Distinct replay signatures for prospective decision-making and memory preservation

Authors: G. Elliott Wimmer[1,2,*], Yunzhe Liu [3,4,1,2], Daniel C. McNamee[1,2,5], Raymond J. Dolan[1,2,3]

Affiliations:

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

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

[3]State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China.

[4]Chinese Institute for Brain Research, Beijing, China.

[5]Neuroscience Programme, Champalimaud Research, Lisbon, Portugal

* Corresponding author. Email: e.wimmer@ucl.ac.uk (G.E.W.)

Abstract

Theories of neural replay propose that it supports a range of functions, most prominently planning and memory consolidation. Here, we test the hypothesis that distinct signatures of replay in the same task are related to model-based decision-making ('planning') and memory consolidation ('preservation'). We designed a reward learning task wherein participants utilized structure knowledge for 'model-based' evaluation, while at the same time had to maintain knowledge of two independent and randomly alternating task environments. Using magnetoencephalography (MEG) and multivariate analysis, we first identified temporally compressed sequential reactivation, or replay, both prior to decisions and following reward feedback. Before choice, prospective replay strength was enhanced for the current task-relevant environment when the benefits of a model-based planning strategy were higher. Following reward receipt, and consistent with a memory preservation role, replay for the alternative 'distal' task environment was enhanced as a function of decreasing recency of experience with that environment. Critically, these planning and memory preservation relationships were selective to pre-choice and post-feedback periods. Our results provide new support for key theoretical proposals regarding the functional role of replay and demonstrate that the relative strength of planning and memory-related signals are modulated by on-going computational and task demands.

Significance statement

The sequential neural reactivation of prior experience, known as replay, is considered to be an important mechanism for future planning as well as preserving memories of the past. Whether, and how, replay supports both of these functions remains unknown. Here, in humans, we found that prior to a choice, rapid replay of potential future paths was enhanced when planning was more beneficial. By contrast, after choice feedback, when no future actions are imminent, we found evidence for a memory preservation signal evident in enhanced replay of paths that had been visited less in the recent past. The results demonstrate that distinct replay signatures, expressed at different times, relate to two dissociable cognitive functions.

Introduction

Humans have a remarkable ability to process information that extends beyond the immediately perceptible, including simulation of prospective plans and retrieval of past memories. It has been hypothesized that hippocampal replay contributes to both these abilities (Diba and Buzsaki, 2007; Buzsaki, 2015; Foster, 2017; Joo and Frank, 2018; Olafsdottir et al., 2018). In rodents, replay is strongly linked to the hippocampus, where cells encoding distinct locations reactivate in a coordinated sequential manner, recapitulating past or simulating potential future trajectories (Diba and Buzsaki, 2007; Buzsaki, 2015; Wikenheiser and Redish, 2015; Foster, 2017; Joo and Frank, 2018; Olafsdottir et al., 2018). A similar phenomenon of sequential reactivation has been identified in humans using decoding techniques in conjunction with high temporal resolution MEG data (Kurth-Nelson et al., 2016; Liu et al., 2019; Michelmann et al., 2019; Eldar et al., 2020; Wimmer et al., 2020; Liu et al., 2021b; Nour et al., 2021; Wise et al., 2021).

Prominent theories of neural replay propose that it is important for planning future actions (Buckner, 2010; Dolan and Dayan, 2013; Pezzulo et al., 2017; Pezzulo et al., 2019) in addition to supporting memory preservation (Wilson and McNaughton, 1994; McClelland et al., 1995; Carr et al., 2011; Kumaran et al., 2016; van de Ven et al., 2016; van de Ven et al., 2020; McNamee et al., 2021). One hypothesis is that task demands, operationalized as temporal proximity to action versus feedback, determine the contribution of replay to planning and memory, respectively (Olafsdottir et al., 2017; Olafsdottir et al., 2018). However, the contribution of awake on-line replay to these two functions has largely been addressed in the context of separate experiments (Johnson and Redish, 2007; Gupta et al., 2010; Pfeiffer and Foster, 2013; Olafsdottir et al., 2015; Wikenheiser and Redish, 2015; Ambrose et al., 2016; Papale et al., 2016; Olafsdottir et al., 2017; Joo and Frank, 2018; Olafsdottir et al., 2018; Shin et al., 2019; Wimmer et al., 2020). Here we directly address the contribution of replay to both these roles within a single task context.

Replay of trajectories leading to a goal has been proposed to underpin decision making that exploits structure knowledge of an environment, referred to as model-based decision-making (Dolan and Dayan, 2013; Mattar and Daw, 2018; Pezzulo et al., 2019). A number of rodent studies indicate a link between hippocampal neural sequences and subsequent path choice selection, consistent with a role for replay in planning (Pfeiffer and Foster, 2013; Singer et al., 2013; Wikenheiser and Redish, 2015; Wu et al., 2017; Xu et al., 2019). However, an inconsistency in such findings raises the possibility that any relationship between replay and subsequent choice might differ across evaluation strategies and reward environments (Pfeiffer and Foster, 2013; Wikenheiser and Redish, 2015; Wu et al., 2017; Carey et al., 2019; Shin et al., 2019; Xu et al., 2019; Gillespie et al., 2021; Wise et al., 2021). Critically, and regardless of any relationship to choice identity, brain lesion studies highlight a necessary role for the hippocampus in model-based behavior (Miller et al., 2017; Vikbladh et al., 2019). This suggests that hippocampal replay may be enhanced when model-based decision-making (planning) is beneficial. Thus far, however, there is no clear evidence linking demands for model-based control and neural replay preceding choice (Kurth-Nelson et al., 2016; Eldar et al., 2020; Liu et al., 2021b; Wise et al., 2021).

Beyond planning, replay is considered critical for memory preservation, where replay of previous experiences might serve to strengthen memory and prevent interference from newer experiences (‘catastrophic forgetting’) (McCloskey and Cohen, 1989; Ratcliff, 1990; McClelland et al., 1995; Kumaran et al., 2016; van de Ven et al., 2020). Studies that have disrupted hippocampal activity support the idea that offline place cell reactivation is critical for learning, memory consolidation, or both (Girardeau et al., 2009; Ego-Stengel and Wilson, 2010; Jadhav et al., 2012; Fernandez-Ruiz et al., 2019). It has been conjectured that human rest-period replay subserves a similar function (Schapiro et al., 2018; Liu et al., 2019; Schuck and Niv, 2019; Nour et al., 2021). Studies of replay in rodents navigating a single environmental context provide initial, but inconclusive, evidence for a link between recent experience with an environment and replay (Gupta et al., 2010; Carey et al., 2019).

Here, we address a role for replay in both planning and memory preservation in a context where participants needed to retain a memory of a 'distal' environment while at the same time learning within a local one. To do this we adapted a reward learning task originally designed to study model-based decision-making (Daw et al., 2011; Doll et al., 2015; Kool et al., 2016; Miller et al., 2017), where distinct start states converge upon shared paths. Critically, to study memory, we included two independent randomly alternating environments. Both the early convergence on shared paths and the alternation of environments strongly favors online planning, and these features distinguish the current paradigm from a recent related report (Liu et al., 2021b). Using recently-developed MEG analytic methods we first identify sequential neural reactivation (Kurth-Nelson et al., 2016; Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a; Liu et al., 2021b; Nour et al., 2021; Wise et al., 2021) and then ask whether replay strength varies as a function of task demands and recent experience. We hypothesized replay would be boosted during pre-choice path planning when model-based decision making was more beneficial (Dolan and Dayan, 2013; Doll et al., 2015; Miller et al., 2017; Pezzulo et al., 2017; Pezzulo et al., 2019; Vikbladh et al., 2019). By contrast, following receipt of choice feedback, we hypothesized replay for an alternative environment would relate to the infrequency of recent experience, consistent with a role in supporting memory preservation (Olafsdottir et al., 2018).

Results

Behavior.

Participants navigated two separate, independent environments ('worlds') in order to earn reward points (**Fig. 1a**). Each world contained two path options, where a top-level shape led deterministically to one of the sequential paths comprising 3 unique stimuli. Knowledge of these paths was tested after each scanning block. In a given world, each path led to a separate stream of reward feedback which drifted over time (**Fig. S2a**). Importantly, within a given world, each trial could start in one of two equivalent start states leading to the two paths. For an agent to perform well in the task, outcomes should have an equivalent influence on subsequent choices regardless of whether an

agent starts in the same start state or the alternative, equivalent, start state. This design feature, combined with a sufficient drift in reward across trials (Kool et al., 2016), allows us to characterize how well participants use structure knowledge to guide model-based behavior (**Fig. 1c**) (Doll et al., 2015; Kool et al., 2016; Kool et al., 2017; Patzelt et al., 2019).

By way of example, imagine an agent is faced with a choice in world 1, start state A (**Fig. 1a**). The agent selects the diamond over the crown and receives an unexpected high reward. If next faced with a choice in start state B, a model-based agent will generalize this experience to promote the subsequent choice of the magenta pentagon to reach the same just-rewarded path. In contrast, a model-free agent does not have access to this recent relevant experience. Thus, only a model-based agent can exploit structure knowledge to allow generalization of reward feedback across equivalent start states. Such behaviour has been proposed to involve looking ahead to values associated with terminal states, possibly using prospective neural reactivation or replay (Doll et al., 2015). Thus, this design allows us to identify behavior reflective of model-based versus model-free learning (Doll et al., 2015; Kool et al., 2016; Kool et al., 2017; Patzelt et al., 2019), analogous to variants of the paradigm that use probabilistic state transitions (Daw et al., 2011; Miller et al., 2017; Vikbladh et al., 2019).

One feature of the our deterministic task variant is that trials are divided between those where model-based behavior is beneficial (different start state) versus not (same start state) (Doll et al., 2015), allowing us to test for an association between neural replay and the benefits accruing from model-based behavior. Furthermore, a deterministic transition structure and an absence of branching paths increases our ability to detect evidence of sequential neural reactivation (Liu et al., 2021b). With respect to planning, the fact that our design includes multiple worlds decreases the predictability of an upcoming trial, thus promoting deployment of planning-related processes at choice as opposed to after outcome feedback (Konovalov and Krajbich, 2016; Liu et al., 2021b). Critically, including multiple worlds also allows us to examine replay signatures of memory preservation for more distal (non-local) experiences.

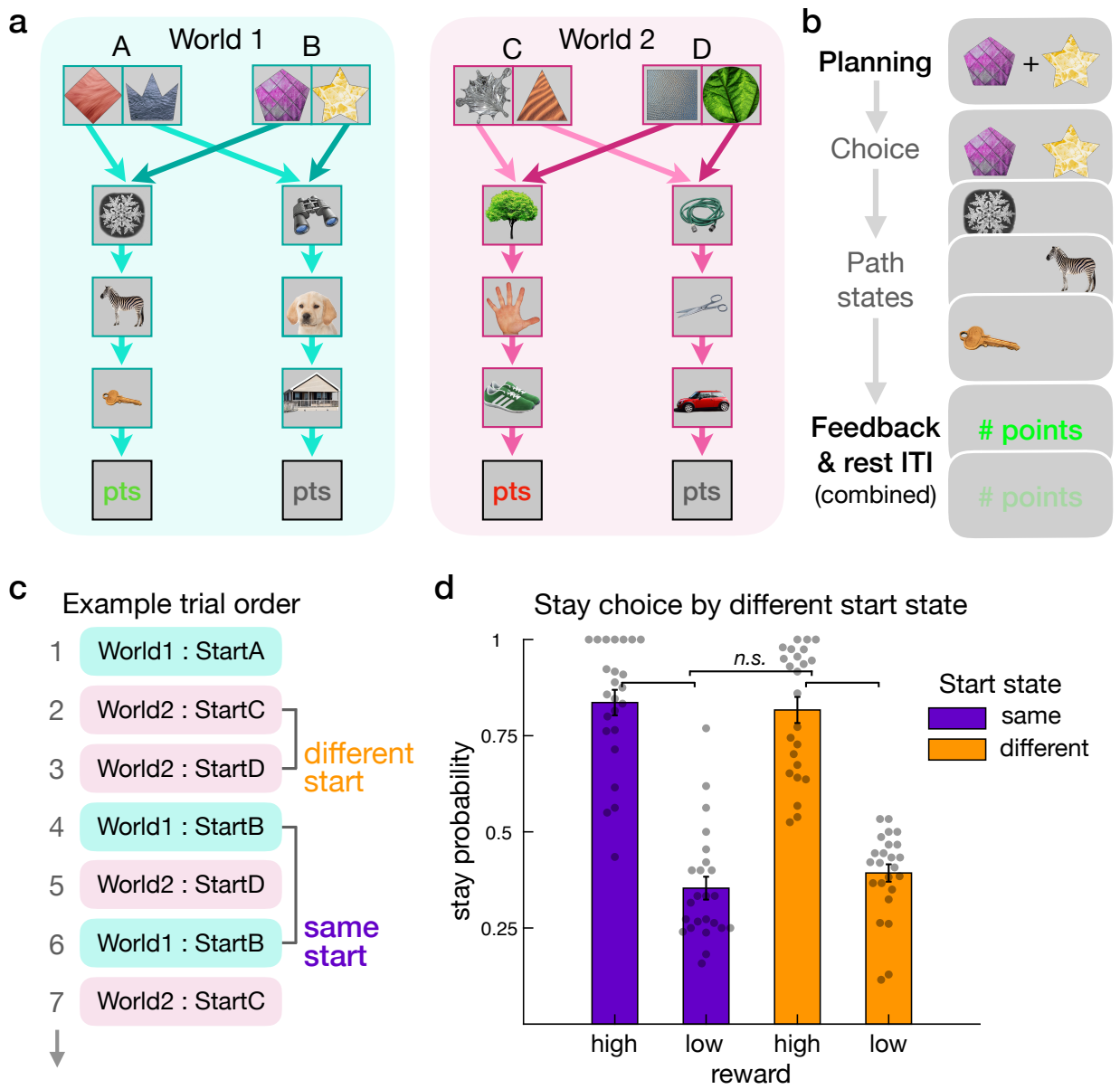

Figure 1. Two-environment reward learning task and generalization behavior. (**a**) Task schematic showing two alternative 'worlds' and their two equivalent start states. Trials in each world start at one of two equivalent shape pair options; to illustrate these connections, arrows from different states differ in color saturation. The shape options then lead deterministically to the same paths and reward outcomes (0-9 points (pts)). To learn this general structure participants engaged in a pre-scanning training session. For the MEG scanning session, participants then learned two worlds populated with new images. Participants' memory for the path sequences was at asymptote after an initial no-reward exploration period. (**b**) Key trial periods in the reward learning phase. Replay was measured prior to choice in a time window we subsequently refer to as the

