1,2] and Christian Büchel[3]

[1] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London, England

[2] Wellcome Centre for Human Neuroimaging, University College London, London, England

[3] Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Corresponding author:
G. Elliott Wimmer
Max Planck UCL Centre for Computational
Psychiatry and Ageing
Wellcome Centre for Human NeuroImaging
10-12 Russell Square
University College London
e.wimmer@ucl.ac.uk

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgments: The authors would like to thank Lara Austerman, Lea Kampermann, and Sven Schönig assistance in behavioral and fMRI data collection and helpful discussions. The authors are grateful to the University of Minnesota Center for Magnetic Resonance Research for providing the image reconstruction algorithm for the simultaneous multi-slice acquisitions. This work was supported by the Deutsche Forschungsgemeinschaft (Research Fellowship to G.E.W.; DFG SFB TRR 58 and SFB 936 to C.B.), and the European Research Council (Grant ERC-2010-AdG_20100407).

Abstract

Representations of our future environment beyond what is immediately perceptible are critical for planning and decision making. Previous research in humans has demonstrated that the hippocampus is critical for forming and retrieving associations and that the orbitofrontal cortex - ventromedial frontal cortex (OFC-VMPFC) is an important region for representing information about recent states. However, we do not know how the brain acquires predictive representations during goal-directed based learning. When goals are more than a few seconds in the future, such predictive representations are critically important for adaptive decision making. We examined the acquisition of predictive information by recording brain activity with fMRI while participants learned to find rewards in multiple different Y-maze environments, where exploration of each maze lasted more than 40 seconds. Behaviorally, participants exhibited rapid acquisition of optimal choices after only one or two experiences. Neurally, we found that activity in the hippocampus was highest during initial exposure and then decayed across the remaining repetitions of each maze, consistent with a role in rapid encoding of the maze experience. Importantly, we found that across learning, multivariate patterns in the OFC-VPFC but not the hippocampus or visual cortex came to represent predictive information about states ~30 seconds in the future. By the end of learning, these representations were of similar strength as representations of the current state. Our findings provide a mechanism by which the brain can build models of world that span long timescales, allowing predictions to be made about distant future events.

Introduction

Making predictions about the future has been proposed to be a core organizing principle of brain function (Rao and Ballard, 1999; Friston, 2005; Clark, 2013). Over short timescales, sensory and motor systems enable prediction of immediate stimuli and consequences of our actions (Rao and Ballard, 1999; den Ouden et al., 2012; Summerfield and de Lange, 2014). Over longer timescales, predictions about the future state of the world – and associated positive and negative consequences of our potential actions – are essential for goal-directed decision making and planning for the future. While the neural mechanisms of prediction in sensory and motor systems have been the subject of extensive research, we know little about the neural systems involved in learning to make predictions from experiences that extend beyond several seconds.

Neural systems including the hippocampus and the orbitofrontal cortex (OFC) have emerged as potential substrates for representing both spatial and abstract maps of the environment (Eichenbaum and Cohen, 2001; Wilson et al., 2014; Behrens et al., 2018). Building on the original discovery of place representations in hippocampal neurons (O'Keefe and Nadel, 1978), the hippocampus has been proposed to encode relations between stimuli that can combine to form a cognitive map of the environment (Eichenbaum and Cohen, 2001). Supporting a role for hippocampal representations in active prediction, research in animals has shown that when freely navigating an environment, activation of ‘place cell’ sequences can predict future movement trajectories (Pfeiffer and Foster, 2013; Olafsdottir et al., 2018). In humans, the hippocampus has also been shown to support the reactivation of visual associations originally separated by a few seconds (Wimmer and Shohamy, 2012). Other fMRI studies have found evidence for visual cortex reactivation for similar short timescale

associations (Doll et al., 2015; Ekman et al., 2017), however, especially over longer timescales, such predictive knowledge must be supported by a still undetermined learning system outside sensory cortex (Summerfield and de Lange, 2014).

Research on the orbitofrontal cortex has recently found support for related cognitive functions. While the OFC has long been viewed as critical for value representation and flexible learning, a unifying account has been proposed that the OFC represents a cognitive map of relevant states in the environment (Wilson et al., 2014; Bradfield et al., 2015; Stalnaker et al., 2015; Schuck et al., 2016; Schuck et al., 2018). Recently, using fMRI in humans Schuck et al. (2016) demonstrated that the OFC represented recently-observed hidden contextual information that was necessary to for task performance. Interestingly, results from a recent study in rodents indicate that the OFC and the hippocampus interact to represent task-relevant structure and guide behavior (Wikenheiser and Schoenbaum, 2016; Wikenheiser et al., 2017). However, it is not known whether and how the hippocampus and/or OFC acquire and represent predictions during goal-directed behavior.

From a computational perspective, learning longer timescale predictions poses difficulties for reinforcement learning models that track only the value of states. Temporal difference learning models are constrained by the slow propagation of errors back through preceding states (Sutton and Barto, 1998), and neurally, the reward prediction error signal carried by midbrain dopamine neurons loses its fidelity beyond 5-10 s (Fiorillo et al., 2008; Kobayashi and Schultz, 2008). Alternative computational models that increase learning speed have been proposed, including eligibility traces (Sutton and Barto, 1998) as well as models that explicitly represent state transitions such as

successor representations and model-based learning (Daw et al., 2005; Stachenfeld et al., 2017; Gershman, 2018). Positive evidence for learning of long timescale prediction in the hippocampus or OFC would lend novel neural support to successor representation or model-based accounts.

To investigate learning mechanisms supporting the prediction of distal future states in humans, we focused on initial learning as participants navigated multiple unique Y-maze environments, each of > 40 s duration. By studying initial learning in relatively long-duration experiences, our goal was to better understand learning in a situation like those outside the lab where rapid learning is evident after only one or a few experiences. We predicted that the hippocampus would show higher activity during initial exposure to each maze, reflecting initial encoding of a representation of the environment. Critically, we also hypothesized that predictions of distal states ~ 30 s in the future would be reflected in patterns of activity in medial OFC / ventromedial PFC (OFC-VMPFC) or hippocampus.

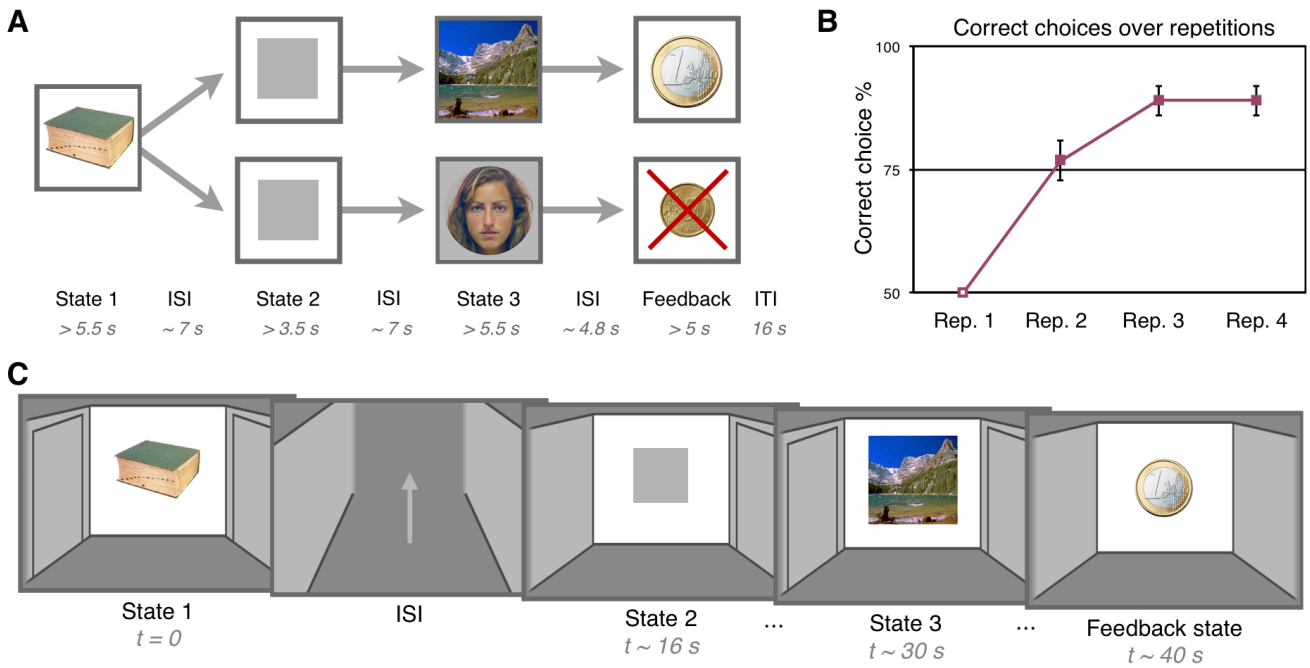

Figure 1. Learning phase task design and behavioral results. (A) Learning phase maze trial structure. Participants made a left or right choice in the first state and then proceeded through guided choices in state 2 and state 3, followed by reward or loss in the feedback state. Maze-unique stimuli were presented in state 1 and state 3. (B) Mean learning phase performance across four repetitions of each maze. Half of initial repetitions ended in reward and half in loss, and this mean 50 % level of performance is indicated by the open square at repetition 1. (C) Illustration of the participant's screen view, a cartoon room with potential left and right door options in each state followed by a path through a hallway between states.

Materials and methods

Participants. A total of 38 subjects participated in the experiment. Participants were fluent German speakers with normal or corrected-to-normal vision and no self-reported neurological or psychiatric disorders or depression as indicated by questionnaire. Three participants were excluded based on behavior: one for previous participation in a related experiment, one for sleeping during the task, and one for chance-level performance.

Thus, data from 35 participants were included in the behavioral analysis (21 female; mean age, 25.7 years; range, 19-31 years). After additional fMRI-based exclusion (detailed below), data from 32 remaining participants (18 female) were included in the univariate fMRI analysis. One additional participant was excluded from the multivariate analysis due to an error in the localizer phase. All participants were remunerated for their participation, and participants were additionally compensated based on the reward received during maze learning. The Ethics committee of the Medical Chamber Hamburg approved the study and all participants gave written consent.

Learning phase. The navigation learning task consisted of a learning phase followed by a localizer phase (**Figure 1A, C**). In the learning phase, across four repetitions for each of eight Y-maze sequences (plus two final non-analyzed mazes), participants attempted to learn the correct initial choice at state 1 in order to progress through the correct state 3 and reach to a reward at the terminal feedback state. Maze- and state-unique stimuli from 3 categories (faces, scenes, and objects) were presented at state 1 and state 3 to enable decoding of state representations across learning. For the sequence design, we utilized an adapted Y-maze structure with a choice at state 1 and instructed choices at states 2 and 3 to both allow for rapid learning (by limiting the structure to two branches) and relatively long maze durations without excessive scan duration.

In each maze, the participant navigated through an illustrated series of rooms (states) with hallways in-between states (**Figure 1A, C**). At each stage in a maze, focal stimuli were presented in the center wall of an illustrated room. Two walls with space for doors extended to the left and right of the center wall. In state 1 and state 3, a maze- and

state-unique stimulus was presented on the center wall. In state 2, a gray square was presented on the central wall. In state 1 and 3, both the left and right door options were initially visible. In state 2, only the single available pseudo-randomly determined (and fixed throughout learning) left or right door option was visible. In state 3, the non-available door was hidden after a short duration. In the reward state, no door options were visible. Between states, an illustrated hallway was presented with an occasional central upward-pointing arrow indicating to participants to press a key to continue advancing to the next state.

Maze duration, from state 1 until the end of the feedback state, was on average 48.5 s for the eight mazes of interest (range 38.3-77.4 s). The mean time between state 1 onset and reward feedback indicating the accuracy of the state 1 choice was 42.8 s (range 32.9-71.4 s). This delay from stimulus and action to feedback is well past the short range of time where dopamine responses show fidelity to a reward prediction error signal (Fiorillo et al., 2008; Kobayashi and Schultz, 2008). The mean time between state 1 onset and state 3 onset was 31.2 s (range 22.1-58.9 s). For the first participant, total maze duration was ~10 s shorter due to abbreviated inter-state periods (36.7 s mean duration), leading to shorter durations between state 1 and state 3 or reward. The delays between repetitions of the same maze were on average more than 4 min. Specifically, the delay between repetition 1 and repetition 2 was 4.4 min (mean 266.0 s, range 104.4 - 430.2). The delay between repetition 2 and repetition 3 was 4.6 min from (mean 278.5 s, range 102.2 - 1294.6). The delay between repetition 3 and repetition 4 was 6.9 min (mean 414.0 s, range 250.5 - 1266.0). This large delay strongly reduces the likelihood of between-repetition working memory maintenance as an explanation of learning

performance (Collins and Frank, 2012; Wimmer and Poldrack, 2018), a common problem in reward learning paradigms where learning repetitions are separated only by several seconds on average.