‘planning period’. After the disappearance of a central cross, participants entered their response. Participants then sequentially viewed the state images corresponding to the chosen path. Finally, participants received reward feedback (0-9 points), the amount of which drifted across trials. Replay was again measured in the post-feedback time window. For interpretation of subsequent results, in this example World 1 is the ‘current world’ while World 2 is the non-presented ‘other world’. (See also **Fig. S1**.) (**c**) Example of a trial sequence of World 1 and World 2 trials, highlighting an example of a trial with a different and a same start state as the previous trial in a world, with labels following panel a. (**d**) Illustration of the dependence of repeated choices (stay) on previous rewards, conditional on whether the start state in the current world was the same as in the previous trial or not. The plot depicts the probability of a stay choice (when participants repeated a previous path selection in a given world) following above-average (high) versus below-average (low) reward. This difference was equivalent for same (purple) versus different (orange) starting states, consistent with behavior being model-based (‘n.s.’ represents non-significant effects in a regression model using continuous reward data). For display purposes, graded point feedback was binarized into high and low and trials with near-mean feedback were excluded; alternative procedures yield the same qualitative results. Grey dots represent individual participants. (See also **Fig. S2**.)

After an initial scanned exploration period without reward feedback, participants performed the primary reward learning task. To ascertain the degree to which behavior was guided by model-based and model-free learning we used a regression approach in combination with a set of reinforcement learning models. Our regression analyses quantify a model-based influence on behavior by testing for an effect of generalization: whether a prior reward has a different effect on choice when starting in the same start state versus a different state than previous experience (Doll et al., 2015; Kool et al., 2016). A model-based controller acts to generalize reward feedback across equivalent starting states, potentially using structure knowledge to look ahead and evaluate expected terminal rewards. Alternatively, updating the equivalent start state can occur after feedback (prior to choice), a form of non-local learning (Liu et al., 2021b). In our analysis, if the effect of previous reward on choice is similar when starting in a different versus same state, then this indicates that participants’ behaviour reflects an

exploitation of structure knowledge for a task environment (Daw et al., 2011; Doll et al., 2015). Qualitatively, we observed that receipt of high versus low reward enhanced the likelihood of choice repetition (**Fig. 1d**) while a high degree of model-based behavior was evident in equivalent stay probabilities for choices starting at the same versus a different state (**Fig. 1d**).

Using logistic regression we quantified the effect of reward and start state on choice (Kool et al., 2016), finding a robust overall effect of previous reward on stay choices (β (regression coefficient) = 0.536 [0.424 0.647]; z = 9.434, p < 0.0001). Importantly, there was an equivalent influence of previous reward on stay choices in the same versus a different start state, consistent with the absence of a significant model-free contribution (interaction between reward and same start state, β = 0.0313 [-0.0365 0.0991]; z = 0.905, p = 0.365). As a non-significant effect does not provide evidence in support of the null hypothesis, we employed a two-one-sided test (TOST) equivalence procedure to enable us to reject the presence of a medium- or larger-sized effect (Lakens, 2017). Indeed, based on this we can reject a medium- or larger-sized model-free interaction effect (TOST equivalence test p = 0.027). While reward-guided choice behaviour was unaffected by start state changes, we found that that reaction times were overall slower for choices on start state change trials (different versus same start state trials β = 0.009 [0.0017 0.0163]; z = 2.423, p = 0.0155), even though the delayed choice limits reaction time variance. In a complementary regression approach, testing the effect of previous reward on the identity of option selection (**Supp. Results**) (Doll et al., 2015), we again found no interaction between reward and start state, consistent with a model-based learning signature (Daw et al., 2011; Doll et al., 2015). Overall, this allows us to infer that behavior is guided by reward to the same degree irrespective of generalization demands, likely reflecting participants' behavior being strongly model-based.

Next, we compared fully model-based and model-free reinforcement learning models to a hybrid model commonly used to assess the relative strength of model-based and model-free learning (Daw et al., 2011; Kool et al., 2016). The hybrid model includes a weighting parameter *w* which controls the degree of model-based influence on choice.

Overall, behavior was best explained by a model-based controller, which outperformed the hybrid model, while the fit of the model-free controller was poor (**Table S1**-**S2**). While the hybrid model exhibited a numerical benefit in raw fit, when penalized for model complexity the pure model-based controller provided an equivalent (measured via AIC) or better fit (measured via BIC; **Table S2**). At the individual participant level, the model-based controller provided a better fit for more than 83% of participants (using either AIC or BIC). Finally, given that good performance requires model-based generalization throughout the task, as expected we found no evidence for a change in model-based behavior over time (**Supp. Results**). The high degree of model-based behavior provided a robust context for us to next examine the associated neural processes.

We also obtained memory tests for sequential path stimuli to confirm that participants learned the path structure. Our task was designed to achieve high memory performance by the start of reward learning, as this would serve to both boost model-based behaviour as well as our ability to detect sequential reactivation. Memory performance, already at a high level at the end of the incentivized structure learning phase (second half, 87.5% [78.2 96.8]), was maintained at this level across the reward learning phase (mean 95.0% [91.3 98.7, with all blocks above 92%; effect of block, $p > 0.64$).

Sequenceness identification.

In our neural analyses, we first established reliable decoding from neural patterns evoked by the unique stimuli that indexed individual path states (**Fig. 1a**). A classifier trained during a pre-task localizer showed successful discrimination of all path stimuli, with peak decoding evident from 140-210 ms post stimulus onset. Based on this, and to maintain consistency with our prior studies, we selected a post-stimulus 200 ms time point for subsequent replay analyses (Kurth-Nelson et al., 2016; Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021b). The trained classifier generalized from the localizer to the actual presentation of path objects during reward learning, showing significant across-task classification ($t_{(23)} = 7.361$, $p < 10^{-7}$; **Fig. S3**). We also found evidence for significant reactivation of stimuli both during planning and after feedback

(as compared to reactivation derived from permuted classifiers; p-values < 0.01; see **Methods**).

We next used the trained classifiers to seek evidence for time-compressed sequential reactivation of path elements. First, we applied the classifiers to reward learning task MEG data to derive measures of state reactivation in each trial separately for each state and at each timepoint, focusing on pre-choice planning and post-feedback rest periods (**Fig. 1b**). In a following step we tested for time lagged cross-correlations between state reactivations within these periods, yielding a measure of 'sequenceness' in both forward and backward temporal directions at each lag (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a) (**Fig. S3**). We use the term sequenceness (or 'sequence strength') to refer to the prediction strength of state *i* to state *j* at some time lag, while we operationally refer to any reactivation of state sequences here as replay. This sequence detection method, validated in previous work, quantifies the average predictivity of state *i* to state *j* within a period, reflecting both the frequency and fidelity of replay events (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a; Liu et al., 2021b; Nour et al., 2021).

Our a priori hypothesis is that signatures of replay are differentially modulated by task environment and recency of experience with an environment; the presence of forward or backward replay during planning and feedback (e.g. Diba and Buzsaki, 2007; Kurth-Nelson et al., 2016) is a necessary precondition for testing this hypothesis. To increase power, sequences included all possible paths, the (to-be) chosen and (to-be) nonchosen paths in the current trial, as well the two paths for the 'other world'. We identified sequenceness time lags of interest by comparing evidence across lags for all valid sequences, using a significance threshold determined by a permutation of stimulus assignment to paths (following previous work; Kurth-Nelson et al., 2016; Liu et al., 2019; Liu et al., 2021a; Liu et al., 2021b). These analyses revealed that during planning there was significant forward sequenceness, with a state-to-state time lag of 70 and 80 ms and also of 190-200ms. For the shorter lags, this indicates that, on average across participants, a given state was followed by reactivation of an adjoining state within the

same path at a delay of 70 to 80 ms (**Fig. 2c** and **Fig. S4**). We found no significant evidence for backward sequenceness during planning.

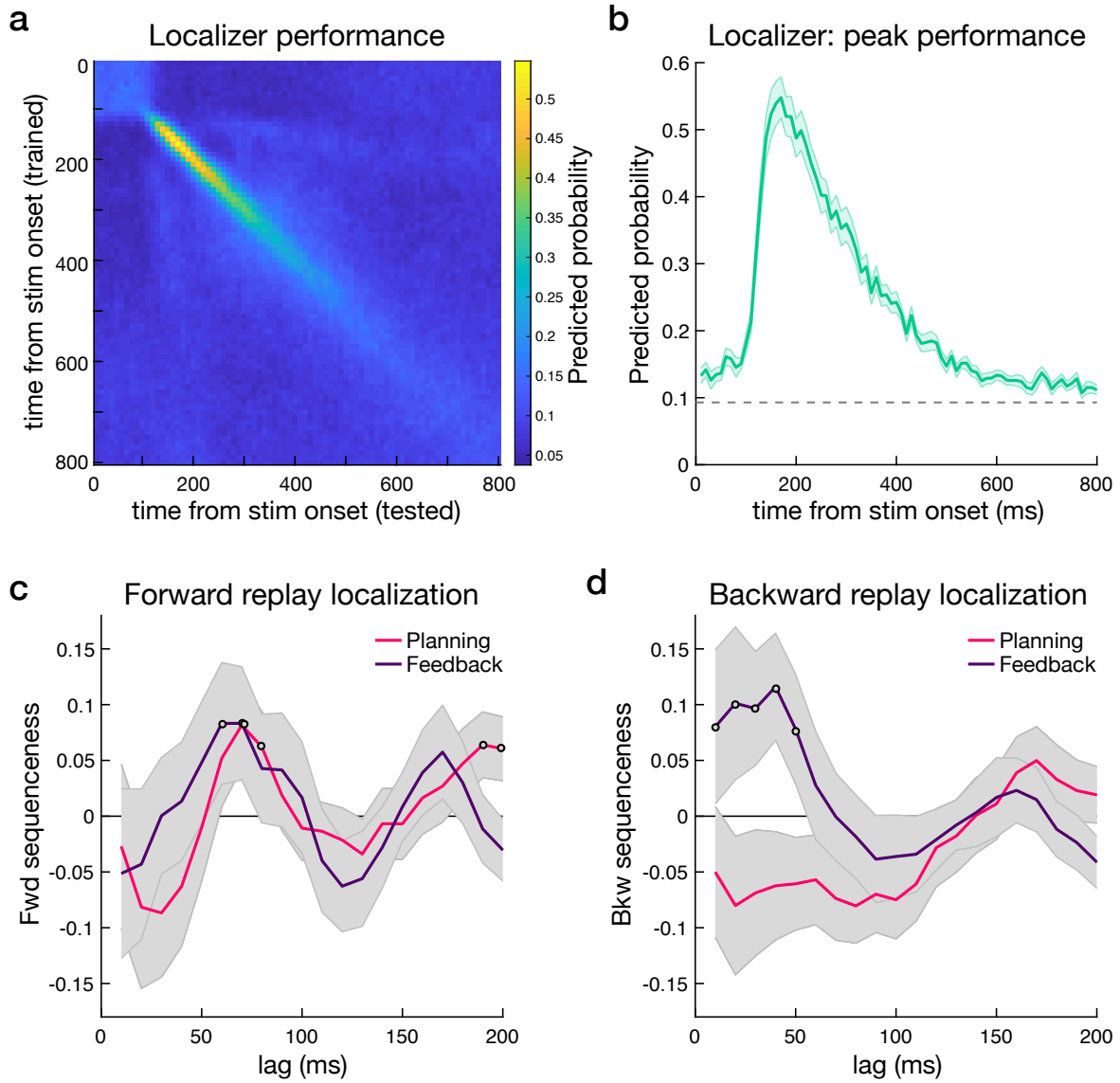

Figure 2. Training of state localizer and sequenceness time lag identification. (**a**) Classifier performance for path state stimuli presented during a pre-task localizer phase, training and testing at all time points. This revealed good discrimination between the 12 path stimuli used in the learning task. Color bar indicates predicted probability. Note that start state shape stimuli were not included in the pre-task localizer and are not included in sequenceness analyses. (**b**) Peak classifier performance from 140-210 ms after stimulus onset in the localizer phase (depicting the diagonal extracted from panel a). See also **Fig. S3**. (**c**) Forward sequenceness for all learned paths during planning and

feedback periods was evident at a common state-to-state lag of 70 ms in both trial periods. Open dots indicate time points exceeding a permutation significance threshold, which differs for the two periods. (**d**) Backward sequenceness for all learned paths during planning and feedback periods was evident at state-to-state lags that spanned 10-50 ms in feedback period alone. Note that the x-axis in the sequenceness panels indicates the lag between reactivations, derived as a summary measure across seconds; the axis does not represent time within a trial period. Open dots indicate time points exceeding a permutation significance threshold, which differs for the two periods. Shaded error margins represent SEM. See **Fig. S5** for example sequenceness events and **Fig. S4** for extended time lags.

We then examined replay following outcome feedback, a period when the displayed reward points faded from the screen toward a brief inter-trial interval rest (**Fig. 1b**; **Fig. S1a**). This corresponds to a time when replay has previously been identified in rodents (e.g. Ambrose et al., 2016). Here we identified significant sequenceness with a peak state-to-state time lag at 40 ms in a backward direction, and 60 to 70 ms in the forward direction (**Fig. 2c**-**d** and **Fig. S4**). To focus on a period with less cognitive demands arising from actual feedback processing and value updating, with similarities to a procedural step in a related rodent study (Olafsdottir et al., 2017), our analysis was focused on the latter 3.5 s of the 5 s post-feedback period; note, however, qualitatively similar results were found when using the full feedback period. In general, our finding of forward and backward replay events, intermixed across seconds, echoes results reported in rodent studies (Davidson et al., 2009; Gillespie et al., 2021) as well as in a recent human study (Liu et al., 2021b).

Based on these initial replay temporal localization analyses, we focused our primary analyses on forward sequenceness with a 70 ms lag between states identified in both the planning and feedback periods. At feedback, we selected the peak lag of 40 ms from the above-threshold lags for backward sequenceness analyses, informed by our previous work (Kurth-Nelson et al., 2016; Liu et al., 2019). For completeness, we also analyzed feedback period forward replay at a lag of 70 ms. To examine links between

replay and task experience, we estimated sequenceness separately for path transitions in the current world and other world, for each period on each trial.

Replay during planning and model-based generalization.

We first tested a link between planning and relative sequence strength, leveraging the fact that in our task model-based generalization was more beneficial when a start state changes relative to when it remains the same (**Fig. 1d**) (Doll et al., 2015). A planning account predicts that when the start state changes the benefit of model-based generalization should be reflected in a boosting of replay relative to when the start state remains the same (**Fig 1c**). To increase power, our sequenceness analyses combined evidence for the two current world paths.

In line with predictions, sequenceness for current world paths was significantly stronger when the start state changed versus remained the same (i.e. when generalization was likely to be more beneficial; multilevel regression β (regression coefficient) = 0.1411 [0.0625 0.2197]; z = 3.519, p = 0.0004; **Fig. 3a**). There was no relationship between sequenceness for other world paths and generalization (β = -0.0201 [-0.0979 0.0577]; z = -0.506, p = 0.613; other world TOST equivalence test p = 0.012; current versus other difference z = 2.829, p = 0.0023, one-tailed). Analyses across different state-to-state lags indicated the presence of a selective effect of generalization centered on the independently selected lag of 70 ms (**Fig. 3b**). Further, the effect of generalization was stable across trials (positive interaction with trial, t = 1.055, p = 0.292; see **Supp. Results** and **Table S3** for additional control analyses). This relationship was also qualitatively similar for both the to-be chosen and the to-be non-chosen paths (p-values < 0.033). In follow-up control analyses, we also tested whether higher pre-choice replay was better explained by our planning account versus an alternative memory retrieval account. The planning account predicts a stepwise increase in replay on trials when generalization was more beneficial, whereas an alternative memory retrieval account predicts a graded increase in replay related to how far in the past a start state was last seen. We found that a planning account provided a better fit to the data (**Supp. Results**). Overall, the link between replay and model-based generalization rests on the

detection of evidence for the same sequential neural representations being triggered by two different sets of choice cues, demonstrating a form of neural generalization underlying behavioural generalization.

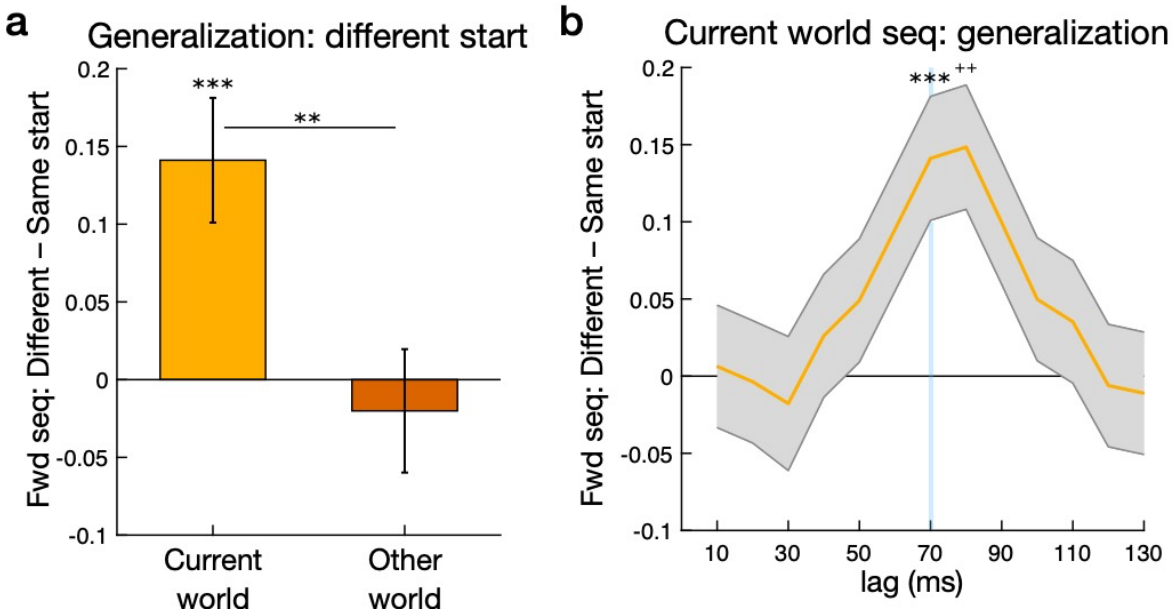

Figure 3. Planning period replay increases and the benefit of model-based generalization. (**a**) Stronger forward replay on trials where the start state is different from the previous trial, where there is a greater benefit in utilizing model-based knowledge (**b**) Timecourse of regression coefficients for variables of interest, showing effects at state-to-state lags from 10 to 130 ms. The light blue line highlights the 70 ms time lag of interest shown in panel a. Y-axes represent sequenceness regression coefficients for binary different versus same start state. See **Fig. S4** for extended time lags. Seq = sequenceness. ** $p<0.01$; *** $p<0.001$; + $p<0.01$, corrected for multiple comparisons.

We next examined whether planning-related replay was modulated by option value, in light of previous imaging and electrophysiology studies in humans reporting correlations between non-sequential hippocampal activity and value (e.g. Kumaran et al., 2009; Lebreton et al., 2009; Wimmer et al., 2012; Lopez-Persem et al., 2020). If replay is involved in deriving value estimates, then we would not necessarily expect a modulation of replay by value, though it is possible that replay might be biased by option value when values are directly informed by recent experience.

Examining the relationship between replay and mean state value, the average model-predicted value across the two options, we found that current world replay strength significantly correlated with mean state value (β (regression coefficient) = 0.0479 [0.0130 0.0829]; t = 2.688, p = 0.0072). However, this relationship was only found on trials where there were no generalization demands (where the start state remained the same; same start trial value β = 0.1128 [0.0484 0.1771]; t = 3.440, p = 0.006; different start trial value β = 0.0110 [-0.0310 0.0530]; t = 0.513, p = 0.608; interaction β = 0.0510 [0.0140 0.0880]; t = 2.702, p = 0.0069). These results were also selective to current world forward replay during planning (**Supp. Results** and **Table S3**). Thus, replay positively related to option value in conditions where start state options were susceptible to direct reinforcement on the preceding trial, where we speculate value inference is less demanding.

Backward replay prioritization at feedback and memory preservation.

We next tested a prediction that replay during periods of low cognitive demand, specifically following reward feedback, relates to automatic memory maintenance processes, which we refer to as 'preservation. Here we focused our analyses on backward replay with a 40 ms time lag, a signal selective to the feedback period (**Fig. 2d**). Behaviorally, we found consistently high levels of path memory during reward learning (see above) and this precluded examining direct links between feedback replay and memory variability. However, in the preceding brief structure learning phase when participants first experienced the sequential paths mean memory performance was lower (78.6% [68.0 89.3]), allowing us to test for a link with backward replay within the inter-trial interval (as no reward feedback was presented). Backward replay exhibited a numerical, but non-significant, increase across time (40ms lag, second half – first half trials; 0.101 [-0.034 0.237]; p = 0.136). Notably, increased backward replay from early to late trials in this initial phase correlated significantly with individual differences in memory performance during this phase (r = 0.409, p = 0.0470; **Fig. S4**). Although the number of trials here is much lower than the primary reward learning phase, this provides initial evidence consistent with a link between replay and memory.

After initial experience, memory preservation can be considered to be an automatic process driven in part by the infrequency, or rarity, of recent experience for a given environment (McCloskey and Cohen, 1989). In the primary reward learning phase, we operationalized rarity as an exponentially weighted average of past exposures to each environment. We found that backward replay of other world paths was greater when they had been experienced less frequently over recent trials (rarity effect for other world paths, all trials $\beta = 0.0513$ [-0.0921 -0.0104]; $t = 2.463$, $p = 0.0139$; **Fig. 4a**). We found no relationship between current world replay and rarity (current $\beta = -0.0101$ [-0.0307 0.0509]; $t = 0.486$, $p = 0.627$; TOST $p = 0.011$; other versus current difference $z = -2.076$, $p = 0.0190$, one-tailed; **Fig. 4a**). Moreover, the relationship with rarity was stable across trials (interaction with trial, $t = 0.487$, $p = 0.627$; see **Supp. Results** and **Table S3** for additional control analyses). We also confirmed this experience-replay relationship in a basic model that makes no assumptions about learning, finding that other world replay was stronger when the other world had been experienced more than one trial ago versus when experienced on the previous trial ($p = 0.0469$).

To further explore this putative memory preservation signal, we also examined a link between the rare experience replay effect and planning forward replay. Stronger replay of rare experiences after feedback might be expected to decrease the need for planning replay. The planning replay signal was extracted from trials where the world changed from trial-to-trial alone, as this captures where any preceding feedback period ‘other world’ replay effects may relate to planning. Consistent with this we found an inverse relationship between the strength of the modulation of backward replay by rarity and planning replay across participants (current world forward replay, world change trials $r = -0.521$, $p = 0.009$; **Fig. 4**).

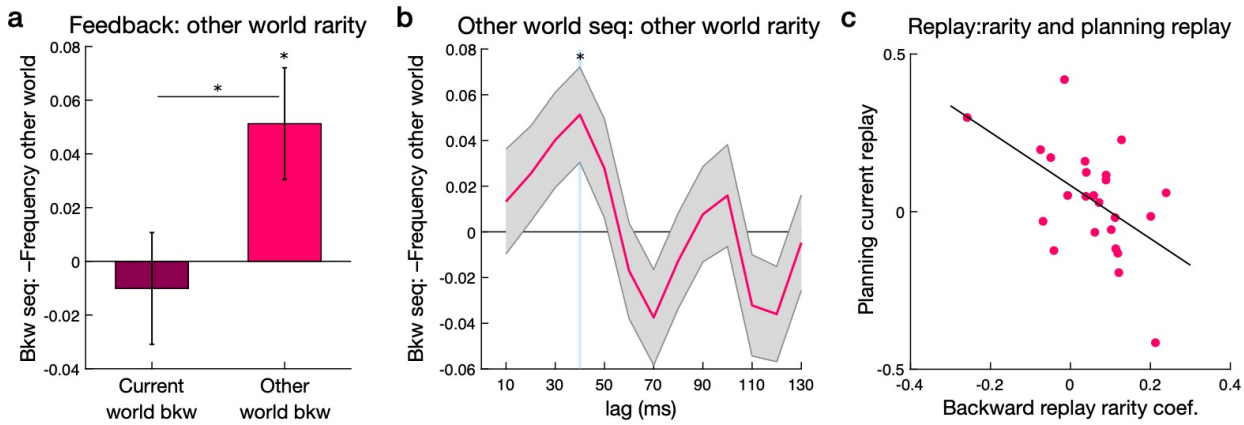

Figure 4. Feedback period backward replay increases with the rarity of recent other world experience. (**a**) Rarity (lower recent experience) of the other world correlated with greater backward replay of other world paths. (**b**) Timecourse of regression coefficients for the rarity effect of interest, showing effects at state-to-state lags from 10 to 130 ms. The light blue line highlights the 40 time lag of interest shown in panel a. Y-axes represent sequenceness regression coefficients for rarity of other world. See **Fig. S4** for extended time lags. Seq = sequenceness; * $p < 0.05$. (**c**) Across-participant relationship between the replay-rarity effect and lower planning period forward replay (world change trials; $p = 0.009$).