At state 1, an illustration of a room with a superimposed central stimulus was presented for 4 s while a small white fixation cross was shown over the stimulus. When the fixation cross disappeared, participants were allowed to make their choice. After making a choice, the selected door opened slightly and the central stimulus moved to the selected door for 1.5 s. Next, the screen changed to a hallway phase for the inter-state period (described below). In states 2 and 3, the room illustration and central stimulus were shown for 2 s (state 2) or 4 s (state 3), after which the stimulus moved to the computer-determined (and still closed) door. Next, participants selected the indicated correct left or right button (for a maximum 30 s duration). A warning appeared if the participant was too late or if they selected the incorrect door. The selected door then opened slightly and the central stimulus moved to the door for 1.5 s. The screen then changed to the hallway view. When participants reached the feedback state, the center wall was blank. They were required to press the bottom button to reveal the feedback. If the correct choice was made (or for the first repetition, if the maze had been assigned to be rewarded), a euro coin was displayed. Alternatively, a 50-cent euro coin with an overlaid red ‘X’ was displayed to indicate loss feedback. The feedback image was shown for 5 s. During the inter-state period between all states, after a 0.25 s blank screen, an illustration of a hallway was presented. An upward-pointing arrow appeared after 0.25 s. Participants progressed to the next state by pressing the top button when the arrow appeared. A series of two arrows, separated by a short delay, followed the first and

second state, while one arrow followed the third state. Trials were followed by a mean 16 s (range 15-18.5) inter-trial-interval during which a central white fixation cross was presented; the fixation cross changed to black for the 1 s preceding the start of the next maze.

For the eight mazes of interest, reward feedback was given on the first exposure for 4 mazes (independent of the state 1 left or right choice of the participant) and miss feedback was given on the first repetition for the other 4 mazes. Based on the initial state 1 left or right choice of the participant and whether the maze was assigned to end in an initial reward or loss, the ‘correct’ and ‘incorrect’ choice was assigned for subsequent repetitions for each maze. The mazes were presented in a pseudo-random order with staged introduction of new mazes as follows. The first four mazes were shown through three repetitions, such that each maze was repeated one time before another maze could be repeated again. Intermixed with the fourth repetition, two mazes from the second set were interleaved. At the fourth repetition for the second set of mazes, two additional mazes were interleaved. These additional mazes always ended in a loss in order to better balance the reward and loss associations for probe questions in the subsequent localizer. During scanning, the experiment was divided into four blocks of 10 maze trials per block.

To measure internal activation of future state representations, each maze in the experiment included three categories of stimuli: faces, scenes, and objects. These were assigned to state 1 and the two alternative state 3 states. Color pictures of female faces set on a gray background were drawn from the Stanford Face Database, while scenes were drawn from an internal database; both sets of stimuli have previously been utilized

to identify internal reactivation of state representations (Wimmer and Shohamy, 2012). Color pictures of objects were drawn from a previously used set of images compiled via internet search (Wimmer et al., 2014), composed of medium-sized items that can be manually interacted with (e.g. a book, water bottle, or guitar) set on a white background. Three pseudo-random orderings of the maze- and state-unique picture stimuli were used for counterbalancing. To indicate reward or loss feedback, a picture of a 1 euro coin was used for reward feedback while a picture of a 50 cent euro coin with a red 'X' drawn through it was used for miss feedback. Additionally, for reward feedback, outwardly-moving gold-colored dots were shown behind the coin to allow us to examine a potential motion-sensitive cortical response to rewards and anticipated rewards. To achieve this effect, a rapidly presented sequence of 12 static images started with only a few visible dots around the coin and then progressed as these and additional dots moved outward, giving the impression of a burst of dots moving away from the coin. These 12 images were repeated in sequence 3 times during the 5 s feedback period.

Before entering the MRI scanner, participants completed 1 trial of a practice maze. This practice maze was repeated again inside the MRI before the start of the experiment to ensure participants were familiarized with the screen and button box.

Localizer phase. To measure patterns of activity associated with the stimulus categories seen at states 1 and 3 in the learning phase, we next collected a "localizer" phase in the MRI scanner. During each trial, participants viewed either a stimulus from the learning phase, a scrambled object stimulus, or moving dot stimulus. For the face, scene, and object stimuli from the learning phase, participants were instructed to try to remember

whether that stimulus was followed by reward or loss during learning so that they could perform well on occasional probe questions.

The localizer phase included 155 total trials, composed of 120 face, scene, and object trials, 20 scrambled object trials, and 15 “motion” trials with a moving dot background. The motion stimuli were composed of a gray hexagon with black outline and blue dots shifting across 12 static images, modeled after the reward feedback background gold dots seen during the learning phase. On each trial, static stimuli were presented for 2 s. For motion trials, the stimulus duration was 3 s. If a memory probe followed the picture, the text “Reward? ^” and “Loss? v" appeared above and below the stimulus picture, respectively. Participants indicated their response with corresponding the top and bottom buttons. After a 0.5 s delay followed by a 0.25 s fixation period, a confidence rating scale appeared along with the text “How certain?”. The rating scale was composed of 4 levels of confidence: “not at all”, “a little”, “medium”, and “very”, which participants navigated using the left and right buttons on the box, confirming their answer with the bottom button. After a 0.25 s pause, the task continued with the inter-trial-interval indicated by a fixation (mean 2.5 s duration; range, 1.5-4.5 s).

The localizer phase was composed of 3 mini-blocks, each containing 40 face, scene, and object stimuli, 4-8 scrambled object stimuli, and 5 motion stimuli. Face, scene, and object trials were pseudo-randomly ordered, while scrambled object stimuli were shown at the end of the mini-block in sets of 4 preceding (and following) a set of 5 motion stimulus trials. Probe questions that assessed memory for the reward or loss associated with that stimulus during learning were included after 25% of the face, scene, and object trials to increase participant attention and to assess memory for the reward or

loss association. The base list was modified for each participant, such that stimuli that were not seen during the learning phase were replaced with an additional 2 s inter-trial-interval (mean, 9.9; range, 3-12). One participant was excluded from localizer analyses and multivariate analyses because an error in list creation led to an exclusion of 57 trials, yielding too few trials for accurate category classification.

After exiting the MRI scanner, participants completed two additional measures. We collected the Beck Depression Inventory (BDI). The second measure we collected was the operation-span task (OSPAN), which was used as an index of working memory capacity (Lewandowsky et al., 2010; Otto et al., 2013). In the OSPAN, participants made accuracy judgments about simple arithmetic equations (e.g. ‘2 + 2 = 5’). After a response, an unrelated letter appeared (e.g. ‘B’), followed by the next equation. After arithmetic-letter sequences ranging in length from 4 to 8, participants were asked to type in the letters that they had seen in order, with no time limit. Each sequence length was repeated 3 times. In order to ensure that participants were fully practiced in the task before it began, the task was described in-depth in instruction slides, followed by 5 practice trials. Scores were calculated by summing the number of letters in fully correct letter responses across all 15 trials (Otto et al., 2013; Wimmer et al., 2018).

The experimental tasks were presented using Matlab (Mathworks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997). The task was projected onto a mirror above the participant’s eyes. Responses during fMRI scanning were made using a 4-button interface with a “diamond” arrangement of buttons (Current Designs, Philadelphia, PA). At the end of the experiment, participants completed a paper questionnaire querying their knowledge of the task instructions and learning strategy. For all parts of

the experiment, verbal and on-screen instructions were presented in German; for the methods description and task figures, this text has been translated into English.

fMRI Data Acquisition. Whole-brain imaging was conducted on a Siemens Trio 3 Tesla system equipped with a 32-channel head coil (Siemens, Erlangen, Germany). fMRI measurements were performed using single-shot echo-planar imaging with parallel imaging (GRAPPA, in-plane acceleration factor 2; TR = 1240 ms, TE = 26 ms, flip angle = 60; 2 x 2 x 2 mm voxel size; 40 axial slices with a 1 mm gap) (Griswold et al., 2002) and simultaneous multi-slice acquisitions ("multiband", slice acceleration factor 2) (Feinberg et al., 2010; Moeller et al., 2010; Xu et al., 2013) as described in (Setsompop et al., 2012). The corresponding image reconstruction algorithm was provided by the University of Minnesota Center for Magnetic Resonance Research. Slices were tilted approximately 30° relative to the AC–PC line to improve signal-to-noise ratio in the orbitofrontal cortex (Deichmann et al., 2003). Head padding was used to minimize head motion. Three participants were excluded for excessive head motion (an average of 5 or more > 2.0 mm framewise displacement translations from TR to TR per learning phase block).

During the learning phase, four functional runs of an average of ~10 min and 45 s were collected, each including 10 maze trials. During the localizer phase, one functional run of an average of ~15 min was collected. For each functional scanning run, four discarded volumes were collected prior to the first trial to allow for magnetic field equilibration and to collect reference data for multiband reconstruction.

Structural images were collected using a high-resolution T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) pulse sequence (1 x 1 x 1 mm voxel size) between the learning phase and the test phase.

Behavioral Analysis. Behavioral analyses were conducted in Matlab 2016a (The MathWorks, Inc., Natick, MA). Results presented below are from the following analyses: t-tests vs. chance for learning performance, within-group (paired) t-tests comparing differences in reward- and loss-associated stimuli across conditions, Pearson correlations, and Fisher z-transformations of correlation values. We additionally tested whether non-significant results were weaker than a moderate effect size using the Two One-Sided Test (TOST) procedure (Schuirmann, 1987; Lakens, 2017) and the TOSTER library in R (Lakens, 2017). We used bounds of Cohen's d = 0.53, where power to detect an effect in the included group of 32 participants is estimated to be 80% (d was adjusted accordingly when analyzing subsets of participants).

fMRI Data Analysis. Preprocessing was conducting using FMRIPREP version 1.0.0-rc2 (http://fmriprep.readthedocs.io) on openneuro.org, which is based on Nipype (Gorgolewski et al., 2011). Slice timing correction was disabled due to short TR of the input data. Each T1 weighted volume was corrected for bias field using N4BiasFieldCorrection v2.1.0 (Tustison et al., 2010), skullstripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template), and coregistered to skullstripped ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al.,

2009) using nonlinear transformation implemented in ANTs v2.1.0 (Avants et al., 2008). Cortical surface was estimated using FreeSurfer v6.0.0 (Dale et al., 1999).

Functional data for each run was motion corrected using MCFLIRT v5.0.9 (Jenkinson et al., 2002). Distortion correction for most participants was performed using an implementation of the TOPUP technique (Andersson et al., 2003) using 3dQwarp v16.2.07 distributed as part of AFNI (Cox, 1996). Functional data was coregistered to the corresponding T1 weighted volume using boundary based registration 9 degrees of freedom - implemented in FreeSurfer v6.0.0 (Greve and Fischl, 2009). Motion correcting transformations, T1 weighted transformation and MNI template warp were applied in a single step using antsApplyTransformations v2.1.0 with Lanczos interpolation. Framewise displacement (Power et al., 2014) was calculated for each functional run using Nipype implementation. For more details of the pipeline see http://fmriprep.readthedocs.io/en/1.0.0-rc2/workflows.html. The data were not resampled; voxel size remained at 2 x 2 x 3. For univariate analyses, images were then smoothed with a 6 mm FWHM Gaussian kernel.

General linear model analyses were conducted using SPM (SPM12; Wellcome Trust Centre for Neuroimaging). Prior to modeling, data from the four learning phase blocks was concatenated, using a method developed by T. Sommer and adapted from Schultz et al. (2012). An autoregressive model, AR(1) (adapted to account for the concatenated sessions), was used to model serial autocorrelations in the data. fMRI model regressors were convolved with the canonical hemodynamic response function and entered into a general linear model (GLM) of each participant's fMRI data. Six scan-to-scan motion parameters (x, y, z dimensions as well as roll, pitch, and yaw) produced

during realignment were included as additional regressors in the GLM to account for residual effects of participant movement.

A univariate GLM was constructed to examine BOLD correlates of learning across repetitions and responses to reward feedback. We predicted that brain regions involved in encoding the stimuli and the associations in the maze episode would show a rapid decrease (or decay) in activity across repetitions. To capture this decrease, we created a decaying encoding predictor that decreased exponentially across repetitions. Specifically, a geometric decay of 50 % was applied across the four repetitions, with values for the four repetitions of: 1, 0.50, 0.25, and 0.125. This encoding decay regressor was entered as a separate parametric modulator of activity at states 1, 2, and 3, although our primary prediction was about activity at the start, in state 1. At the reward state, reward and loss feedback were modeled as 1 and -1. We additionally modeled reward feedback during the subsequent rest period to test for lasting effects of reward on BOLD. States 1 and 3 and the feedback state were modeled as a 5 s duration events. State 2 was modeled as a 4 s duration event. The rest period was modeled as a 12 s duration event. The 5 s duration of the State 1 and 3 primary events of interest captures the 4 s presentation of the central stimulus plus 1 additional second where the stimulus had shifted to the left or right door and matches the duration of the modeled feedback state events. Responses to the 8 mazes of interest in addition to the first exposure to maze 9 and 10 were included in the model. (Later repetitions of maze 9 and 10 were excluded because both choices were predetermined to lead to a loss.)

Our navigation learning task, where reward associations remained fixed across repetitions, did not allow for us to examine effects of choice value at state 1 or feedback

due to collinearity with learning and reward variables. As a consequence, we could not examine whether responses to delayed reward feedback were better captured by reward magnitude versus reward prediction error, as suggested by neurophysiological studies in non-human primates (Fiorillo et al., 2008; Kobayashi and Schultz, 2008). Specifically, reward and choice value, the components of reward prediction error, were highly positively correlated (median r = 0.57) and thus we could not include them together in a model at the same time period. At stage 1, decaying encoding across repetitions and choice value were also highly negative correlated (median r = -0.71) and could not be included together in a model. At feedback, while the correlation between decaying encoding and reward was lower than the above correlations (median r = -0.46), we opted to not model decaying encoding at feedback or rest.

Next, we analyzed the localizer phase to identify univariate patterns of activity that discriminated between the stimulus categories. These contrasts were used to identify regions of interest for multivariate analyses. The initial GLM included separate regressors for faces, scenes, objects, scrambled objects, and motion stimuli. Static localizer stimuli were modeled as 2 s events while motion stimuli were modeled as 3 s events. From these first-level results, we computed contrasts with each of the three categories compared versus either of the two alternatives as well as each category versus the combination of both alternatives.

Multivariate fMRI analyses. For multivariate classification analyses, in the learning phase we estimated a mass-univariate GLM using the non-smoothed BOLD data. Each stage as well as the rest period were each modeled with individual regressors, yielding

200 regressors of interest for the learning phase. The learning phase event durations were as above: 5 s for states 1 and 3 and the reward state, 4 s for state 2, and 12 s for the rest period. In the localizer phase we estimated a separate mass-univariate GLM using the non-smoothed BOLD data. Each trial was modeled as a 2 s event with an individual regressor, yielding 155 total regressors. Models included the six motion regressors; for the learning phase, block regressors were included as effects of no interest.

Multivariate analyses were conducting using The Decoding Toolbox (Hebart et al., 2014). Classification utilized a L2-norm learning support vector machine (LIBSVM; Chang and Lin, 2011) with a fixed cost of $c = 1$. The classifier was trained on the full localizer phase data, which was divided into 3 mini-blocks. In the case of imbalanced numbers of face, scene, and object trials per mini-block (due to omitted stimuli not seen during learning) other trials were randomly left out to achieve balance. The trained classifier was then tested on the full learning phase data. Note that for the primary across-phase classification analysis, no cross-validation is necessary for training because no inferences are drawn and no results are reported on the localizer phase data. Based on the strength of discriminability in the localizer phase (using cross-validation), classification of faces was based on the comparison of faces versus scenes. Classification of scenes was based on the comparison of scenes versus faces. Classification of objects was based on the comparison of objects versus faces. While scrambled objects were included in the localizer phase to compare to objects, classification performance for objects versus scrambled objects was worse than that for objects versus faces. For objects in general, across OFC-VMPFC, hippocampal, and

visual ROIs, classification performance was lower than that for faces or scenes. Learning phase classification is reported as area under the curve (AUC) which uses graded decision values and better accounts for biases in classification that may arise due to the different processes engaged by the localizer and learning phases.

The representation of current and future state information was analyzed by applying classifiers trained on the localizer phase to the patterns of activity evoked during maze learning, with a focus on state 1. As described above, one classifier for each category was trained based on the localizer phase data. These trained classifiers were applied to each stage of each maze trial to derive decision values, a parametric variable ranging from 0 to 1 representing how well the pattern of multivoxel activity in regions of interest at each state was described by the classifier.

The classifier decision values for each classifier and for each state were then entered into regression models to measure representations of current and future state. We computed three regression models, one for each category, to examine the information present in classifier decision values at state 1. These models estimated whether decision values were related to state 1 category, state 3 category, the number of correct maze repetitions (derived from behavior), and the two interactions between state 1 and 3 category and the number of correct repetitions. State 1 and state 3 category were represented as binary variables, where 1 indicated that the state matched the category of interest and 0 otherwise. The number of correct repetitions variable was based on learning performance for each maze, focusing on whether the state 3 stimulus of interest had been reached during learning. For the first repetition, the value was set to 0. The value remained at 0 until the repetition following when the participant reached the

correct state 3 and was able to observe the stimulus (and category) present in that state. Subsequently, the value incremented by 1 for every subsequent repetition. The value was incremented for every repetition after the critical future state 3 had been visited, which allowed for this state information to be integrated into decision making at state 1. We made no assumptions about whether future predictions would be stronger or weaker if participants returned again to the incorrect choice. An alternative model where the correct repetition variable was only incremented on correct trials yielded qualitatively similar results.

Within-participant, the results from the above regression models were averaged across the three categories to allow for group-level analyses (via t-tests). This approach was used instead of a single hierarchical mixed-effects model because each trial would be included three times in a single model, violating independence of observations. Participants' data for individual categories were excluded from averaging across models if the classifier failed to generalize across task phases and classify learning phase stimuli. Performance was calculated as the regression coefficient between (concatenated) decision values for state 1 and state 3 and the contrast between the presence of the actual category of interest (e.g. faces) versus the comparison category (e.g. scenes) at state 1 and state 3. If the regression coefficient for states 1 and 3 for a given participant was below zero for all three categories that participant was excluded from the analysis.

Additional analyses examined current, future, and past state representations at different states in the task (state 2, state 3, the feedback state, and post-feedback rest). We did not explore timepoint-by-timepoint modeling of feedback and rest period effects

due to signal-to-noise concerns, as the category decoding of current states was at a relatively low level and the search through many timepoints of data is less constrained than modeling specific states as above.

Control analyses examined different versions of the primary model above. First, we examined four separate models, one for each correct repetition bin, with current state and future state as predictors. This allowed us to examine the pattern of change in current and future state representation over time without the interaction included in the primary model. Second, we examined two alternative models where either current state or future state was omitted from the model. This allowed us to examine whether the inclusion of both current and future state in the model led to potentially spurious results. Note that while the localizer followed learning, any significant effects of learning on future state representation cannot be due to the classifier detecting activation of learned associations at test: any such confounding effects would only affect the main effect of future state representation, not the change of this representation across learning.

Regions of interest. For univariate results, linear contrasts of univariate SPMs were taken to a group-level (random-effects) analysis. We report results corrected for family-wise error (FWE) due to multiple comparisons (Friston et al., 1993). We conduct this correction at the peak level within small volume ROIs for which we had an a priori hypothesis or at the whole-brain cluster level (both after an initial thresholding of $p < 0.005$ uncorrected). For univariate and multivariate analysis, we created two regions of interest in the OFC-VMPFC and hippocampus. We constructed the a priori OFC-VMPFC ROI (x, y, z: 0 44 -14; radius 10mm) based on a previous finding of short-term preceding

state representations in the OFC (Schuck et al., 2016). The bilateral hippocampus ROI was derived from the AAL atlas (Tzourio-Mazoyer et al., 2002). We focused on the hippocampus instead of the combined hippocampus and parahippocampal cortex to limit the size of the ROI for multivariate analyses. As control regions, we also constructed ROIs based on univariate responses to faces, scenes, and objects in the localizer phase. The face ROI was constructed based on the group contrast of faces versus scenes (derived from smoothed data at the first level) at a threshold of $p < 0.001$, selecting clusters surrounding (±42 -42 -22, including $Z < -14$ and $Y < -68$; $Z < 0$ and $Y > -68$). The scene ROI was constructed based on the contrast of scenes versus faces at a threshold of $p < 0.001$, selecting clusters encompassing the bilateral parahippocampal gyrus (including $Z < 4$ and $Y < -68$; $Z < 16$ and $Y > -68$). The object ROI was constructed based on the contrast of objects versus faces in symmetric clusters in the posterior occipital lobe at a threshold of $p < 0.025$ uncorrected (centered at ±58 -58 -10). All voxel locations are reported in MNI coordinates, and results are displayed overlaid on the average of all participants' normalized high-resolution structural images using the xjView toolbox (http://www.alivelearn.net/xjview) and AFNI (Cox, 1996).

Data availability.

Behavioral data will be made publicly available on the Open Science Framework (www.osf.io). Whole-brain fMRI results are available on NeuroVault (https://neurovault.org/collections/4420/) and the full imaging dataset will be publicly available on Openneuro (openneuro.org).

Results

In the navigation learning experiment, participants learned to select the correct choice in an initial state in order to receive a delayed reward (and avoid a loss) after moving through two intervening states (**Figure 1A**). The average duration from start to the end of a maze experience was 48.5 s. Across repetitions, participants showed a rapid increase in mean performance, rising to 76 % correct at the second exposure (CI [70.0 83.6]; $t_{(34)} = 22.81$, $p < 0.0001$, t-test versus chance; **Figure 1B**). Initial feedback also exerted an effect on learning: if a reward was received on the first exposure, performance on the next repetition was significantly higher than performance following an initial loss (repetition 2, initial reward, 82.1 % CI [76.0 88.2]; $p < 0.0001$; repetition 2, initial loss, 71.4 % (CI [62.1 80.8]; $p < 0.0001$; $t_{(34)} = 2.77$, CI [2.8 18.6]; $p = 0.009$). Supporting the involvement of multiple cognitive systems in learning, we found that individual differences in working memory capacity as measured by the OSPAN task significantly correlated with mean accuracy over the learning phase ($r = 0.39$, $p = 0.026$; OSPAN performance mean 38.3, range 0-86; arithmetic component performance mean 89.2%). While episodic memory and working memory are often viewed as a distinct functions, the hippocampus is also known to be involved in critical aspects of working memory maintenance (Olsen et al., 2012).

Our fMRI analysis first examined univariate learning-related brain responses. At the first state in a maze, state 1, we expected to find an initially high and then (exponentially) decreasing response across repetitions that might be related to learning in regions supporting memory and associative encoding. We indeed found a significant decreasing effect of exposure number (repetition) on BOLD activity in a cluster centered

in the right hippocampus (z = 4.42; p = 0.014 SVC; 18 -10 -21; **Figure 2A, Figure S1**) as well as right parahippocampal gyrus (p<0.05 whole-brain FWE-corrected; **Table S1**). This effect was relatively selective, with only three other above-threshold clusters of activity in this contrast: two in the occipital lobe and one in the left dorsal premotor cortex (whole-brain FWE-corrected; **Table S1**). The pattern of decreasing activity across repetitions was also selective to state 1; we found only a single significant cluster in the precuneus at state 2 and no significant clusters of activity in state 3 (**Table S1**). Overall, the pattern of initially high and then decreasing activity over time in the hippocampus and parahippocampal cortex supports a role for the hippocampus and broader MTL in rapid episodic or relational learning.

Next, to confirm expected responses to relatively delayed reward receipt, we examined responses to feedback. We found that reward versus loss feedback, arriving ~ 40 s after the state 1 choice for each maze, was correlated with BOLD activity in a cluster including the ventral striatum and OFC-VMPFC, as expected (**Figure 2B; Figure S2**). We also found reward-related activation in the left hippocampus (z = 3.94; p = 0.028 SVC; -40 -22 -12; this effect was also part of a larger significant cluster in the whole-brain analysis), in line with predictions from research on fMRI responses to relatively delayed feedback (~ 7 s; Foerde and Shohamy, 2011). Interestingly, we found that neural processing related to the effect of a reward versus loss persisted through the 16 s rest period following feedback. During rest, feedback continued to differentially affect activity in a cluster including the OFC-VMPFC and ventral striatum, as well as a region of the right posterior hippocampus (OFC-VMPFC and ventral striatum, p < 0.05 whole-brain

FWE-corrected; right hippocampus z = 4.75; p = 0.044 SVC; 32 -34 -3; **Figure S2, Table S1**).