Next, we asked whether there was a link between this rare experience replay signature and choice behaviour. If feedback replay supports memories for more distant structure and value, we might expect the strength of this replay signal to positively influence choice in the other world. In an augmented reinforcement learning model, we tested this connection by allowing replay-related memory preservation to decrease choice uncertainty (or noisiness). The feedback replay measure was extracted from trials preceding a world change, as this is where preceding feedback period replay may relate to a following choice. The model included two additional softmax inverse temperature parameters that applied to world change trials with high versus low preceding feedback replay (**Methods**). A higher inverse temperature parameter can reflect lower uncertainty such that choices are more strongly guided by estimated prospective values. We found a significantly higher inverse temperature when choices were preceded by high replay versus low feedback replay (world change trials; high replay median = 19.40; low replay = 14.27; $p = 0.015$, one-tailed, Wilcoxon signed rank test). Control analyses demonstrated that the replay effect on choice was selective to backward replay of the

relevant world (**Supp. Results**). Further, while backward replay related to the experienced rarity of a world, we found no modulation of choice noisiness by experienced rarity itself, consistent with the internal variability of backward replay underlying the observed effect. Thus, supporting a potential memory preservation mechanism, we found 1) that backward replay was positively modulated by rarity of experience, 2) that a stronger rarity replay effect was linked to lower planning replay strength, and 3) that the strength of backward replay on a trial-to-trial basis decreases uncertainty in subsequent choices.

We then compared the task and experience links to replay that we identified during planning and after feedback to determine if these signals were distinct. We found no correlation between planning period forward replay and rarity of recent experience effect (current world 70 ms lag, $p = 0.815$), with the feedback period significantly stronger than the planning period effect (difference, $z = 1.778$, $p = 0.0378$, one-tailed). Conversely, we found no significant correlation between feedback period backward replay and the benefit of generalization (other world 40 ms lag, $p = 0.119$), while the planning period effect was significantly stronger than the feedback period (difference, $z = 2.891$, $p = 0.002$, one-tailed). Together, these planning and feedback comparisons represent a double dissociation, with feedback period replay being selective for an expected signature of memory preservation.

In additional control analyses, we found no relationship between backward replay and reward feedback or reward prediction error (**Supp. Results** and **Table S3**). Further, during planning, we identified significant forward replay with a 190 and 200 ms time lag (**Fig. 2c**), but found no correlation between this signal and any variables of interest (**Table S3**). At feedback, we also identified significant forward replay with a 70 ms time lag (**Fig. 2c**), but found no correlation between current world forward replay with feedback or other variables of interest (**Supp. Results** and **Table S3**). Finally, in exploratory analyses of a longer 160 ms lag replay signal identified recently (Liu et al., 2021b), we found no relationship with variables of interest at feedback or during planning (**Supp. Results**).

Replay onset beamforming and time-frequency analyses.

To explore the spatial source of sequenceness events we conducted supplemental beamforming source localization analyses to test whether replay onset is associated with increased power in the hippocampus, as previously found (Liu et al., 2019; Wimmer et al., 2020; Higgins et al., 2021; Liu et al., 2021b; Nour et al., 2021). As the interpretation of sources of MEG signal is complex, especially for putative deep regions such as the MTL, we emphasize the supplemental nature of these analyses. Candidate replay onsets were identified by locating sequential reactivation events at time lags of interest, applying a stringent threshold to these events, and conducting broad-band beamforming analyses (as in Wimmer et al., 2020).

At both planning and feedback periods, we identified power increases associated with replay onset in clusters that included peaks in visual cortex extending into the hippocampus (**Fig. 5**; planning forward replay peak MNI coordinates [x, y, z] 20, -81, -2; $p < 0.05$ whole-brain permutation-based cluster correction; feedback backward replay peak; -10, -91, -7; **Fig. S6**; **Table S4**). Using a hippocampal ROI analysis, during the planning period there was a significant power increase at replay onset evident in the right hippocampus (30, -36, -4; $p < 0.05$, ROI permutation-based cluster correction). Within the feedback period, we found bilateral hippocampal power increases (right 30, -21, -12 and 26, -36, 0; left -25, -16, -17; $p < 0.05$, corrected). Similar whole brain and hippocampal ROI results were found when separately looking at planning period current world replay onsets and feedback period other world replay onsets (**Table S4**). These results, along with recent related reports (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021b; Nour et al., 2021), support an interpretation that sequential reactivation is tightly linked to enhanced hippocampal activity.

In separate time-frequency analyses of power changes associated with replay onset events, we found that replay onset for both the planning and feedback periods related to increased power centered at 5-40 Hz ($p< 0.05$, permutation-based cluster-correction; **Fig. S7**). While we found no main effect of replay-related power increases in the ripple band (120-150 Hz) (Liu et al., 2019; Liu et al., 2021b; Nour et al., 2021), at feedback, we found that individual differences in other world replay onset ripple power

correlated with the strength of the modulation of backward replay by rarity (other world – current world; r=0.527, p=0.008; **Fig. S7**).

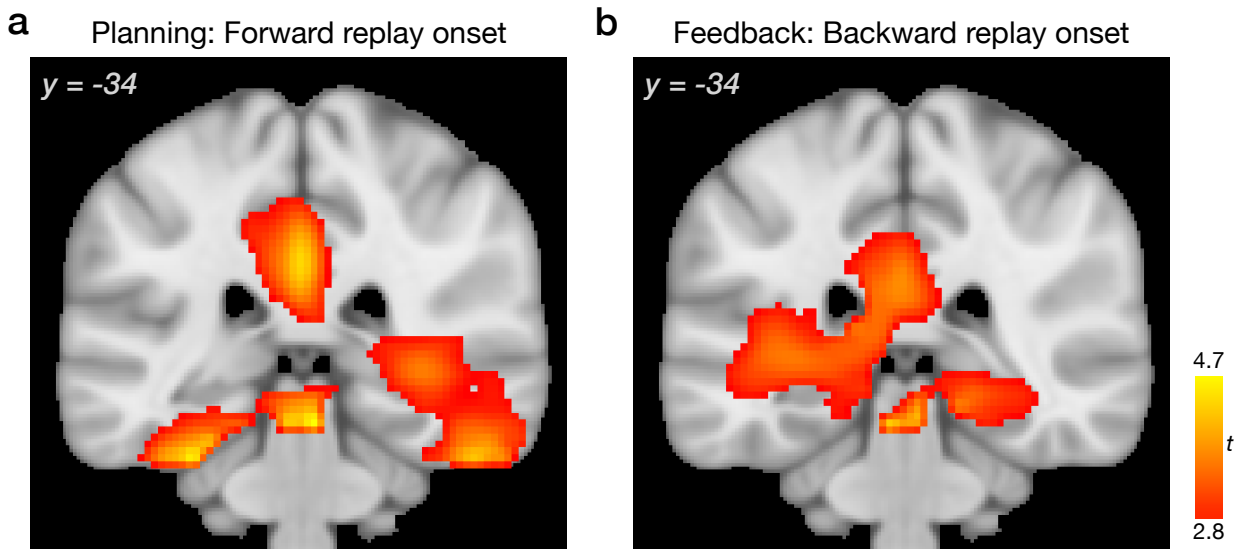

Figure 5. Exploratory replay onset beamforming analyses. (**a**) In the planning period, beamforming analyses revealed power increases associated with replay onset in the right MTL, including the hippocampus. (**b**) After reward feedback, power increases associated with replay onset were found in the bilateral MTL, including the hippocampus. See also **Fig. S5** and **Table S4**. The y coordinate refers to the Montreal Neurological Institute (MNI) atlas. For display, statistical maps were thresholded at P < 0.01 uncorrected; clusters significant at P < 0.05, whole-brain corrected using non-parametric permutation tests. For unthresholded statistical maps and results within the hippocampus ROI mask, see https://neurovault.org/collections/11163/.

Generalized position representations across worlds.

Finally, we asked whether learning led to changes in neural representations that reflected abstract, generalized, information about task structure (Liu et al., 2019; Sun et al., 2020; Nour et al., 2021). We predicted that neural representations of stimuli in the same path position (1, 2, and 3) would become more similar after learning (Liu et al., 2019; Nour et al., 2021), an abstraction which might aid planning. We used Representational Similarity Analysis (RSA) to index representation changes from the pre-task localizer to reward learning path navigation (Deuker et al., 2016; Diedrichsen and Kriegeskorte, 2017; Nour et al., 2021).

When path stimuli were presented during learning, we found evidence for a significant representation of position information (**Fig. S8**; p< 0.05 FWE corrected). In

particular, a position representation was evident for stimuli presented in different worlds ('across-world'), consistent with position information generalizing across stimuli shown on separate trials. Thus, for example, stimuli occupying position 1 in world 1 showed greater similarity to stimuli in position 1 in world 2 than to stimuli in positions 2 or 3 in world 2. The effect was evident in an initial peak from 180–250ms ($p < 0.0001$ uncorrected), where across-world position information was present for all three individual positions when examined separately (p-values < 0.0012). Our design did not include a post-learning localizer, so these representations were necessarily assessed during reward learning. Because of this, it is possible that some position representation information during the learning task could arise from a preceding trial phase, or indeed from cognitive expectation effects. We note that while previous findings showed encoding of position information for stimuli interleaved during learning (Liu et al., 2019; Nour et al., 2021), our results demonstrate that position information generalizes across structurally similar environments even where stimuli never overlapped.

We also investigated path identity representations. For decision-making, it can be helpful to differentiate between grouped stimuli so as to facilitate distinct reactivation during learning; alternatively, it may be helpful to increase similarity for grouped stimuli to allow for linking or chunking. Supporting a differentiation effect, path information emerged at 800ms after stimulus onset, as reflected in a significant decrease in similarity for stimuli in the same path (800-910ms, $p < 0.05$ FWE corrected; peak 820ms, $p = p = 0.00038$ uncorrected; **Fig. S8**). We did not find any relationship between position or path representation strength and behavioural or other neural measures; these null effects could partly be due to noise added by cognitive expectation effects during learning.

Discussion

Two proposed roles for neural replay relate to prospective decision-making and memory preservation. We show that during decision-making, replay strength was related to the relative benefit of model-based, goal-directed, control of behavior. By contrast, after outcome feedback, replay of alternative environment paths positively related to the rarity

in recent experience of a more distal environment. Furthermore, consistent with a putative role in memory preservation, stronger replay following feedback related to a subsequent decrease in behavioural choice uncertainty for the alternative environment. Thus, we find selective links between replay strength and the benefits of planning and the recency of experience, demonstrating distinct roles for replay within a single task.

By manipulating the benefit derived from model-based generalization of reward value across trials, we identified a relationship between planning replay and a reliance on model-based inference. Building on previous imaging work which linked future state reactivation and individual variability in model-based behavior (Doll et al., 2015; Wise et al., 2021) we show here, on a trial-by-trial basis, a boosting of replay when making a choice in a different, but functionally equivalent, state as on the preceding trial, potentially supporting value inference. As planning-related replay onsets were associated with power increases in a region of the MTL consistent with a localization to the hippocampus, this suggests a contribution of the hippocampus to model-based inference, as previously demonstrated in lesion studies (Miller et al., 2017; Vikbladh et al., 2019).

Model-based generalization can be accomplished using different strategies in addition to planning, such as updating ‘non-local’ options following feedback (Konovalov and Krajbich, 2016; Mattar and Daw, 2018; Liu et al., 2021b) or by forming overlapping representations through extensive experience. To illuminate these aspects of planning, our design included two unique features. First, unlike a majority of learning experiments which repeat the same environment on each and every trial (e.g. Daw et al., 2011; Konovalov and Krajbich, 2016), we employed two separate environments, intermixed in an unpredictable manner across trials, thereby limiting the utility of planning for the next choice immediately after feedback. Furthermore, in our design, two alternative start states in an environment converge upon shared paths at the very first step, potentially increasing the degree of inference required during planning to differentiate between trajectories. These two features, environment alternation and early path convergence, distinguish the current task from a recent report that focused on feedback linked replay

signals (Liu et al., 2021b), which revealed that generalization involved updating different non-local start states after feedback. Here, by contrast, we found no significant reward- or planning-related responses after feedback either for our replay signatures of interest, or for the signature identified in Liu et al. (2021b; **Supp. Results**). Thus, our current results, and those of Liu et al. (2021b), both demonstrate a link between neural replay and model-based inference, but at different time periods. It is often the case that the decisions we face arise unpredictably, and our results here support the idea that replay is of benefit in such situations.

To successfully generalize reward feedback and perform well in our task, participants need a model-based strategy. However, similar to a design used in a recent related report (Liu et al., 2021b), choices were only made at the first level, and thus evaluation of path steps was not strictly necessary for model-based behavior. Nevertheless, we found robust evidence for path replay, identifying a new link between replay and model-based decision-making (planning). Further, this relationship was specific to sequential replay and we found no effects related to the reactivation of individual states. We speculate that in many environments outside the lab, replay-assisted planning of trajectories is advantageous, and may be a default strategy employed even when not strictly advantageous. Future experiments could usefully study these neural mechanisms in environments where sequential step-by-step choices are required, although a previous study in this domain did not identify connections to participants' behavior (Kurth-Nelson et al., 2016). Similarly, to link sequential replay to value-guided choice, it will be important for future experiments to explore whether pre-choice replay events are linked to value estimates. Our task was not optimized for value decoding. However, in simpler contexts, replay has been linked to outcome representation during a post-learning rest period in humans (Liu et al., 2019), with similar links reported between hippocampal activity and activity in the ventral striatum or ventral tegmental area in rodents (Gomperts et al., 2015; Sosa et al., 2020).