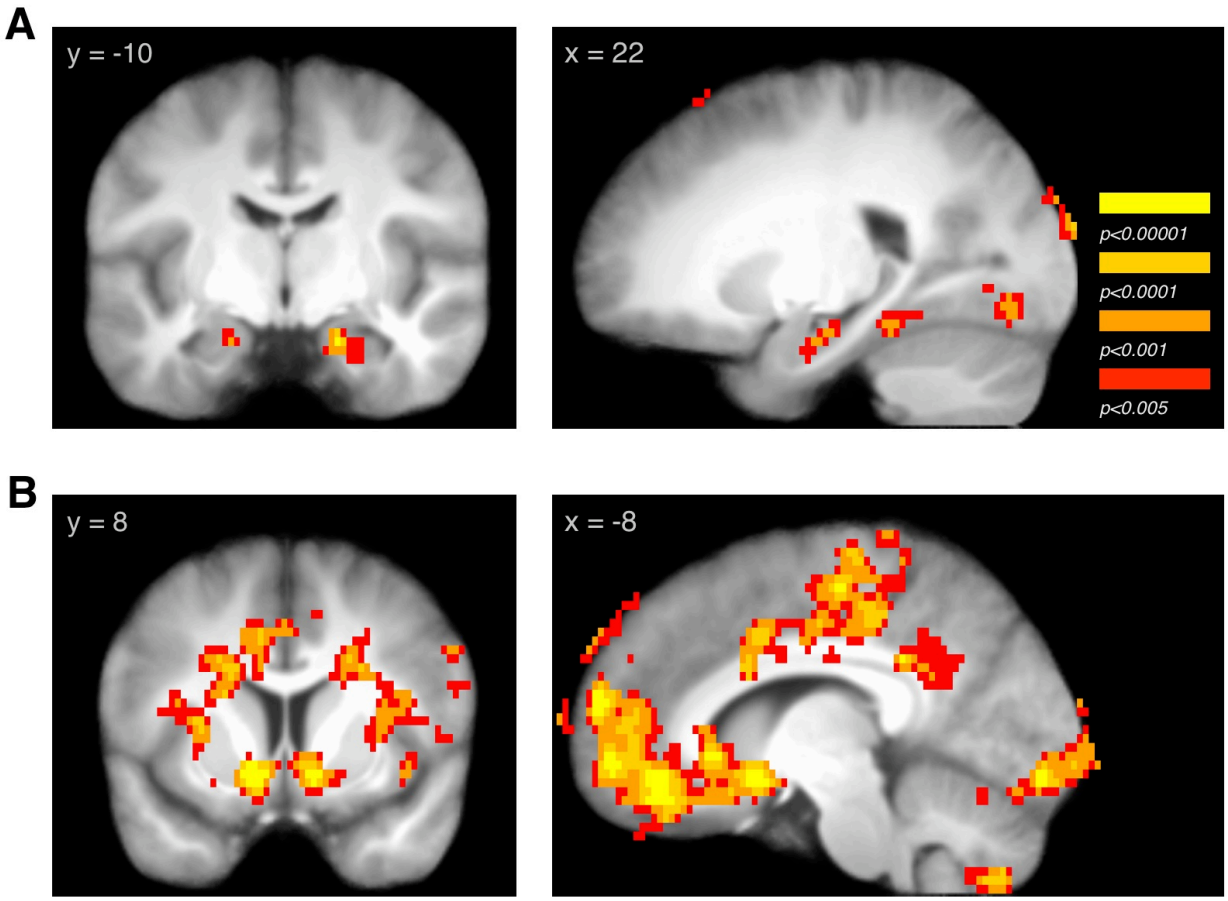

Figure 2. Learning-related univariate responses. (A) Across repetition of mazes during the learning phase, activity in the right hippocampus / MTL significantly decreased, supporting a role for the hippocampus in maze encoding (right hippocampus, $p < 0.05$ SVC; image threshold $p < 0.005$ unc.) (B) At feedback, beginning > 40 s after the start of a maze, activation in the ventral striatum and OFC - ventromedial PFC, among other regions, was significantly reward versus loss. (Images $p < 0.05$ whole-brain FWE-corrected.)

We next turned to our primary question of whether learning mechanisms in the OFC-VMPFC and hippocampus acquire predictive information about future states. We utilized multivariate methods, with the OFC region of interest defined based on Schuck et al. (2016) and the hippocampal ROI based on an anatomical atlas. As a control, we also analyzed responses in visual category-selective regions in the occipital and ventral temporal cortex. We first verified that a classifier trained on multivoxel patterns of activity from the post-learning localizer phase was able to generalize to accurately classify stimuli at state 1 and state 3 during the learning phase (state 2 did not contain any image stimuli; **Figure 1A**). Averaging across the three categories (faces, scenes, and objects), we found significant classification in all regions of interest (OFC-VMPFC mean AUC 9.88 CI [6.45 13.31]; $t_{(30)} = 5.88$, $p < 0.001$; hippocampus 10.63 CI [7.68 13.57]; $t_{(30)} = 7.37$, $p < 0.001$; visual ROIs 158.42 CI [47.61 48.85]; $t_{(30)} = 158.42$, $p < 0.001$).

We then used the trained classifier to test for evidence of prospective future state representations at the beginning of each maze, ~ 30 s before the expected future state became visible. Our method tests how well patterns of activity at state 1 were related to the actual current stimulus category on the screen as well as the future state 3 stimulus category. The future state 3 information would be initially experienced upon the first rewarded repetition of a maze. In our analyses, we refer to this effect as ‘correct repetitions’; this variable is coded from 0 (no experience with the future state) to 3 (maximum possible experiences with the future state). Using raw classifier decision values derived from the neural pattern at state 1, which reflect the strength of evidence for each tested category, we analyzed whether decision values were explained by the

current state (state 1) and future state (state 3) and whether this changed with learning across repetitions (see **Methods**).

As expected, in the OFC-VMPFC distributed patterns of activity accurately discriminated the stimulus visible at the current state (including all participants; 0.092 CI [0.050 0.134]; $t_{(30)} = 4.48$, $p < 0.001$; **Figure 3B**). We also found an effect of future state overall (0.052 CI [0.001 0.103]; $t_{(28)} = 2.08$, $p = 0.047$; in this and following results, analyses exclude participants with poor current state classification performance) and a negative effect of correct repetition (-0.016 CI [-0.029 -0.003]; $t_{(28)} = -2.48$, $p < 0.02$). Critically, we found a significant increase in the representation of the future state across repetitions (interaction between future state and correct repetition 0.051 CI [0.01 0.093]; $t_{(28)} = 2.53$, $p = 0.017$; **Figure 3B**).

This effect of future state representation was selective, as the interaction between current state and correct repetition was near zero (0.007 CI [-0.041 0.055]; $t_{(28)} = 0.30$, $p = 0.77$). However, this effect was not statistically equivalent to a null effect, as indicated by an equivalence test using the TOST procedure (Lakens, 2017): the effect was not significantly within the bounds of a medium effect of interest (Cohen's $d = \pm 0.55$, providing 80% power with 29 participants in the current analysis; $t_{(28)} = -1.34$, $p = 0.095$), so we cannot reject the presence of a medium-size effect. In a control analysis modeling each correct repetition bin separately, we found no pattern of change over time in state 1 representation in the OFC-VMPFC (**Figure S3**) but an increase in future state 3 information over time (**Figure 3C**) such that the strength of representation of the current visible state and the state anticipated > 30 s in the future was no longer significantly different by the last learning repetition ($t_{(28)} = 1.91$; CI [-0.005 0.152]; $p = 0.066$). In two

additional control models, we removed the current state or the future state from the model and found similar results for the included variables as in the full model above (**Figure S3**).

In the hippocampus, in contrast, while we observed evidence for representing the current state (0.0455 CI [0.022 0.069]; $t_{(30)} = 4.00$, $p < 0.001$), we did not find any evidence for increasing represention of future states (interaction 0.004 CI [-0.018 0.026]; $t_{(28)} = 0.35$, $p = 0.73$, **Figure S4**), although this effect was not statistically equivalent to a null effect (TOST $t_{(28)} = -1.06$, $p = 0.15$). Additionally, we found that the future state interaction effect in the hippocampus was significantly weaker than the effect in the OFC-VMPFC ($t_{(26)} = 2.16$; CI [0.002 0.094]; $p = 0.040$). In our control regions in the posterior and ventral occipital cortex that responded to the categories of visual stimuli, while we observed significant representations of the current state (0.318 CI [0.292 0.345]; $t_{(30)} = 24.62$, $p < 0.001$), we did not find evidence for increasing representation of future states (interaction 0.011 CI [-0.009 0.031]; $t_{(30)} = 1.14$, $p = 0.26$; **Figure S4**), and this effect was statistically equivalent to a null effect ($t_{(30)} = -3.08$, $p = 0.005$). The future state interaction effect in the visual cortex was also significantly weaker than the effect in the OFC-VMPFC ($t_{(28)} = 2.76$; CI [0.014 0.098]; $p = 0.01$).

Later in the maze, we did not find any evidence for the representation of preceding states (information reflecting state 1 during state 3 experience) in the OFC-VMPFC or other regions (**Figure S6**). In an exploratory analysis of activity during the other states in a maze, we found a positive but non-significant change in the representation of the future state 3 across learning in the OFC-VMPFC during state 2 (interaction 0.040 CI [-0.002 0.082]; $t_{(28)} = 1.95$, $p = 0.062$), an effect in the same

direction as the state 1 interaction. In the hippocampus, during the rest period following the maze, we found an increase in the representation of state 3 stimuli across learning (interaction between state 3 state and correct repetition 0.015 CI [0.003 0.027]; $t_{(28)}$ =2.74, p = 0.011).

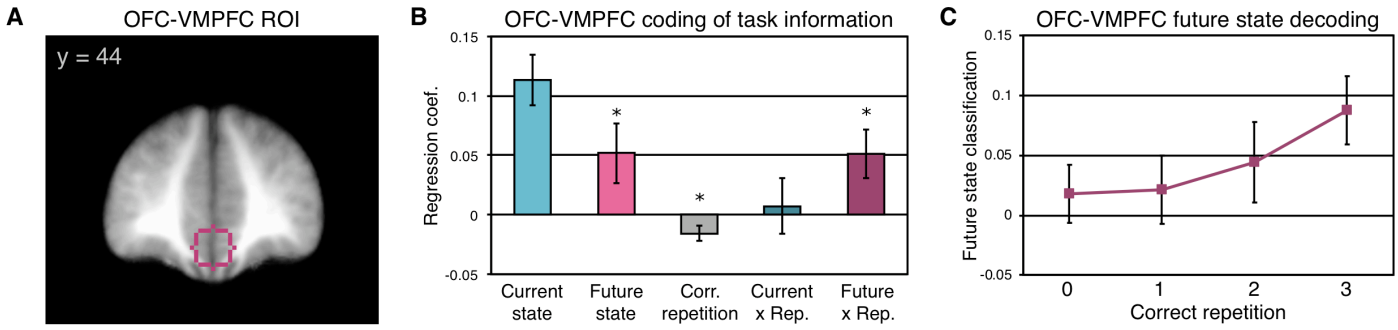

Figure 3. OFC-VMPFC representation of anticipated future state. (A) Outline of OFC-VMPFC region of interest shown over the across-participant average anatomical MRI. (B) Regression coefficients for decoded of activity in the OFC-VMPFC for the current state (state 1), future state (state 3), the repetition since performance was correct, and the interaction of current state and future state with correct repetition. (C) For illustration, future state classification is plotted separately for each correct maze repetition, derived from control models examining current and future state separately for each correct repetition bin. (* p < 0.05; **p < 0.01; *** p < 0.001; Statistical comparisons were not completed on the current state representation in (B) because the plotted data represents results after exclusion of participants with poor current state decoding.)

In a final analysis, we examined the relationship between activity in the OFC-VMPFC and the hippocampus to explore whether these learning-related regions may interact. In a PPI analysis, we found robust positive connectivity between these two regions across different maze states, including state 1, the feedback state, and rest (p <

0.0001 SVC; **Figure S9**). In a conjunction of results across these three states, in addition to the hippocampus, activity in only a few clusters including the posterior cingulate and dorsomedial PFC remained at stringent uncorrected thresholds. However, we found no relationship between OFC-VMPFC - hippocampal connectivity and behavioral or with neural measures of learning and predictive representations.