In contrast to planning, a rest period after outcome feedback entails minimal cognitive demands with respect to the current environment, rendering it likely that activity at this

time-point might support preservation of weaker memories (Olafsdottir et al., 2018). Our results, selective to the feedback period, are consistent with this. Previous studies of hippocampal replay in rodents have suggested a link to less recent experiences (Gupta et al., 2010; Carey et al., 2019), though in these studies experience was confounded with low value, and in one case, a shift in replay was observed even before a reduction in experience (Carey et al., 2019). By parametrically varying experience in two distinct environments, we provide a quantitative link between replay and infrequent experiences. Such a neural replay mechanism can act to reinforce memories that are at risk of becoming weaker, effectively serving as internally-generated interleaved training (Kumaran et al., 2016). While we found a link to memory preservation during ongoing behaviour, we speculate that such a mechanism may also operate offline for memory consolidation.

Our experiment necessitated robust task structure knowledge and very high memory performance for the sequential paths. Behaviourally, previous research has shown that a strong understanding of task structure promotes model-based learning while, conversely, a poor understanding leads to idiosyncratic learning strategies (Feher da Silva and Hare, 2020). One potential limitation of high memory performance is that this precludes linking memory and post-feedback replay. However, during reward learning, a memory preservation account does not necessarily predict a simple relationship between the two: if internal evidence of weak memory drives higher post-feedback replay, an effective replay mechanism may remediate any memory compromise before it is observed. A feature of our design is that it allows us to identify a neural mechanism that may naturally assist in maintaining high memory levels for distant experiences. Further, our reinforcement learning models throw light on a link between trial-to-trial strength in feedback replay and lower noise in the following choice. In general, we speculate that replay-supported memory preservation also supports adaptive behaviour in situations outside the laboratory where learning is more variable, and these are avenues to explore in future experiments.

We suggest that there may be a trade-off between the two separate functions of planning and memory preservation that we identify. One idea is that the content and function of replay is modulated by task demands, which differ between planning versus resting after feedback (Olafsdottir et al., 2018). Consistent with this, it has been reported that replay content that is temporally proximal to active navigation is task-directed, while replay content during reward consumption is undirected, potentially related to preserving memory of the entire environment (Olafsdottir et al., 2017). Computationally, our results suggest the expression of replay may reflect an arbitration of resources as a function of a reliance on planning versus memory (Kali and Dayan, 2004). In line with such a tradeoff, we found that across-participant strength of the memory replay effect related to a lower expression of planning period forward replay. Such arbitration between functions would also influence the degree to which replay supports updating of values, as in a recently proposed computational model of replay (Mattar and Daw, 2018; Liu et al., 2021b). We did not formally manipulate task demands such as cognitive load during planning, but our results provide potential pointers for future targeted studies (Otto et al., 2013).

In conclusion, we provide evidence that prospective replay is enhanced when model-based behavior is beneficial, while replay consistent with memory preservation is observed when demands are low, consistent with distinct signatures for key proposed functions of neural replay (Wilson and McNaughton, 1994; McClelland et al., 1995; Buckner, 2010; Carr et al., 2011; Dolan and Dayan, 2013; Kumaran et al., 2016; van de Ven et al., 2016; Pezzulo et al., 2017; Schapiro et al., 2018; Pezzulo et al., 2019; van de Ven et al., 2020; McNamee et al., 2021). While we dissociate these functions, both planning and memory functions are necessary for adaptive behavior, not least because a stable memory of the one's environment aids successful decision making (Behrens et al., 2018; Joo and Frank, 2018; Olafsdottir et al., 2018). By identifying replay signatures for planning and recency of experience, our results have relevance for targeting an understanding of common or separable disruptions to these functions in psychiatric disorders and disorders that impact on memory, such as Alzheimer's disease (Brewin et

al., 2010; Suh et al., 2013; Jones et al., 2019; Huys et al., 2021; Nour et al., 2021; Prince et al., 2021).

Acknowledgments: The authors thank Rani Moran for helpful discussions, Matt Nour for assistance with analysis, and Daniel Bates and the Imaging Support team at WCHN for assistance with scanning. This work was supported by a Wellcome Investigator Award (098362/Z/12/Z) to R.J.D. G.E.W is supported by a research fellowship from the Deutsche Forschungsgemeinschaft (DFG) and an MRC Career Development Award (MR/V032429/1). Y.L. is supported by the Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning. D.M is supported by a Sir Henry Wellcome Trust Postdoctoral Research Fellowship (110257/Z/15/Z). The Max Planck University College London Centre is a joint initiative supported by University College London and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

This research was funded in whole, or in part, by the Wellcome Trust 098362/Z/12/Z. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions: G.E.W., Y.L., and D.C.M. designed the experiment. G.E.W. collected and analyzed the data. G.E.W and Y.L. wrote the analysis code. G.E.W, Y.L., D.C.M and R.D. contributed to data interpretation. G.E.W. wrote the paper with Y.L., D.C.M., and R.J.D.

Competing interests: The authors declare no competing interests.

Methods

Twenty-seven healthy volunteers participated in the experiment. Participants were recruited from the UCL Psychology and Language Sciences SONA database and from a group of volunteers who had participated in previous MEG studies. Of this participant group, the MEG session was not conducted in three participants: one due to scheduling conflicts, one due to technical problems with the MEG scanner, and one due to poor performance in the behavioral training session (see below). This resulted in the inclusion of data from 24 participants for behavioral and MEG analyses (14 female; mean age 23.8 years (range 18-34). Participants were required to meet the following criteria: age between 18-35, fluent English speaker, normal or corrected-to-normal vision, without current neurological or psychiatric disorders, no non-removable metal, and no participation in an MRI scan in the two days preceding the MEG session. The study was approved by the University College London Research Ethics Committee (Approval ID Number: 9929/002). All participants provided written informed consent before the experiment. Participants were paid for their time, for their performance in the reward learning task as well as memory for the state-to-state sequences (up to £10 based on percent correct performance above chance), and a bonus for performance in the localizer phase target detection task (up to £2).

Experimental task

We designed our reward learning experiment to investigate the potential role of replay in prospective planning and memory preservation (**Fig. 1a**). We adapted a reward-based learning task used in previous experiments to target model-based decision making (Doll et al., 2015; Kool et al., 2016), itself an adaptation of a common ‘two-step’ task (Daw et al., 2011). In this version of the task, there is an equivalence between the two alternative top-level start states in a world. This equivalence provides us with the ability to dissociate model-based and model-free behavior. An additional condition is that potential reward points drift across trials at a sufficient rate that allows for model-free and model-based expectations to often differ. We refer to the decision process as ‘planning’. Closely related studies (Otto et al., 2013; Miller et al., 2017; Vikbladh et al.,

2019) have defined planning as the engagement of model-based decision-making, and our results are consistent with a strong role for model-based decision-making. Of note, and similar to the current design, Miller et al. (2017) use the term ‘planning’ in a two-step task where a decision is made solely at the top level.

Additionally, this version utilizes deterministic transitions between states. The deterministic version of the task has been shown to both incentivize and increase model-based behavior (Kool et al., 2016; Kool et al., 2017). A task using probabilistic transitions would have decreased our ability to detect clear evidence of sequential reactivation, given the increase in number of transitions to evaluate and a constrained number of total trials, as reasonable scanning duration was already maximized. This reasoning also led to using a task with a non-branching path structure (i.e. where states involved no forced choices). We also constrained choice points and branching paths in order to decrease the number of total states, which maintained our ability to sufficiently decode visual states using state-of-the-art imaging technology and analysis techniques (Liu et al., 2021a). Importantly, this limitation of branching choices is similar to the environments in the majority of rodent studies of replay, where animals navigate linear tracks with few, if any, subsequent choices (e.g. Diba and Buzsaki, 2007; Jadhav et al., 2012; Ambrose et al., 2016; Papale et al., 2016; Wu et al., 2017; Carey et al., 2019; Shin et al., 2019; Xu et al., 2019). While these situations do not strictly require sequential evaluation (Liu et al., 2021b), such human and animal designs are able to take advantage of any default tendency to utilize such a neural mechanism. Finally, as measures of model-based behavior are dependent on start state alone, the current task omitted choice at the terminal state, allowing participants to better track rewards and thereby likely increase model-based behavior (following recent work; Gillan et al., 2015; Kool et al., 2016; Kool et al., 2017; Miller et al., 2017).

We augmented this task by adding a second multi-step structure or ‘world’ in order to study the relationship between neural replay signatures and memory preservation (**Fig. 1a**). Further, by employing two worlds we 1) decreased the predictability of the upcoming trial, which in turn increased the utility of planning at choice onset versus

immediately after feedback (Konovalov and Krajbich, 2016), and 2) decreased the dependence of reward learning on short-term working memory for immediately preceding feedback (Collins and Frank, 2012; Wimmer et al., 2018).

We ensured that participants understood the structure of the task in four ways, as poor understanding of a related paradigm has been reported to result in apparent model-free behavior (Feher da Silva and Hare, 2020). First, we provided extensive instructions to minimize misunderstanding. Second, we employed an initial non-reinforced structure exploration learning phase. Third, across all sessions, accurate knowledge of task structure itself was incentivized by periodic memory test trials that allowed participants to harvest additional payment. Fourth, participants learned the abstract task structure during an initial training session (with different stimuli). This session was followed by a break before the MEG session, allowing for robust learning of the structure prior to scanning and potentially allowing for memory consolidation. We conducted the first session on a preceding day for almost all participants (n = 22, range 1-13 days) or after a 2-hour break in two participants, resulting in a 2 day median separation between sessions. Overall, these features helped ensure that behavior was quite model-based, as our goal was to understand planning. Further, by studying behavior reliant on well-learned structure knowledge, our results may better connect to how memory is utilized outside the lab.

Session 1 behavioral procedure

Instructions

In the first experimental session, participants were first instructed in the two-world task, then completed a structure learning phase, and finally completed a brief reward learning phase. This session utilized the same abstract structure as the subsequent MEG session but the structures were populated with different visual stimuli.

Participants were given detailed instructions about the reward learning task and the underlying ‘world’ structure in order to maximize understanding. In brief, they were first instructed that on each trial they would face a choice between two different shape options. Each shape would lead to a different path made up of three sequentially-

presented images. Paths ended in reward points, which ranged from 0-9, and these reward points would drift over time. Their goal was to choose the shape that led down the path to the currently greater number of reward points to earn more bonus money in the experiment. They were also told that along with the presentation of shapes, they would see either '1x' or '5x' above the shape options, which indicated whether the end reward points on this trial would be multiplied by 1 or by 5 (Kool et al., 2016).

Next, participants were told that there were actually two pairs of shape options that would converge to lead down the same two paths and reach the same source of rewards. The shapes in the two pairs were equivalent, in that if shape 1a in start state A led to the path with a snowflake, shape 1b in start state B would also lead to the path with the snowflake (**Fig. 1a**). Importantly, participants were instructed that the rewards at the end of the paths would be reached by either of the potential starting shapes (e.g. 1a and 1b). Consequently, if they were first choosing between the shapes in start state A and found a high reward after choosing shape 1a, then if they were subsequently starting with pair B, they should choose shape 1b in order to get to the same source of reward as they had just experienced when they chose shape 1a. Participants can only accomplish this by leveraging their knowledge of the task structure, classifying this as model-based behavior.

Finally, participants were instructed that these shapes and paths made up 1 'world' and that the real experiment would have two independent worlds, each with the same structure. Trials would start with a pair of shapes from one of the two worlds at random. In general, participants' goal was to keep track of what paths led to the highest rewards at a given time and to choose the shape that led to those paths, while at the same time maintaining their memory for the paths in each world.

Structure learning session 1

Before starting the full reward learning task with point feedback, participants were given the opportunity to learn the structure of the worlds. This phase was composed of two blocks of 20 trials, with learning incentivized by rewarding performance on memory questions about the structure of the worlds. Trials started pseudorandomly in one of the four potential start states across the two worlds.

Participants' goal in this structure learning phase was to explore the different paths in order to learn the sequence of images that followed each shape. Trial events were the same as in reward learning phase trials (see below; **Fig. 1b**; **Fig. S1a**) with the exception that no reward points were shown at the end of the paths and no stakes information was shown during shape presentation. In each structure learning trial, after a planning period, participants made a choice between two shape options (shown randomly on the left and right of the screen). After this selection, the three following states in the path associated with that shape were presented sequentially. Participants were instructed to remember the complete sequence from the chosen shape through to the third path picture. Each path stimulus was randomly presented on the left or right side of the screen and participants needed to press the corresponding 1 or 2 button to indicate the stimulus location on the screen. In this phase, no reward feedback was presented. A fixation cross was presented during a 4-6 s inter-trial interval (ITI). (See detailed timing in the MEG reward learning phase description.)

After each of the two structure learning blocks of 20 trials, participants completed eight memory test probe trials (**Fig. S1b**). Each of the eight start shapes cued one probe trial. On a memory test trial, a single shape was presented on the left or right side of the screen. After the participant selected the shape, they were presented with four potential stimuli from the first state of each of the four paths. Participants selected the stimulus that came next using the 1-4 buttons. After framing the selected stimulus in blue for 0.25 s, this probe structure was repeated for the second and third states in the path. At each stage, one of the four stimuli was correct, while the other three stimuli came from the same state (first, second, or third) across the other three paths. After the response for the third path stimulus, participants were asked to rate their confidence in their set of answers according to the following scale and indicated button response: "Guess (1) Low (2) Medium (3) Certain (4)". Correct performance was based on accurately selecting the correct picture at each of the three stages. Structure memory performance increased from 56.9% (range, 0-100) after the first 20 learning trials to 90.1% (range, 12.5-100) after the second 20 learning trials. Similarly, mean confidence ratings increased from 2.89 to 3.65.

Reward learning session 1

Next, participants engaged in a short reward learning phase to provide experience in maximizing reward earnings. The reward learning phase was the same as the scanned reward learning phase in session 2 (below), with the exception that structure memory questions were pseudo-randomly interleaved with the regular reward learning trials instead of being segregated to breaks between blocks. Trials started pseudorandomly in one of the four potential start states across the two worlds. All trials proceeded in the same way as trials in the preceding structure learning phase, with the addition of reward feedback at the end of each path as well as cued stakes information during planning and choice (**Fig. 1b**; **Fig. S1a**). Reward feedback was presented after a 2 s inter-stimulus-interval (ISI). Reward points flickered in brightness for a period of 1.5 s and then a 3-5 s blank inter-trial interval followed, a slight difference in procedure from the subsequent MEG session. (See detailed timing in the MEG reward learning phase description.)

The reward learning phase length in session 1 was initially based on the free time remaining in the scheduled session but was then set to be a maximum of 40 trials, resulting in a mean of 50 trials across participants (range, 29-105 trials). The memory probe questions were made more difficult in the reward learning phase than the structure learning phase by randomizing the incorrect lure stimuli to be from any stage and any path. Performance on the interleaved structure memory probes was 92.2% (range 70-100%). As noted above, one participant was not invited for MEG scanning based on very poor session 1 memory probe performance (40% correct).

Session 2 MEG procedure

Localizer

After initial setup in the MEG room, participants were given instructions for the localizer phase. The purpose of the functional localizer was to derive participant-specific sensor patterns that discriminated each of the 12 stimuli that made up the world paths by presenting each stimulus many times. The localizer design was adapted from those

used previously, where a picture name identification task followed the presentation of each picture stimulus (Liu et al., 2021b). Participants were instructed to pay attention to a picture shown in the center of the screen and then after the picture disappeared, to select the correct name for the picture from two alternatives. The instructions were followed by 12 practice trials. At the mid-point of each localizer block, a 20 s rest was inserted (after which participants pressed a key to continue). After each block, performance was indicated by 1-4 yellow stars indicating performance >65% correct through >95% correct. If performance was lower than 65%, participants were shown the text: “Please try to perform better in the next part!”

In a localizer trial, participants were presented with a stimulus in the center of the screen for 0.85 s. Then the stimulus disappeared and two words appeared, one above the screen midpoint and the other below. One of the words correctly named the preceding object, while the other was incorrect; the top / bottom locations of the correct and incorrect word were randomized. Participants were instructed to select the correct word using the 1st or 2nd buttons on a 4-button response pad, which was rotated in orientation such that the 1st button was above the 2nd. Words were presented for 0.45 s, followed by a fixation ITI during which button responses for the words were still recorded. The ITI duration mean was 1.2 s (range 0.7 – 2.7 s). If performance on the naming task fell below 70% correct (across missed responses and false alarms in the preceding 32 trials), a warning was presented: “Please increase your performance on the picture identification!”

The stimulus images were presented in a pseudo-random order, with the constraint that no stimulus repeat in subsequent trials. Each stimulus was presented 50 times. The localizer was divided into 4 blocks, with 150 trials per block for a total of 600 trials.

Structure learning session 2

Participants then engaged in a structure learning phase where the new stimuli from the preceding localizer populated the two worlds. As during structure learning in session 1, no rewards were present at the end of the paths. This phase was composed of two blocks of 20 trials, with the four potential start states across the two worlds

pseudorandomly interleaved. Participants' goal was to explore the different paths in order to learn the sequence of images that followed each shape. They were reminded, based on the previous day's experience, that the task contained two 'worlds' with two pairs of starting shape options in each world. Further, each pair of shape options led to two different paths, and two pairs of shape options converged onto the same two paths.

As in the subsequent reward learning trials (**Fig. 1b**; **Fig. S1a**), participants made a choice between two shape options. After this selection, the three following states in the path associated with that shape were presented sequentially. Each stimulus was randomly presented on the left or right side of the screen and participants needed to press the corresponding 1 or 2 button to indicate location. After each path, no reward feedback was presented; a 4-5 s fixation cross inter-trial interval divided trials. Please see the detailed description of a reward learning phase trial below for additional timing information.

After each of the two structure learning blocks of 20 trials, memory test trials were presented following the same procedure as session 1 (**Fig. S1b**). For analyses, an accurate trial included a correct response at each of the three states. In the structure learning phase, participants explored all of the eight potential shape-path combinations (most-explored path per-participant, mean 6.8 trials out of 40 total [range 6.0 – 8.4]; least-explored path per-participant, mean 3.3 trials [range 2.0 – 4.2]).

Reward learning session 2

Participants then engaged in the primary reward learning phase. The design of this phase followed that of the preceding structure learning phase, with the addition of drifting point rewards at the end of the paths (**Fig. 1b**; **Fig. S1a**). Participants aimed to earn as many points as possible in the two different worlds. This phase was composed of 6 blocks of 24 trials for MEG data collection, yielding 144 total reward learning trials.

A reward learning trial began with a 2.5 s planning period where a response was not allowed. To indicate this period, a black cross was presented in the center of the screen. Above the two shape options, the reward stakes on the current trial (1x or 5x) were shown. When the fixation cross disappeared, participants could select a shape for a maximum of 1.5 s. Note that this enforced delay before choice execution limits the

ability to interrogate choice reaction time. If no choice response was recorded for this choice, the trial ended with text indicating a loss of nine points: “Too late or wrong key! -9 pts”. After a 0.5 s inter-stimulus interval (ISI) the stimulus from the first path state was presented randomly on the left or right side of the screen for a minimum of 0.5 s up until a response was recorded, or 1.5 s max. Participants responded with a button corresponding to the screen location of the stimulus. After a 0.5 s ISI, this procedure was repeated for the second and third path state stimuli. If participants failed to respond to any path stimuli, the trial ended with the above no-response error message. Either a 0.5 s or 7.0 s ISI preceded the feedback period. This pre-feedback delay in one world (randomly assigned) was always 0.5 s while the delay in the other world was 7.0 s. (This delay had no effect on behavioral or neural results and thus all analyses combine results across worlds.) The feedback points were then presented in colored text determined by the point amount, where a bright green color was used for the maximum amount of 9 points and bright red color for the minimum amount of 0 points (with intermediate point values colored along a green-to-red gradient). Below the display of the points, the total point value accounting for the stakes multiplier was shown. The feedback text initially flickered in brightness for 0.75 s. Next, the text faded away across a period of 2.25 s after which the screen was blank for the remaining ITI of 2-3 s, for a total feedback period of 5-6 s.

We generated a set of counterbalanced lists in order to decrease variability between participants. Trial lists and counterbalancing assignments for the task were generated for reward point drifts, world order, start state order within world, stimulus assignments to states, shape assignment to paths, and stakes. First, reward points drifted from trial-to-trial according to a process using a standard deviation of 1.75 and reflecting boundaries of 0 and 9. Additional criteria included a mean point value between 4.45 and 4.55 for each path in each half of the task and a weak negative correlation between reward points in a given world of between -0.175 and -0.225. Two pairs of reward point drifts of 72 trials were generated, corresponding to the two worlds. One of the paths in each world was initialized with a start value of 8 while the other was initialized with a start value of 2. The reward point drifts were counterbalanced across worlds across participants. Second, world order was pseudorandomized with

predominant alternation across trials and a maximum repetition of the same world of 3 trials. Third, within a given world, the start state was pseudorandomized with predominant alternation of start state across trials and a maximum repetition of the same start state in a world of 4 trials. Fourth, stimuli were assigned to states based on four counterbalance lists. Fifth, shapes were assigned to paths based on five counterbalance lists. Finally, per-trial stakes were pseudorandomly assigned with a maximum repetition of the same stakes value of 4 trials. For the cued reward stakes manipulation, unfortunately, we identified an unintended significant correlation between stakes and estimated option values (**Supp. Results**). Because of this limitation, we could not clearly assess any relationships between stakes and behavior or neural activity. Further, as we found no benefit from including stakes information in the learning models (**Supp. Results**), the effect of stakes was not included in any behavioral or MEG analyses; i.e. any analyses based on reward points used the unmodulated point value.

At the end of each block of 24 learning trials, participants engaged in a set of memory probe trials during a break in MEG data acquisition. The first two learning blocks were each followed by 8 memory probe questions to ensure the existence of robust structure knowledge, while the remaining four blocks were followed by 5 memory probe questions. As in session 1, the memory probe questions were made more difficult in the reward learning phase than the structure learning phase by randomizing the incorrect lure stimuli to be from any stage and any path. (In the first two participants, confidence ratings were not collected.) See the preceding description in session 1 for memory probe trial timing information. After the memory questions, participants could rest and stretch until ready for the next block.

In the first two participants, only five learning blocks were collected and the learning list also varied, resulting in different numbers of trials with MEG data (100 and 88). In these participants, the memory test questions were randomly interspersed between choice trials instead of being segregated into mini-blocks. In the third participant, only five learning blocks were collected, leading to 120 trials with MEG data. All analyses were adjusted to account for the difference in total trials.

For 21 participants, an additional localizer scan was collected after the reward learning blocks; these data are not analyzed here. Following scanning, all participants then completed a brief written post-experiment questionnaire.

Behavioral analysis

Learning behavior was analyzed using computational models, following prior studies (e.g. Daw et al., 2011; Doll et al., 2015; Kool et al., 2016). To verify learning and determine how reward influenced choice, we used computational models that seek to explain a series of choices in terms of previous events. First, we used logistic regression models, which test for local trial-to-trial adjustments in behavior guided by minimal assumptions about their form, yet at the same time qualitatively capture aspects of model-free and model-based behavior. We then used Q-learning reinforcement learning models, which use a structured set of assumptions to capture longer-term coupling between events and choices and can capture model-free and model-based contributions.

Primary analyses examined behavior independent of whether a trial (or trials) from the alternative world were interleaved between current world trials. All missed response trials (where no response was recorded within the response time window; mean = 3.75 trials per participant) were excluded from the below analyses. In the stay and switch analyses, the previous non-missed trial was counted as the last choice. For reinforcement learning models, the Q-value estimates were carried forward, ignoring missed trials.

Regression

In the regression analysis, we used logistic regression to account for each participant's sequence of choices. For all regression analyses, we focus on the influence of preceding events within the same world (ignoring intervening trials related to the other world). We used multi-level regression functions implemented in R (lme from the nlme package for linear regression; glmmTMB from the glmmTMB package for logistic regression). All predictors and interactions were included as random effects, following

the ‘maximal’ approach (Barr et al., 2013). Correlations between random effects were included when convergence was achievable with this structure.

The first model predicted stay choices based on previous reward and whether the current start state was the same or different (Kool et al., 2016). Here, if participants receive a relatively high reward on the preceding trial, it is generally advantageous in the current task to stay with that choice on the current trial. Conversely, if participants receive a relatively low reward, it is generally advantageous to switch to choosing the alternative option. Thus, the strength of this reward effect on stay choices indexes how well participants were guided by preceding reward. Critically, to test whether experience in one start state is carried over to influence actions in the other start state, our model included a binary variable representing same versus different start state. Any generalization was captured in our model by including the interaction between the previous reward variable and the same start state variable. This measures to what degree, if any, an individual is more influenced by preceding reward when a trial begins in the same start state as the previous trial. Such an influence is characteristic of model-free behavior. Alternatively, given that generalizing across equivalent start states requires knowledge of the structure of the task, the lack of an interaction supports the existence of model-based behavior.

A second model examined how option selection was influenced by the choice and reward received on the previous trial. Instead of predicting stay decisions, this model predicts option choice (with options arbitrarily coded as 0 and 1) using the history of preceding rewards and option choices (Lau and Glimcher, 2005; Wimmer et al., 2014; Doll et al., 2015). (Such models approximate reinforcement learning models when longer histories of previous choices and rewards are included, with the decaying influence of previous events reflecting learning rate.) Critically, in addition to preceding choices and rewards, to assess model-based behavior, the model also included a variable representing whether the current start state was the same or different than the preceding trial. As in the first analysis, the interaction of this same start state variable with previous trial reward and choice is able to capture any model-free influence on behavior (Doll et al., 2015). As above, if there is a stronger influence of previous events on choice when the start state is the same, this provides evidence for a model-free

component to behavior. Conversely, the lack of such an interaction supports model-based behavior.

Supplemental regression analyses were conducted within-participants to analyze individual differences in model-based generalization; here, individual participant fits were derived from the bayesglm function in the arm package to constrain extreme coefficients in one participant.

Reinforcement learning

In the reinforcement learning analysis, we fit three different Q-learning reinforcement learning (RL) models to subjects choice behavior (Sutton and Barto, 1998): model-free, model-based, and hybrid models. These models generated participant-specific model parameters and model fit values (log likelihood). These models also allow us to compute trial-by-trial variables related to choice value and reward feedback for use in neural analyses.