Discussion

Learning to predict the future is a critical aspect of goal-directed decision making. While previous research has focused on the relatively short timescale prediction of sensory and motor events (Rao and Ballard, 1999; den Ouden et al., 2012; Summerfield and de Lange, 2014), we do not yet know how the human brain acquires predictive representations of future states during goal-directed learning. In a simple learning task with extended sequences of experience, we found that participants learned over only a few repetitions to make correct choices. Using fMRI, we found that this rapid learning across repetitions – each separated by minutes – was related to initially high and then decreasing activity in the right hippocampus and parahippocampal cortex. Critically, as learning developed, we also found an increasing representation of to-be-visited future states in multivariate patterns of activity the OFC-VMPFC, but not in the hippocampus. By the last learning repetition, information representing a hidden state ~ 30 s in the future was of similar strength as the state currently visible on the screen. Our results indicate that a learning mechanism in the OFC-VMPFC may enable the human brain to represent the distant and unseen future consequences of goal-directed actions.

Our finding that the human OFC-VMPFC acquires distal future state representations provides novel support for the proposal that the OFC represents a cognitive map of the environment (Wilson et al., 2014; Stalnaker et al., 2015; Schuck et al., 2016; Schuck et al., 2018). Computationally, the representation of future state information in the OFC-VMPFC lends support to reinforcement learning models that explicitly represent state-to-state associations, such as model-based or successor representation models (Daw et al., 2005; Stachenfeld et al., 2017; Gershman, 2018) over

additions to model-free temporal difference algorithms such as eligibility traces (Sutton and Barto, 1998). However, we cannot rule out a contribution of a cooperative model-free eligibility trace learning mechanism that could assist in state and action value learning by maintaining preceding state information until feedback (Gmaz et al., 2018; Maggi et al., 2018).

Previous experimental work in humans has shown that recent past (hidden) state information that is necessary for task performance is represented in the medial OFC (Schuck et al., 2016). In contexts that allow for predictions, other work has also shown that the OFC may represent information about the identity of immediately upcoming rewards (Howard et al., 2015; Howard and Kahnt, 2018). Our results indicate that the medial OFC / VMPFC acquires predictive information over time, allowing the OFC to represent critical future state information for goal-directed decision making. The > 40 s duration of mazes in our experiment, while relatively short in comparison to many experiences outside the lab, also goes well beyond the timescale of common human fMRI studies of decision making. Other studies have found evidence for visual cortex reactivation for relatively short timescale predictive associations (Wimmer and Shohamy, 2012; Doll et al., 2015; Ekman et al., 2017), including during reward-based decision making (Doll et al., 2015). However, when state predictions extend beyond several seconds, predictive knowledge must be supported by a learning system outside sensory cortex (Summerfield and de Lange, 2014). Our results indicate that a mechanism in the OFC-VMPFC is capable of fulfilling this role.

In the hippocampus, we found that initially high activity rapidly decreased across learning repetitions. This pattern is consistent with a rapid episodic encoding

mechanism, where activity decreases as repetitions contain less new information (Kaplan et al., 2014) and consistent with the role of the hippocampus in relational encoding (Davachi, 2006). Notably, in the current experiment, repetitions of mazes were separated by at least 4 min, a delay that likely requires a contribution of longer-term memory for successful learning. While the vast majority of previous human studies have examined memory for brief episodes (e.g. static pictures), our results extend what is known about learning-related hippocampal effects by showing that hippocampal activity selectively at the beginning of goal-directed episodes decreases across learning. This finding, however, is in the opposite direction of a previous report which found increases in event-onset hippocampal activity over repeated viewing of passive movies (Ben-Yakov et al., 2014). These different effects could be related to the goal-directed learning component of the current study, where participants are incentivized to predict future consequences. Behaviorally, the rapid learning we observed after even a single repetition fits well with recent research showing that humans can make value-based decisions based on information provided by specific past experiences (Gluth et al., 2015; Wimmer and Buechel, 2016; Bornstein et al., 2017). For goal-directed decision making, this rapid encoding of maze experiences may work together with other learning mechanisms, potentially in the OFC, to allow us to seek out positive or avoid negative future outcomes.

While we found that the OFC-VMPFC represents distal future states, we did not find evidence of future state representations in the hippocampus. Indeed, we found that the future state representation effect in the OFC-VMPFC was significantly greater than that in the hippocampus. Such a null effect is surprising, as research in rodents has

demonstrated that the hippocampus represents future paths (Johnson and Redish, 2007; Pfeiffer and Foster, 2013). Previous research in humans has found that hippocampal activity during reward learning co-varies with the activation of associated items in visual cortex (Wimmer and Shohamy, 2012) and that hippocampal activity patterns represent retrieved information about currently cued target locations (Brown et al., 2016). During rest periods following learning, recent papers have reported reactivation of hippocampal activity patterns (Schapiro et al., 2018; Schuck and Niv, 2018) and that hippocampal activity co-varies with the reactivation of visual stimuli (Momennejad et al., 2017). While exploratory, we did find evidence for hippocampal representation of past states during post-maze rest (**Figure S8**) and we also identified significant functional connectivity between the OFC and hippocampus during learning. Interestingly, in recent work using MEG, an imaging method with much faster temporal resolution, Kurth-Nelson et al. (2016) reported that fast sequences of activity reflected potential trajectories in a non-spatial environment. While no correlations between MEG activity and behavior were identified in that study, such rapid sequence-related activity, likely decoded primarily from visual cortex, could be driven by the hippocampus in humans, similar to findings in rodents (Pfeiffer and Foster, 2013; Olafsdottir et al., 2018).

The brain has been proposed to be a predictive machine (Rao and Ballard, 1999; Friston, 2005; Clark, 2013). Our results provide new support for how the brain is able to make relatively long-timescale predictions that are critical for goal-directed decision making. We find that as models of the environment are initially established, a learning mechanism in the OFC-VMPFC is capable of making long timescale predictions about future states. At the same time, overall activity in the hippocampus correlated with maze

exposure across learning. Our complementary results in the OFC-VMPFC and hippocampus suggest that these regions may interact to support learning and decision making in goal-directed tasks (Wikenheiser and Schoenbaum, 2016; Wikenheiser et al., 2017). Further understanding of these mechanisms and their interaction will be for important for understanding deficits in both learning the structure of the environment and in using these representations for goal-directed decisions that are found in various psychiatric disorders and addiction.

References

Andersson JL, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20:870-888.

Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12:26-41.

Behrens T, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018) What is a cognitive map? Organizing knowledge for flexible behavior. bioRxiv.

Ben-Yakov A, Rubinson M, Dudai Y (2014) Shifting gears in hippocampus: temporal dissociation between familiarity and novelty signatures in a single event. J Neurosci 34:12973-12981.

Bornstein AM, Khaw MW, Shohamy D, Daw ND (2017) Reminders of past choices bias decisions for reward in humans. Nat Commun 8:15958.

Bradfield LA, Dezfouli A, van Holstein M, Chieng B, Balleine BW (2015) Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. Neuron 88:1268-1280.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433-436.

Brown TI, Carr VA, LaRocque KF, Favila SE, Gordon AM, Bowles B, Bailenson JN, Wagner AD (2016) Prospective representation of navigational goals in the human hippocampus. Science 352:1323-1326.

Chang CC, Lin C (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:1-27.

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181-204.

Collins AG, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35:1024-1035.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162-173.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179-194.

Davachi L (2006) Item, context and relational episodic encoding in humans. Curr Opin Neurobiol 16:693-700.

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

Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage 19:430-441.

den Ouden HE, Kok P, de Lange FP (2012) How prediction errors shape perception, attention, and motivation. Frontiers in psychology 3:548.

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015) Model-based choices involve prospective neural activity. Nat Neurosci 18:767-772.

Eichenbaum H, Cohen NJ (2001) From Conditioning to Conscious Recollection: Memory Systems of the Brain. New York: Oxford University Press.

Ekman M, Kok P, de Lange FP (2017) Time-compressed preplay of anticipated events in human primary visual cortex. Nat Commun 8:15276.

Feinberg DA, Moeller S, Smith SM, Auerbach E, Ramanna S, Gunther M, Glasser MF, Miller KL, Ugurbil K, Yacoub E (2010) Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. PLoS One 5:e15710.

Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. Nat Neurosci 11:966-973.

Foerde K, Shohamy D (2011) Feedback timing modulates brain systems for learning in humans. J Neurosci 31:13157-13167.

Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins DL (2009) Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood. Neuroimage 47.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815-836.

Friston KJ, Worsley KJ, Frackowiak SJ, Mazziotta JC, Evans AC (1993) Assessing the significance of focal activations using their spatial extent. Human Brain Mapping 1:210-220.

Gershman SJ (2018) The Successor Representation: Its Computational Logic and Neural Substrates. J Neurosci 38:7193-7200.

Gluth S, Sommer T, Rieskamp J, Buchel C (2015) Effective Connectivity between Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from Memory. Neuron 86:1078-1090.

Gmaz JM, Carmichael JE, van der Meer MA (2018) Persistent coding of outcome-predictive cue features in the rat nucleus accumbens. Elife 7.

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS (2011) Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Frontiers in neuroinformatics 5:13.

Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63-72.

Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A (2002) Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 47:1202-1210.

Hebart MN, Gorgen K, Haynes JD (2014) The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Frontiers in neuroinformatics 8:88.

Howard JD, Kahnt T (2018) Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex. Nat Commun 9:1611.

Howard JD, Gottfried JA, Tobler PN, Kahnt T (2015) Identity-specific coding of future rewards in the human orbitofrontal cortex. Proc Natl Acad Sci U S A 112:5195-5200.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825-841.

Johnson A, Redish AD (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J Neurosci 27:12176-12189.

Kaplan R, Horner AJ, Bandettini PA, Doeller CF, Burgess N (2014) Human hippocampal processing of environmental novelty during spatial navigation. Hippocampus 24:740-750.

Kobayashi S, Schultz W (2008) Influence of reward delays on responses of dopamine neurons. J Neurosci 28:7837-7846.

Kurth-Nelson Z, Economides M, Dolan RJ, Dayan P (2016) Fast Sequences of Non-spatial State Representations in Humans. Neuron 91:194-204.

Lakens D (2017) Equivalence tests: A practical primer for t-tests, correlations, and metaanalyses. Soc Psychol Personal Sci.

Lewandowsky S, Oberauer K, Yang LX, Ecker UK (2010) A working memory test battery for MATLAB. Behav Res Methods 42:571-585.

Maggi S, Peyrache A, Humphries MD (2018) An ensemble code in medial prefrontal cortex links prior events to outcomes during learning. Nat Commun 9:2204.

Moeller S, Yacoub E, Olman CA, Auerbach E, Strupp J, Harel N, Ugurbil K (2010) Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn Reson Med 63:1144-1153.

Momennejad I, Otto AR, Daw N, Norman KA (2017) Offline Replay Supports Planning: fMRI Evidence from Reward Revaluation. bioRxiv.

O'Keefe J, Nadel L (1978) The Hippocampus as a Cognitive Map. New York: Oxford University Press.

Olafsdottir HF, Bush D, Barry C (2018) The Role of Hippocampal Replay in Memory and Planning. Current biology : CB 28:R37-R50.

Olsen RK, Moses SN, Riggs L, Ryan JD (2012) The hippocampus supports multiple cognitive processes through relational binding and comparison. Front Hum Neurosci 6:146.

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941-20946.

Pfeiffer BE, Foster DJ (2013) Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497:74-79.

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE (2014) Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84:320-341.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79-87.

Schapiro AC, McDevitt EA, Rogers TT, Mednick SC, Norman KA (2018) Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. Nat Commun 9:3920.

Schuck NW, Niv Y (2018) Sequential replay of non-spatial task states in the human hippocampus. bioRxiv.

Schuck NW, Wilson RC, Niv Y (2018) A state representation for reinforcement learning and decision-making in the orbitofrontal cortex. In: Goal-Directed Decision Making: Computations and Neural Circuits (Morris RW, Bornstein AM, Shenhav A, eds), pp 259-278. London: Academic Press.

Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron 91:1402-1412.

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15:657-680.

Schultz H, Sommer T, Peters J (2012) Direct evidence for domain-sensitive functional subregions in human entorhinal cortex. J Neurosci 32:4716-4723.

Setsompop K, Gagoski BA, Polimeni JR, Witzel T, Wedeen VJ, Wald LL (2012) Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. Magn Reson Med 67:1210-1224.

Stachenfeld KL, Botvinick MM, Gershman SJ (2017) The hippocampus as a predictive map. Nat Neurosci 20:1643-1653.

Stalnaker TA, Cooch NK, Schoenbaum G (2015) What the orbitofrontal cortex does not do. Nat Neurosci 18:620-627.

Summerfield C, de Lange FP (2014) Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci 15:745-756.

Sutton RS, Barto AG (1998) Reinforcement Learning: an Introduction. Cambridge: MIT Press.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310-1320.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273-289.

Wikenheiser AM, Schoenbaum G (2016) Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. Nat Rev Neurosci 17:513-523.

Wikenheiser AM, Marrero-Garcia Y, Schoenbaum G (2017) Suppression of Ventral Hippocampal Output Impairs Integrated Orbitofrontal Encoding of Task Structure. Neuron 95:1197-1207 e1193.

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014) Orbitofrontal cortex as a cognitive map of task space. Neuron 81:267-279.

Wimmer GE, Shohamy D (2012) Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338:270-273.

Wimmer GE, Buechel C (2016) Reactivation of reward-related patterns from single past episodes supports memory-based decision making. J Neurosci 36:2868-2880.

Wimmer GE, Poldrack R (2018) Reinforcement learning over weeks versus minutes changes neural representations of value and leads to longer-lasting associations. biorXiv.

Wimmer GE, Braun EK, Daw ND, Shohamy D (2014) Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. J Neurosci 34:14901-14912.

Wimmer GE, Li JK, Gorgolewski KJ, Poldrack RA (2018) Reward Learning over Weeks Versus Minutes Increases the Neural Representation of Value in the Human Brain. J Neurosci 38:7649-7666.

Xu J, Moeller S, Auerbach EJ, Strupp J, Smith SM, Feinberg DA, Yacoub E, Ugurbil K (2013) Evaluation of slice accelerations using multiband echo planar imaging at 3 T. Neuroimage 83:991-1001.

Supplementary Information

Figure S1. Illustration of hippocampal activity at state 1 decrease across repetitions, related to **Figure 2**. Sequence onset is at time 0; expected peak of BOLD response to state 1 is between 6.2 – 8.7 s. Repetition 1 in darkest aqua; repetition 4 in lightest aqua.

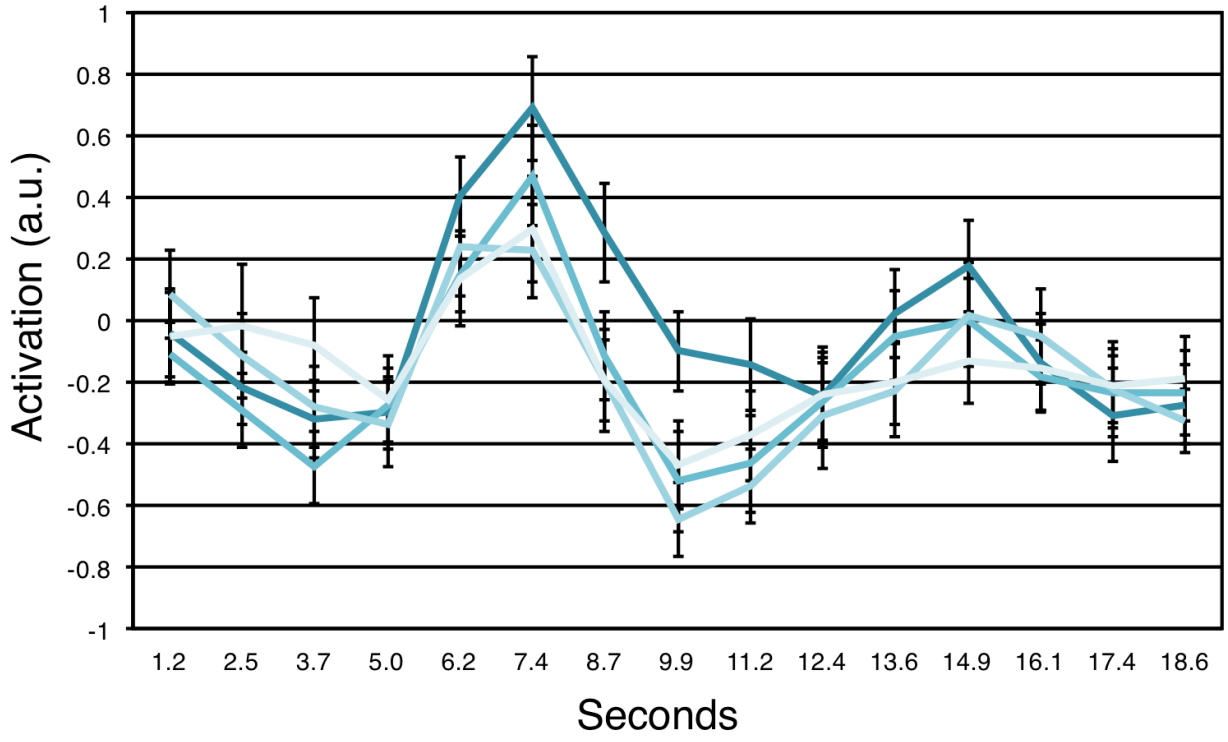

Figure S2. Reward-related univariate responses at feedback and during post-feedback rest. (A) Reward versus loss feedback at the reward stage, including left hippocampus (top), and overlaid with the OFC-VMPFC ROI (bottom) (B) Reward versus loss feedback during the post-feedback rest period, including overlap with the OFC-VMPFC ROI (bottom). (Images $p < 0.05$ whole-brain FWE-corrected.)

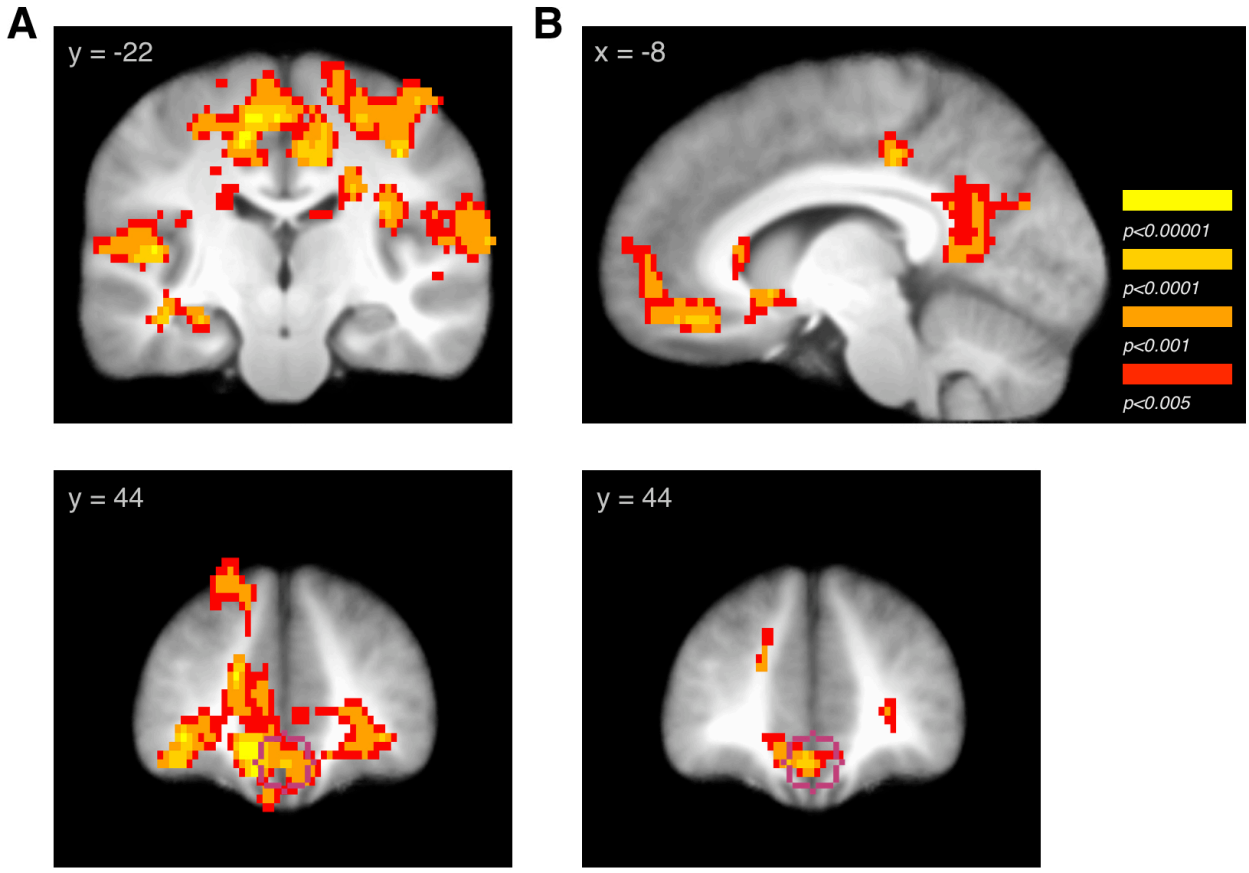

Figure S3. State representation in the OFC-VMPFC. (A) Effects of current and future state modeled separately for each bin of correct repetition, as in **Figure 3C**. Current state in aqua; future state in magenta. (B) Effect of current state, correct repetition, and interaction of current state and correct repetition in a control model omitting future state. (C) Effect of future state, correct repetition, and interaction of future state and correct repetition in a control model omitting current state. (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$)

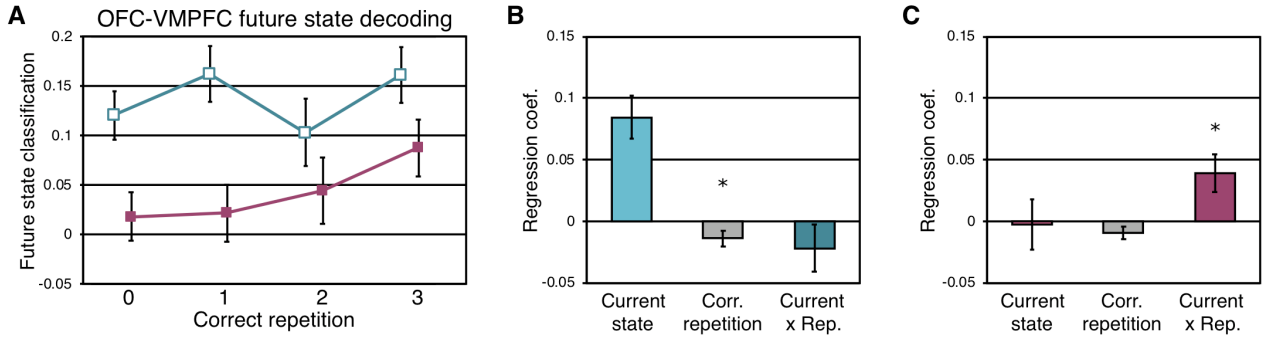

Figure S4. Current and future state representation at state 1 in the hippocampus and in visual control regions, as in Figure 3. Bottom: results from an analysis conducted separately for each correct repetition bin. Current state in aqua; future state in magenta. (* p < 0.05; ** p < 0.01; *** p < 0.001; statistical comparisons for the hippocampus were not completed on the current state representation because the plotted data represents results after exclusion of participants with poor current state decoding.)