The model-free reinforcement learning model learns to assign an independent action value to each state of the task. For simplicity, we assign model-free values to the path as a whole, without considering the value of each intermediate state (see discussion below), yielding two "states" in the experiment: the choice state (state 1) and the path (state 2). In this way, the model below is equivalent to and based on the model applied to two-step tasks lacking the intermediate path states (Kool et al., 2016). Further, with no second stage choice, there are only two actions, a_1 and $a_{2,}$, with one action selected in the start state. Model-free values are updated at stage *i* and trial *t* according to a prediction error δ, modulated by a learning rate α (with range [0 1]):

$$Q_{MF}(s,a) = Q_{MF}(s,a) + \alpha\, \delta_{i,t} \qquad (1)$$

For the model-free strategy, the prediction errors differ after moving from the start states (A and B) to one of the paths (states X1 or X2), and from a path to a reward. The values of the start state options are first updated according to the difference in value between the start and path states:

$$\delta_{1,t} = Q_{MF}(s_{X,t}, a_{1,t}) - Q_{MF}(s_{1,t}, a_{1,t}) \quad (2)$$

After reward feedback (re-scaled for modeling to a range of [0 1]), the values for the path states are updated according to the difference in value between the path state and the reward received:

$$\delta_{2,t} = r_{2,t} - Q_{MF}(s_{X,t}, a_{1,t}) \quad (3)$$

Additionally, the model-free values for the start states are updated after the received reward, modulated by a fractional eligibility trace parameter e (with range [0 1]):

$$\delta_{1,t} = e\,(r_{X,t} - Q_{MF}(s_{X,t}, a_{1,t})) \quad (4)$$

This allows for start states to be updated by the reward received after the second state.

The model-based strategy utilizes the learned path (state 2) model-free Q-values in combination with the learned transition matrix between task states.

$$Q_{MB}(s_A, a_j) = P(s_X \mid s_A, a_j)\, Q_{MF}(s_{X1}) + P(s_X \mid s_B, a_j)\, Q_{MF}(s_{X2}) \quad (5)$$

The hybrid reinforcement learning model combines model-free and model-based learning models and computes per-trial option values based on a weighted sum of the two estimates. Our model was based on that used by Kool et al. (2016), where the values input to the softmax choice rule are calculated according to:

$$Q_{net}(s_A, a_j) = \omega Q_{MB}(s_A, a_j) + (1 - \omega) Q_{MF}(s_A, a_j) \quad (6)$$

Given value estimates on a particular trial where participants were choosing between two options, participants are assumed to stochastically with probabilities according to a softmax distribution (Daw et al., 2006):

$$P(s_{i,t}) = \exp(\beta(Q_{net}\ s_{i,t},a) / \sum_{a'} (\exp(\beta(Q_{net}\ s_{i,t},a')) \quad (7)$$

The free parameter β represents the softmax inverse temperature (with range [0 Inf]), which controls the exclusivity with which choices are focused on the highest-valued option. This exclusivity can reflect certainty in option estimates. Note that since the softmax is also the link function for the logistic regression model discussed above, this analysis also has the form of a regression from Q values onto choices except here, rather than as linear effects, the past rewards enter via the recursive learning of Q, controlled, in nonlinear fashion, by the learning rate parameters.

The purely model-based RL model is a reduced form of the hybrid model where the weighting parameter $\omega = 1$ and the eligibility parameter e have no effect and are dropped. Conversely, the purely model-free RL model is a reduced form of the hybrid model where $\omega = 0$.

We found that models with decaying ("forgetting") Q-values for non-experienced states provided a better fit. Specifically, one parameter decayed the value of non-chosen Q values. In the task, some decay of non-chosen Q values is rational, as the reward values drift for both chosen and non-chosen terminal states. A second parameter decayed Q-values for the non-experienced world. The decay parameters were constrained to the range [0 1]. The parameters decayed values to the median reward value of 0.5. Thus, if the non-experienced world decay parameter was less than 1, when both Q-values were above or below 0.5, both values moved toward the median while the distance between values was maintained. In the situation where Q-values were above and below the median, the distance between values decreased.

Thus, on each trial, the value for the non-chosen Q-values was decayed according to the fractional parameter τ_{ALT}:

$$Q_{MF}(s_i,a_{nonchosen}) = ((Q_{MF}(s_i,a_{nonchosen}) - 0.5) * \tau_{ALT}) + 0.5 \quad (8)$$

Separately, on each trial, all values in the non-experienced world were decayed according to the fractional parameter τ_W:

$$Q_{MF}(s_i,a_j) = (Q_{MF} (s_i,a_j) - 0.5) * \tau_W) + 0.5 \quad (9)$$

Parameters were optimized for each subject using an optimization routine based on the fmincon in Matlab. Option values were initialized with mid-range values of 0.5, which also provided the best fit to the data.

To generate per-participant, per-trial variables based on the reinforcement learning models, the models were simulated on each participant's sequence of experiences with their best-fit parameters.

Reinforcement learning and memory preservation

To test if feedback period other world replay was directly related to the quality of subsequent choice behavior, providing additional evidence of a role in memory preservation, we estimated an additional reinforcement learning model augmented by neural replay. Enhanced memory preservation was expected to benefit the representations of the other world structure and associated values, which could be related to decreased uncertainty on subsequent choices. Conversely, memory decay was expected to be associated with increased uncertainty. To capture a preservation-related decrease in uncertainty in our reinforcement learning model, we focused on the softmax inverse temperature associated with the translation of learned values into choice. While this parameter is often interpreted as a signal reflecting noise or uncertainty specific to choice processes, this parameter can also reflect longer-term shifts in the uncertainty of underlying memory representations. In our best fit model-based RL model, the uncertainty at choice is over the prospective value representations, where the effect of uncertainty is a function of the uncertainty of memory for 1) the links between the shape options and the paths, 2) the paths and terminal values, and 3) the terminal values themselves. The softmax inverse temperature can be viewed as effectively controlling the width of the distribution of the

estimates of the terminal values. With a high inverse temperature, uncertainty (the width of the distribution) is low and choices are strongly influenced by the stored value estimates. Conversely, with a low inverse temperature, choices become more random and undirected. We predicted that transient increases in memory preservation would be associated with decreased uncertainty, captured by a higher inverse temperature.

We extracted trial-by-trial sequenceness values from the preceding feedback period and the planning period. The modulation of uncertainty applied to trials where the 'world' changed from a preceding trial (e.g. from world 1 to world 2, or vice versa). It is only on 'world change' trials where preceding "other world" backward replay represents the path options that are under consideration in the current trial. To determine trials with relatively high backward replay, the sequenceness data were thresholded at the 60th percentile to ensure sufficient choice trials for RL parameter estimation. The threshold was determined by stepping down until the inclusion of at least 10 trials per participant where there was strong replay (minimum number of trials, 11; median, 27). For the replay strength variable, a 1 represented the presence of strong replay and 0 represented the absence. In control analyses, we also examined alternative replay variables derived from the preceding feedback phase as well as the current planning period.

The RL memory preservation model was built upon the above RL model, augmented with two additional softmax inverse temperature parameters. The baseline β applied to trials where the world was the same as the previous trial. $\beta_{wchange}$ applied to trials where the world changed from the previous trial but where there was no strong replay. $\beta_{wchange_replay}$ applied to trials where the world changed from the previous trial and there was strong replay. We predicted that $\beta_{wchange_replay}$ would be higher than $\beta_{wchange}$. The choice equations for the augmented model are as follows. For world-same trials, choice was determined by Eq. 7:

World-change trials where $replay_{t-1} = 0$ (low):

$$P(s_{i,t}) = \exp(\beta_{wchange}(Q_{net}\ s_{i,t},a) / \sum_{a'} (\exp(\beta_{wchange}(Q_{net}\ s_{i,t},a')) \quad (10)$$

World-change trials where $replay_{t-1} = 1$ (high):

$$P(s_{i,t}) = \exp(\beta_{wchange_replay}(Q_{net}\ s_{i,t},a) / \sum_{a'} (\exp(\beta_{wchange_replay}(Q_{net}\ s_{i,t},a')) \qquad (11)$$

Rarity of experience

To examine potential memory preservation effects after reward feedback, we computed a variable representing the relative inverse frequency ('rarity') of the alternative world. Frequency of experience was an exponentially-weighted measure computed for each world on a trial-by-trial basis. This measure was calculated by adapting the fractional updating in Equation 2 to track world frequency based on appearance (1) or non-appearance (0) of a world on a given trial. We expected the learning rate controlling the change in estimates of world frequency to be relatively slow based on related work (Bornstein and Daw, 2013), which led us to set the experience learning rate to 0.10. The resulting experience frequency values were then inverted to provide the infrequency or 'rarity' of each world at each trial. A subsequent control analysis made no assumptions about learning rate, simply testing whether other world replay was stronger when that world had been experienced on the last trial or not.

MEG acquisition

The participants were scanned while sitting upright inside an MEG scanner located at the Wellcome Centre for Human Neuroimaging at UCL. A whole-head axial gradiometer MEG system operating at third order synthetic gradiometry configuration (CTF Omega, VSM MedTech) recorded data continuously at 1200 samples per second, utilizing 272 channels (3 channels of the original 275 channels are not included due to excessive noise in routine testing). Three head position indicator coils were used to locate the position of the participant's head in the three-dimensional space with respect to the MEG sensor array. They were placed on the three fiducial points: the nasion and left and right pre-auricular areas. The coils generate a small magnetic field which is used to localize the head and enable continuous movement tracking. We also used an Eyelink eye-tracking system to record participant's eye movements and blinks; the MEG implementation uses no additional head restraint or support. Eye-tracking was

calibrated prior to scanning and data was stored alongside the MEG data in three additional channels at the MEG sampling rate. The task was projected onto a screen suspended in front of the participants. The participants responded during the task using a 4-button response pad to provide their answers (Current Designs), responding with self-selected digits to the first and second buttons. Trial events were written into the MEG data via a TTL parallel port. Event timing was verified by concurrent photodiode recording of trial-event-associated brightness changes in a bottom-left location of the projected screen (not visible to participants); this channel was also written into the MEG data.

MEG Pre-processing

MEG data were processed in MATLAB using the packages SPM12 (Wellcome Trust Centre for Neuroimaging) and FieldTrip, following previous procedures (Liu et al., 2019; Wimmer et al., 2020). See https://github.com/gewimmer-neuro/multistep_replay for an example preprocessing script. The CTF data were imported using the OSL package (the OHBA Software Library, from OHBA Analysis Group, OHBA, Oxford, UK). Slow drift was removed by applying a first-order IIR high-pass filter at 0.5 Hz. The data were down-sampled (including anti-aliasing low-pass filter) from 1200 Hz to 100 Hz (for sequenceness analyses, yielding 10 ms per sample) or 400 Hz (for time-frequency analyses) for improved signal to noise ratio and to conserve processing time.

Preprocessing was conducted separately for each block. An initial preprocessing step in OSL identified potential bad channels whose signal characteristics fell outside the normal distribution of values for all sensors. Then independent component analysis (FastICA, http://research.ics.aalto.fi/ica/fastica) was used to decompose the sensor data for each session into 150 temporally independent components and associated sensor topographies. Data from the three eye-tracking channels was also included and flagged for artifact monitoring. Artifact components were classified by automated inspection of the combined spatial topography, time course, kurtosis of the time course, and frequency spectrum for all components. For example, eye-blink artifacts exhibit high kurtosis (>20), a repeated pattern in the time course and consistent spatial topographies. Mains interference has extremely low kurtosis and a frequency spectrum

dominated by 50 Hz line noise. Artifacts were then removed by subtracting them out of the data. All subsequent analyses were performed directly on the filtered, cleaned MEG signal from each sensor, in units of femtotesla.

The MEG data were epoched to extract data for each trial. To allow participants to shift from a goal of structure learning to a goal of value-directed choice, the first 4 trials of the reward learning task were excluded from sequenceness analyses; this exclusion had no qualitative effect on the results. In the localizer blocks, a 2.5 s epoch was extracted in each trial, encompassing 0.5 s preceding stimulus onset and continuing past the stimulus and word response. In the pre-choice planning period, we extracted epochs of 2.5 s at the beginning of each trial. In the post-feedback reward period we extracted epochs of 3.5 s, following the first 1.5 s after reward. At this stage, in the epoched data, we further excluded time periods within individual channels that exhibited extreme outlier events in a given trial epoch (defined by values > 7x the mean absolute deviation) by replacing these values with zero.

For planning analyses, we used the 2.5 s pre-choice planning period after excluding the first 160 ms to allow for early visual stimulus processing, following a related previous experiment (Wimmer et al., 2020). For memory analyses in the post-feedback period, where our prediction was that memory processes would be engaged when other cognitive demands are relatively low, with similarities to a procedural step in a related rodents study (Olafsdottir et al., 2017), we focused on the time period following initial reward processing (the latter 3.5 s of the 5 s period). Here we expected that the demands of actual feedback processing and value updating would preclude the engagement of any memory preservation signal. However, we note that our results remain qualitatively the same even without this early time period exclusion step.

In the initial scanned structure learning phase, only 40 trials were presented. To increase power, our sequenceness analyses included all valid paths in the current and other worlds. The contrast of late versus early sequenceness subtracted the mean across the first 20 trials from the second 20 trials. In this phase, the participants' goal was exploration and learning; no reward feedback was presented. Thus, the inter-trial-interval was analogous to the feedback period in the reward learning phase. For MEG

analyses, we examined the 4 s inter-trial-interval period beginning at the offset of the third and final path stimulus.