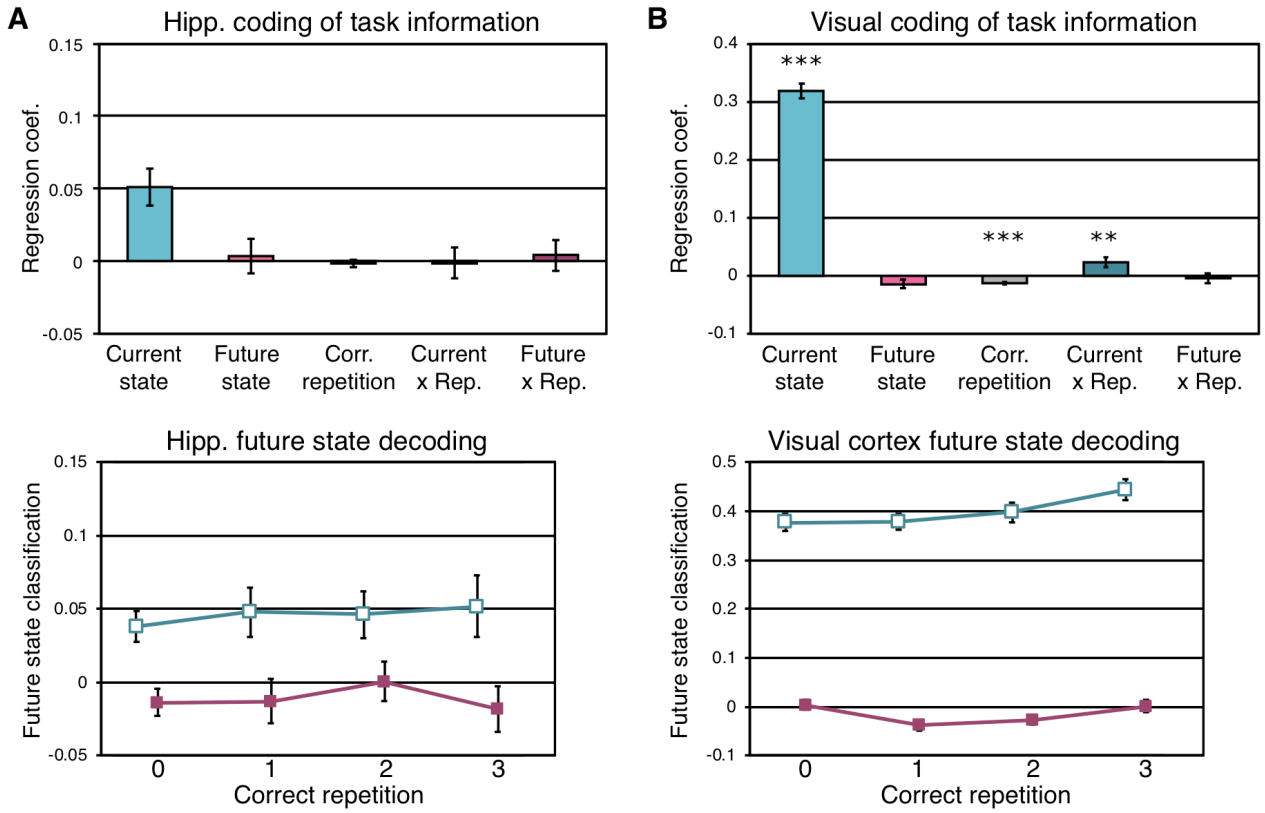

Figure S5. State 2: State representation at state 2 in the OFC-VMPFC, hippocampus, and visual control regions. (* p < 0.05; ** p < 0.01; *** p < 0.001)

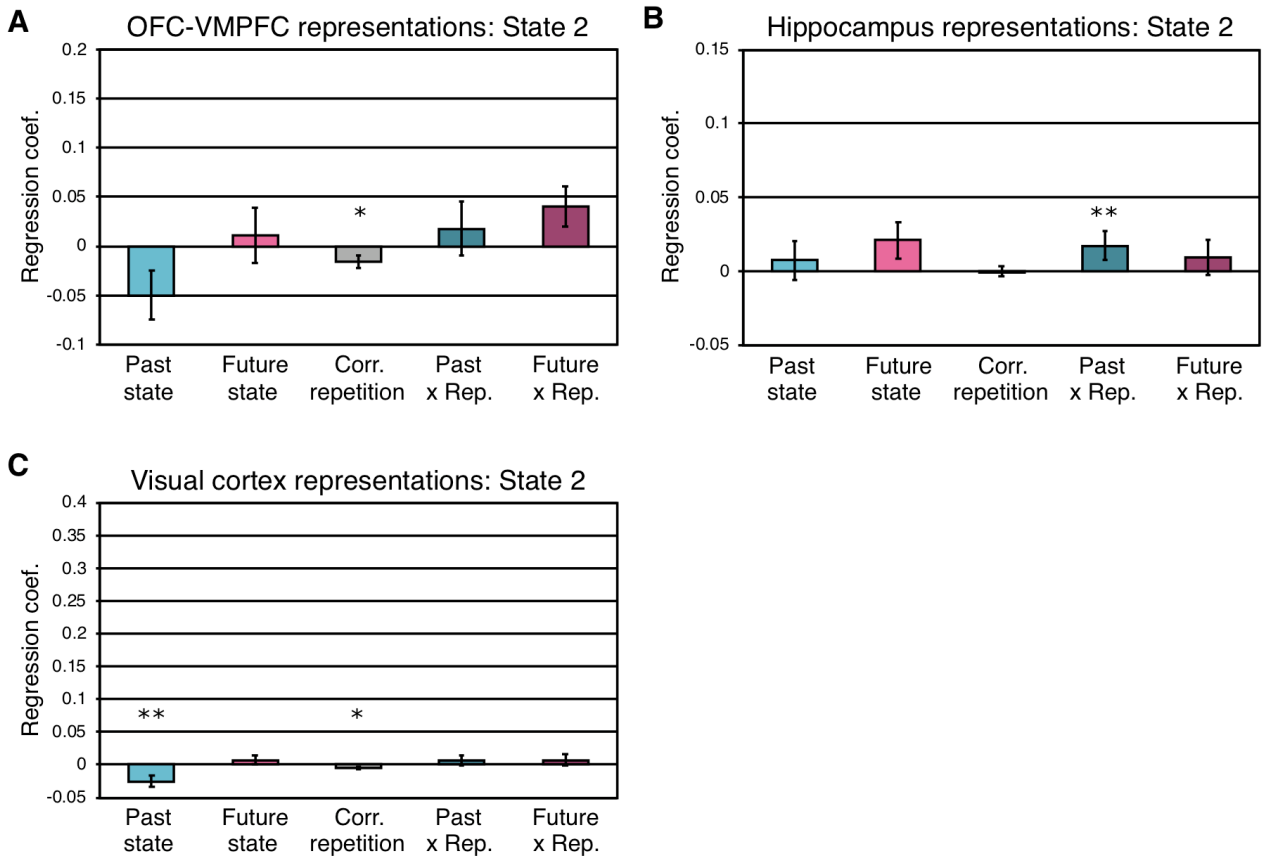

Figure S6. State 3: State representation at state 3 in the OFC-VMPFC, hippocampus, and visual control regions. (* p < 0.05; ** p < 0.01; *** p < 0.001; statistical comparisons for OFC-VMPFC and hippocampus were not completed on the current state representation because the plotted data represents results after exclusion of participants with poor current state decoding.)

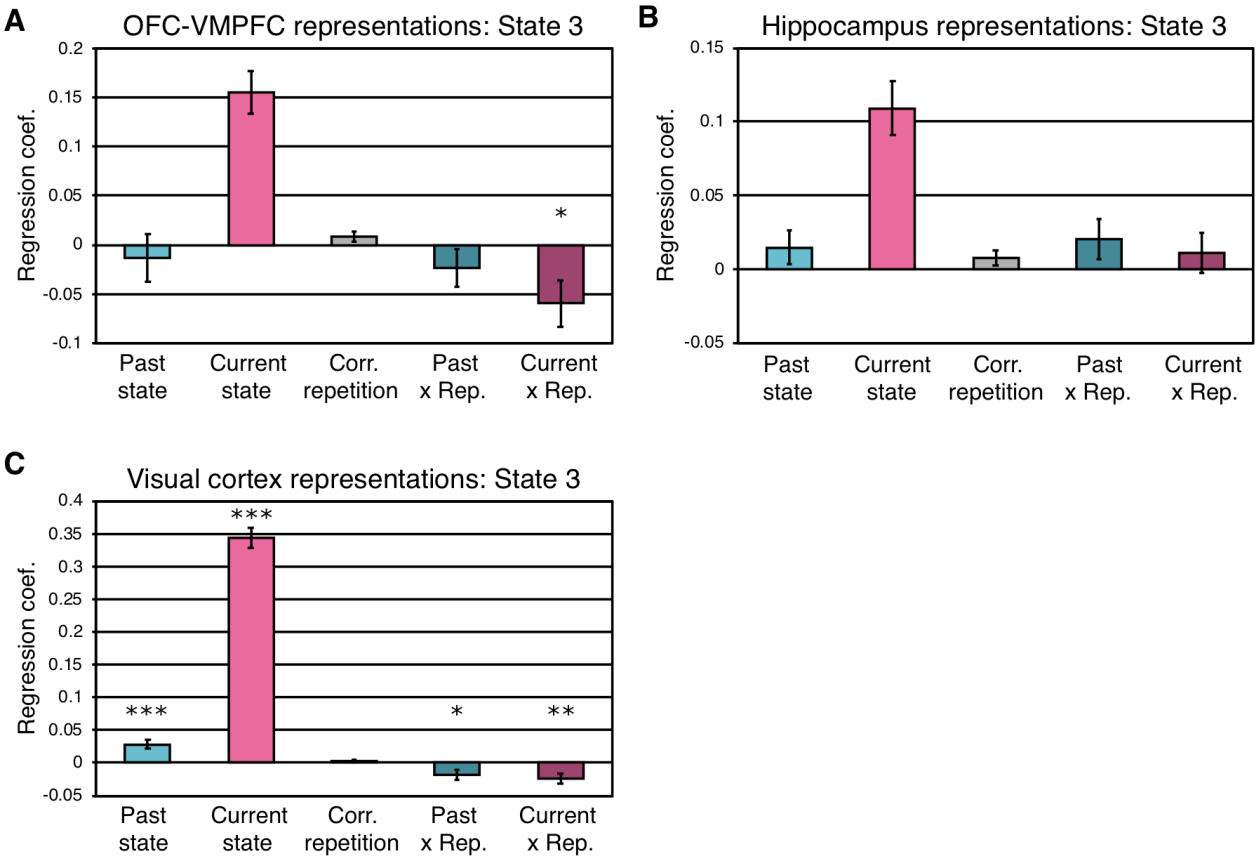

Figure S7. Feedback state: State representation at the feedback state in the OFC-VMPFC, hippocampus, and visual control regions. (* p < 0.05; ** p < 0.01; *** p < 0.001)

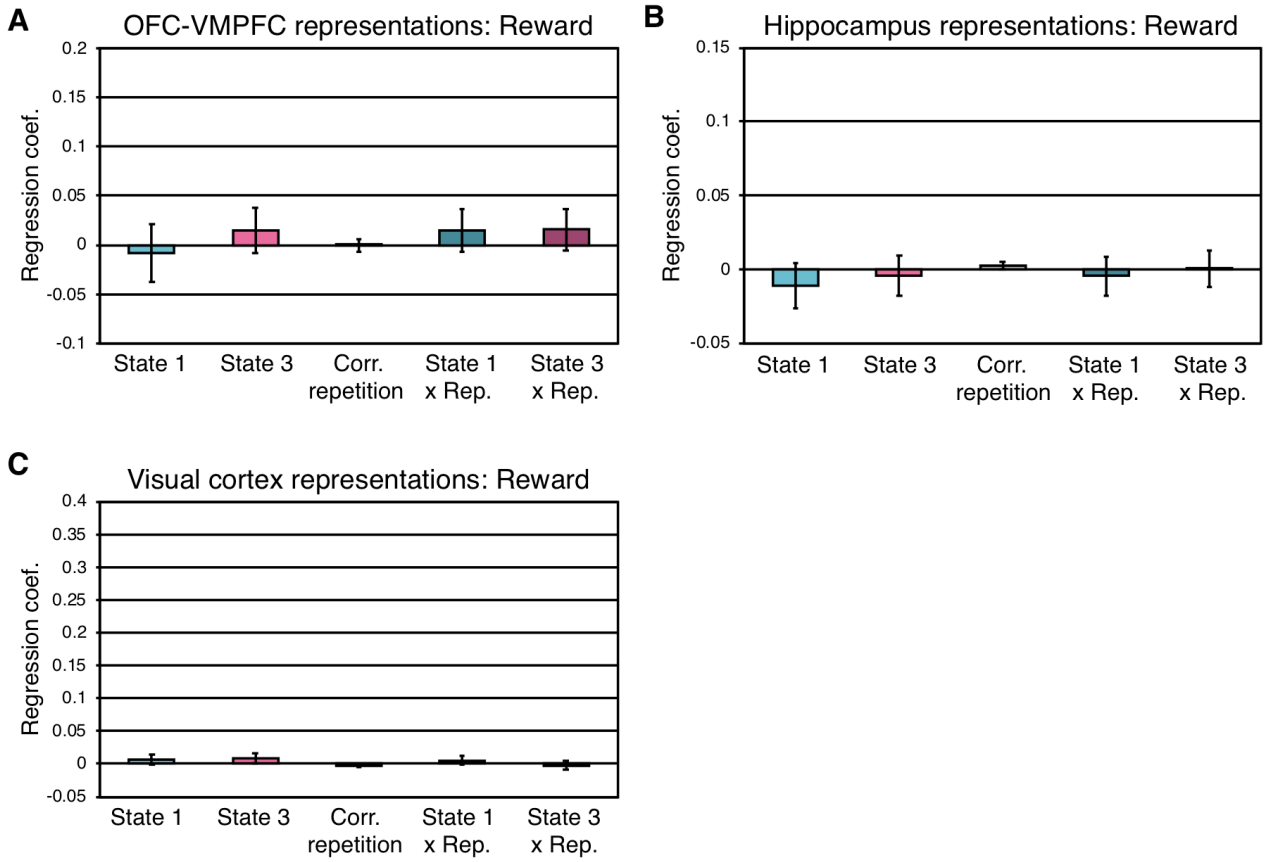

Figure S8. Rest: State representation during the rest state in the OFC-VMPFC, hippocampus, and visual control regions. (* p < 0.05; ** p < 0.01; *** p < 0.001)

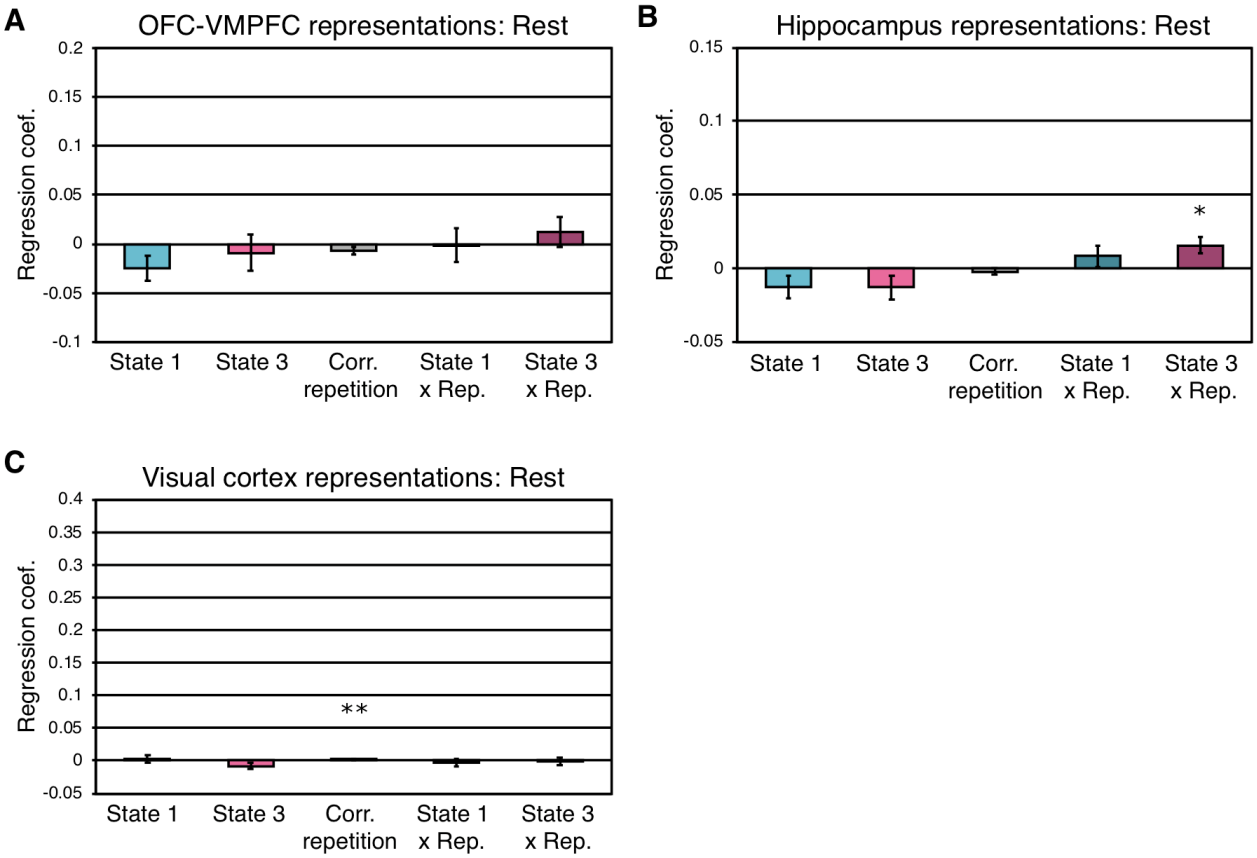

Figure S9. Connectivity (PPI) between the OFC-VMPFC, conjunction of effects at state 1, feedback state, and rest (individual maps thresholded at $p < 0.0001$ unc.).

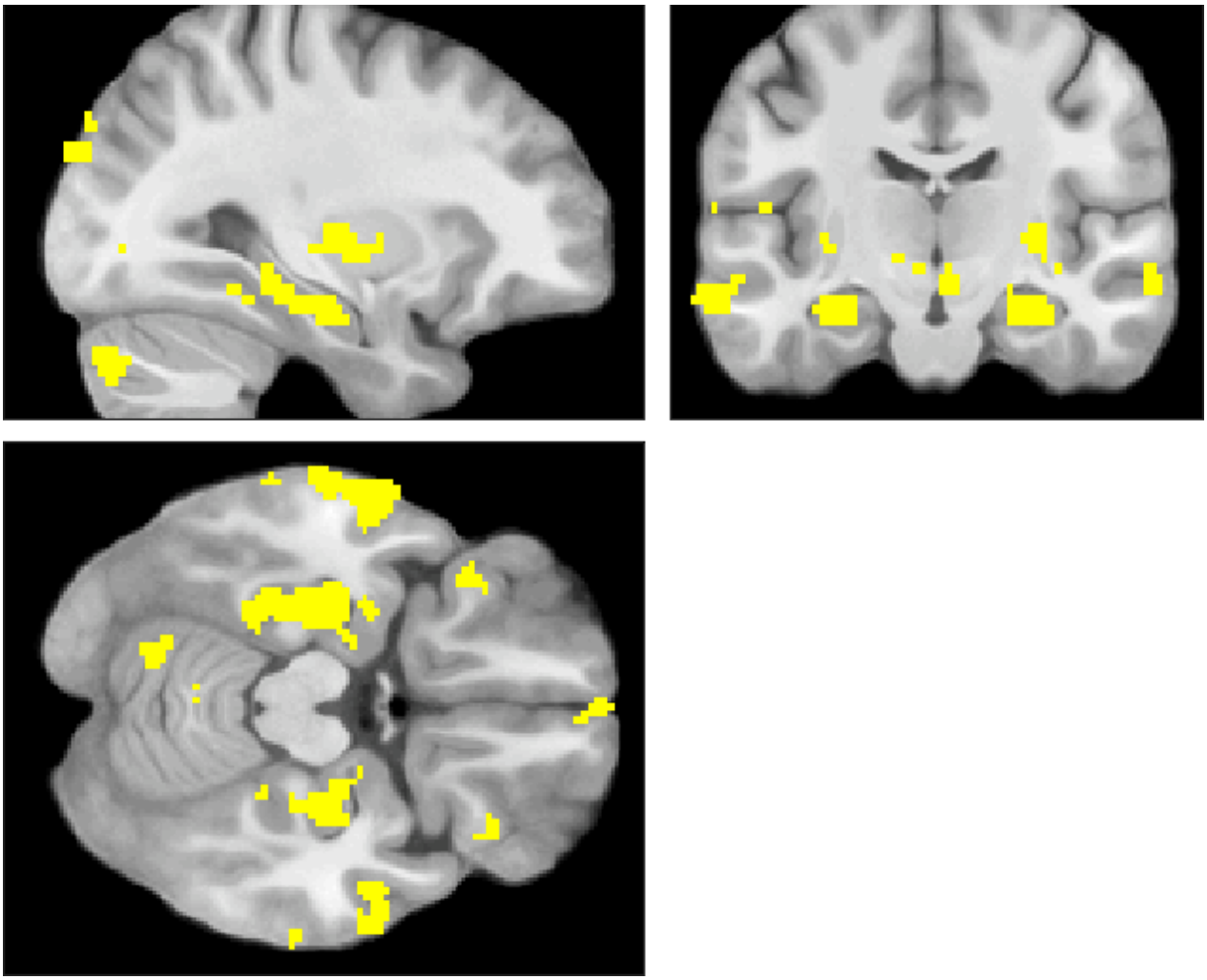

Table S1. Summary of univariate analysis results. Clusters of activity exceeding whole-brain $p < 0.05$ FWE-corrected (cluster-forming threshold $p < 0.005$). Within each cluster, the first 10 regions are listed that include >= 10 voxels of a cluster. For the reward versus non-reward contrast at feedback, the cluster-forming threshold was set to $p < 0.0005$ to allow for informative clusters.

Contrast	Regions	Cluster size	x	y	z	Peak z stat
Decay (State 1)	L Superior Frontal Gyrus	454	-30	26	54	3.78
	R Middle Occipital Gyrus R Middle Temporal Gyrus R Superior Occipital Gyrus	450	18	-100	18	4.1
	R Lingual Gyrus R Fusiform Gyrus R Middle Occipital Gyrus	295	24	-78	-9	3.96
	R Parahippocampal Gyrus R Fusiform Gyrus R Occipital Gyrus	218	28	-44	-12	4.24
	L Middle Occipital Gyrus L Superior Occipital Gyrus	126	-34	-86	18	3.42
Decay (State 2)	R Precuneus L Precuneus	296	8	-66	48	4.11
Decay (State 3)						
Reward	Middle Occipital Gyrus Middle Temporal Gyrus Lingual Gyrus Cuneus L Superior Temporal Gyrus Inferior Temporal Gyrus Inferior Parietal Lobule Inferior Occipital Gyrus R Precuneus R Fusiform Gyrus L Insula	5642	7.01	46	-62	6
	Medial Frontal Gyrus Anterior Cingulate L Middle Frontal Gyrus Superior Frontal Gyrus	1855	5.24	-14	56	9

	L Inferior Frontal Gyrus Ventral Striatum Caudate Head Rectal Gyrus Subcallosal Gyrus					
	Cingulate Gyrus Paracentral Lobule Medial Frontal Gyrus Precentral Gyrus Precuneus Postcentral Gyrus	1785	5.04	-10	12	27
	R Superior Temporal Gyrus R Inferior Parietal Lobule R Precentral Gyrus R Postcentral Gyrus R Insula R Transverse Temporal Gyrus R Inferior Frontal Gyrus	1408	5.55	58	-36	15
	L Precentral Gyrus L Transverse Temporal Gyrus L Postcentral Gyrus L Superior Temporal Gyrus L Insula	239	4.55	-44	-24	9
	R Dorsal Caudate White Matter	176	4.87	22	10	30
	R Precentral Gyrus R Postcentral Gyrus	168	4.35	40	-22	42
	R Putamen R Insula	132	4.39	32	0	3
	R Medial Frontal Gyrus R Middle Frontal Gyrus R Superior Frontal Gyrus	118	4.45	16	54	3
	L Caudate Tail L Posterior Putamen	86	4.35	-18	-2	27
	L Superior Temporal Gyrus	82	4.71	-64	-10	3
	L Cingulate Gyrus L Precuneus	51	4.41	-8	-42	30
Reward (State rest)	L Middle Temporal Gyrus Posterior Cingulate Cuneus Precuneus L Middle Occipital Gyrus Cingulate Gyrus	1953	4.53	-20	-74	27

	L Superior Occipital Gyrus L Lingual Gyrus L Inferior Occipital Gyrus L Parahippocampal Gyrus L Superior Temporal Gyrus					
	Medial Frontal Gyrus Anterior Cingulate Ventral Striatum Caudate Head Rectal Gyrus L Superior Frontal Gyrus	1321	5.35	8	24	3
	R Middle Occipital Gyrus R Inferior Occipital Gyrus R Cuneus R Middle Temporal Gyrus R Inferior Temporal Gyrus R Posterior Cingulate R Lingual Gyrus	995	4.53	32	-62	9
	L Cingulate Gyrus White Matter	458	4.76	-22	-26	36
	R Precentral Gyrus R Postcentral Gyrus	333	4.25	30	-38	72
	R Hippocampus R Parahippocampal Gyrus	250	4.93	32	-34	0
	L Parahippocampal Gyrus L Fusiform Gyrus L Hippocampus L Middle Occipital Gyrus L Middle Temporal Gyrus	250	3.88	-38	-52	-12
	R Precentral Gyrus R Insula R Postcentral Gyrus	207	4.3	56	0	18
	L Medial Frontal Gyrus L Superior Frontal Gyrus	181	4	-16	42	24