MEG data decoding and cross-validation

Lasso-regularized logistic regression models were trained for each of the 12 stimuli from the paths. Methods followed previous studies (Kurth-Nelson et al., 2016; Liu et al., 2019; Wimmer et al., 2020). Only the sensors that were not rejected across all scanning sessions in preprocessing were used to train the decoding models for a given participant. A trained model k consisted of a single vector with length of good sensors n consisting of 1 slope coefficient for each of the sensors together with an intercept coefficient. Decoding models were trained on MEG data elicited by direct presentations of the visual stimuli. Our experimental task was not optimized to detect reactivation of expected value or reward point outcome representations, and as a consequence our analyses focus only on stimuli.

For each stimulus we trained one binomial classifier. Positive examples for the classifier were trials on which that stimulus was presented. Negative examples consisted of two kinds of data: trials when another stimulus was presented, and data from the fixation period preceding stimulus onset. The null data were included to reduce the correlation between different classifiers – enabling all classifiers to report low probabilities simultaneously. Prediction accuracy was estimated by treating the highest probability output among all classifiers as the predicted stimulus. In the functional localizer task, prediction performance of classifiers trained at each 10 ms bin from 0 ms to 800 ms after stimulus onset and performance was tested iteratively on left-out trials. The range of peak classifier performance across participants was then identified, with the 200 ms time period selected among the peak times for all subsequent analyses based on based previous replay experiments (see Results) (Kurth-Nelson et al., 2016; Liu et al., 2019; Wimmer et al., 2020). A new classifier trained at this time point on all the localizer data trials was estimated for use in subsequent decoding analyses.

In a control analysis, prior to the sequential analyses, we examined whether we could find characteristics suggestive of single stimulus reactivation in the reward learning task periods of interest. This control is supplemental, as we do not identify

specific events in any of our analyses and the features of the human MEG signal are best analyzed by regression analyses using all the data in periods of interest (Liu et al., 2021a). For this reactivation analysis, we compared the 12 stimulus time series derived from applying the trained stimulus classifiers to the trial-wise task planning and feedback period data (when no stimuli were actually being presented) to time series derived from applying permuted classifiers. To create permuted classifier data, within each participant and for each of the 12 trained stimulus classifiers, we took the original trained sensor weights (a 1-dimensional vector) and randomized the vector weightings across the non-zero-weighted sensors (preserving the use of informative sensors).

We compared two measures of the resulting time series: mean and skewness. In general, if true spike-like events are added to a random vector, the positive spike deviations will increase the mean signal, and – independent of the mean – they will also increase skewness. Comparing our real data to the permuted data, we indeed found significantly higher mean values in the real versus permuted data (averaged across all stimuli, as we found equal reactivation for current and other word stimuli; p-values < 0.01). Second, we found significantly increased skewness in the real versus permuted data (p-values < 0.01). These differences were found in both the planning period and the feedback period, where none of the relevant trained stimuli were actually presented.

Sequenceness measure

The decoding model described above allowed us to measure spontaneous sequential reactivation of the 12 states either during the planning or after feedback periods. We applied each of the 12 trained classifiers to the MEG data at each time point in each period. This yielded a [time x state] reactivation probability matrix for each period in each trial, containing twelve timeseries of reactivation probabilities each with the length of time samples included in the analysis window.

We then used the Temporally Delayed Linear Modelling framework to quantify evidence of ‘sequenceness’ (Liu et al., 2021a), which describes the degree to which these representations were reactivated in a task-defined sequential order (Kurth-Nelson et al., 2016; Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a; Liu et al., 2021b). TDLM is a multiple linear regression approach that quantifies the degree to which a

time-lagged reactivation time course of state *j*, ($X(\Delta t)_j$, where Δt indicates lag time) predicts the reactivation time-course of state *i*, (X_i). It involves two stages. At the first stage, we use a set of multiple linear regression models to generate the empirical state-to-state reactivation pattern, using each state's ($j \in$ [1: 12])reactivation time course as a dependent variable, and the historical (i.e., time-lagged) reactivation time courses of all states ($i \in$ [1: 12]) as predictor variables. Separate linear models were estimated for each stimulus *i* and each time lag Δt:

$$X_j = \sum_{i=1}^{12} X(\Delta t) i \beta(\Delta t)_{ij} + C \tag{12}$$

where *C* is a constant term and $\beta(\Delta t)_{ij}$ is the coefficient capturing the unique variance in X_i explained by $X(\Delta t)_j$. The resulting regression coefficients quantify the evidence for each empirical state -> state reactivation pattern at a specific lag, Δt. We calculated sequenceness from time lag 10 ms to 600 ms. All such first-level coefficients are in a lag-specific empirical transition matrix $\boldsymbol{B}$, representing evidence for all possible state-to-state transitions at a given time lag.

At the second stage, we quantified the evidence that the empirical transition matrix $\boldsymbol{B}$, can be predicted by the sequences of interest, i.e., the 4 paths across both worlds in the task. All transitions of interest were specified in model transition matrices (of size 12x12, with 1s for transitions of interest and 0s otherwise), separately for a forward direction (T_F, the same as visual experience) and the inverse for a backward direction (T_B), where $T_F = T_B'$. As control variables, the regression included a constant matrix (T_{cons}) that captures the average of all transitions, ensuring that any identified effects were not due to background neural dynamics, and a matrix (T_{auto}) that models self-transitions to control for auto-correlation. The evidence for all sequences of interest was then quantified by:

$$\boldsymbol{B} = \sum_{r} Z(r) * T_r \qquad (13)$$

where r is the number of all regressors included in the second stage, as specified above. Z is the scalar regression coefficient quantifying the evidence that the hypothesized transitions, T_r predict the empirical transitions, ***B*** (i.e. sequenceness strength). Note, this estimate of sequence strength is a relative quantity. An estimate of zero for state *i* to state *j*, for example, does not mean there is no replay of *i* -> *j*, it means there is no stronger replay of i -> j, than that of other transitions. Repeating the regression of Equation 13 at each time lag (Δt = 10, 20, 30, ... , 600 ms) results in time courses of both forward and backward sequence strength as a function of time lag, where smaller lags indicate greater time-compression of replay.

A subsequent model substituted the full transition matrices with matrices that separated the paths from the two worlds (current, other) in the forward ($T_{Fcurrent}$, T_{Fother}) and backward ($T_{Bcurrent}$, T_{Bother}) direction. Per-trial results were z-scored across lags; note that this had no qualitative effect on the results. As human studies of sequential state reactivation have only found evidence for replay at relatively short lags (e.g. Kurth-Nelson et al., 2016; Liu et al., 2019), our analyses only considered lags up to 350 ms. Finally, whilst the sequenceness method was originally tested on longer time periods > 10 s (Kurth-Nelson et al., 2016; Liu et al., 2019), recent work has demonstrated its applicability to shorter time periods (Wimmer et al., 2020; Liu et al., 2021b), where simulations have verified an ability of the method to discriminate between true sequences and noise. Note, unlike rodent electrophysiology, we cannot detect discrete replay events given that we are not measuring spiking data, but an LFP-like continuous signal. This is another reason we utilize linear modelling to assess the sequence strength on average across a trial period.

To ensure that the classification results were not overfit to the regularization parameter of the logistic regression, all results were obtained with the lasso regularization parameter that yielded the strongest mean cross-decoding from the localizer to the sequential presentation of the actual stimuli in the reward learning phase. This independent cross-decoding approach yielded the same parameter value

(l1 = 0.002) as we found previously (Wimmer et al., 2020). As our analysis used trial-based data and not rest, we did not find a strong alpha rhythm and did not control for alpha in the sequenceness analyses (Liu et al., 2019; Wimmer et al., 2020).

To localize forward and backward time lags of interest for the primary trial-by-trial regression analyses, we identified significant lags using permutation tests with shuffled transition matrices. Shuffled transition matrices were randomly generated for each trial to follow the structure of the true matrix, including only invalid transitions. Further, to match the structure of the true paths, each shuffled transition matrix was required to have four independent paths. Thus, each shuffled matrix contained eight entries (two per path). This ensured that each stimulus would appear in a sequence a maximum of once per permutation and that true sequential triplets were formed across states. We ran 100 permutations of this procedure, averaging results across trials within each permutation within each participant. The significance threshold was determined from the resulting sequenceness values in two steps. First, for each permutation, we computed the maximum permutation-derived sequenceness value across all time lags. Then, across these set of 100 maximum values, we applied a 95% threshold to determine the permutation significance threshold. This approach has been validated in both simulation and empirical data (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a).

Identifying Replay Onsets

For follow-up source localization and time-frequency analyses, replay onsets were defined as moments when a strong reactivation of a stimulus was followed by a strong reactivation of the next (or preceding) stimulus in the sequence from a path (Liu et al., 2019; Wimmer et al., 2020). As described in the Results, a different direction and lag was the focus of the planning and reward feedback period analyses. We identified the stimulus-to-stimulus time lag Δt at which there was maximum evidence for sequenceness (as described above). We first generated a matrix Orig as

$$Orig = X \times T \tag{14}$$

where X is the [time x state] reactivation matrix, and T is the task transition matrix. The transition matrix T defines the mapping between the task state corresponding to column i in X, and column i in Orig (specifically, column i in Orig is the reactivation time course of the state that 'precedes' state i in T). We then shifted each column of X by the relevant replay lag, Δt, to generate another matrix Proj,

$$Proj = X(\Delta t) \tag{15}$$

where row i of Proj corresponds to row i+[lag] ms of X. Multiplying Proj and Orig elementwise, and summing over the columns of the resulting matrix, therefore yields a [time x 1] vector, R, where each element, t, corresponds to the evidence for a two-state replay with given lag, starting from any task state at time t.

$$R_t = \sum_i^s Orig_{ti} * Proj_{ti} \tag{16}$$

We then identified putative replay onsets. Within-participants, we thresholded *R* at its 95th percentile to only include high-magnitude putative replay onset events. We also imposed the constraint that a replay onset event must be preceded by a 200 ms pre-onset baseline period exhibiting summed reactivation evidence < 90th percentile at each time point.

MEG Source Reconstruction

Source reconstruction (beamforming) was performed in SPM12 and FieldTrip utilizing OAT. Forward models were generated on the basis of a single shell using superposition of basis functions that approximately corresponded to the plane tangential to the MEG sensor array.

Linearly constrained minimum variance beamforming (Van Veen et al., 1997) was used to reconstruct the epoched MEG data to a grid in MNI space, sampled with a grid step of 5 mm. The sensor covariance matrix for beamforming was estimated using the preprocessed data in broadband power across all frequencies in the lower range 1.5–47.3 Hz.

For the replay onsets analysis, we computed baseline activity as the mean power averaged over -100 ms to -50 ms relative to replay onset. All non-artifactual trials were baseline corrected at source level. We looked at the main effect of the initialization of replay. This analysis was conducted separately to investigate forward replay events in the planning period and backward replay events in the feedback period. To increase power, primary analyses examined putative replay onset events across both worlds, while secondary analyses examined current world replay onsets during planning and other world replay onsets during feedback.

The statistical significance of clusters identified in the beamforming analysis was calculated using non-parametric permutation tests in OSL to identify clusters significant at $P_{FWE} < 0.05$ (whole-brain corrected; cluster-defining threshold $p < 0.01$ ($t = 2.81$); 5000 permutations). We expect power increases in source space to reflect increases in underlying neural activity. Replay onsets are defined by high evidence for internal generation of a visual stimulus. As expected, previous reports find robust replay-onset-related activity in the visual cortex in addition to the MTL (Wimmer et al., 2020; Liu et al., 2021b). Further, this internally-generated activity overlaps with that detected for actual stimulus presentation (Wimmer et al., 2020) and generally corresponds to regions showing increased visual-evoked signal in fMRI and electrophysiological studies. Finally, we assume that continuous regions of activity extending anterior from the visual cortex to the medial temporal lobe would reflect the same positive direction of underlying neural activity.

Time-frequency analyses

A frequency decomposition (wavelet transformation) was computed separately for the planning and reward feedback period in every trial. From this data, we extracted power changes surrounding the putative replay onset events (± 200 ms) identified using the preceding method. To increase power, primary analyses examined putative replay onset events across both worlds, while secondary analyses examined current world replay onsets during planning and other world replay onsets during feedback. We examined evidence across participants for (mean) power increase at replay onset compared to a pre-onset baseline (from –200 to –150ms before replay onset) within a

frequency range of interest of 120-150 Hz, in the ripple band that has been previously associated with replay events (Liu et al., 2019; Liu et al., 2021b; Nour et al., 2021) as well as a theta band range of interest (8-12 Hz). Within these bands of interest, we tested individual difference correlations between replay onset (± 2.5ms) and individual differences in the replay-behaviour relationships previously identified. A cluster-based permutation analysis implemented in OSL separately tested whether there were significant clusters associated with replay onset across the full time-frequency range. The clustering algorithm used an initial threshold of $p < 0.01$ (corresponding to a t-threshold of 2.81) and an alpha level of 0.05.

In control analyses to test for the specificity of our results to sequential replay, we used the above trial-wise frequency decomposition for the planning and reward feedback periods. Within each participant, at each timepoint and frequency, we estimated the correlation across trials with the planning and feedback variables of interest: at planning, the benefit of generalization and state value; at feedback, rarity of other world experience. These results were submitted to cluster permutation analyses as in the preceding replay onset time-frequency analyses.

Non-sequential reactivation analyses

In two additional control analyses, we tested for the specificity of our results to sequential replay. We used the same time by state classifier evidence for the planning and reward feedback periods that also served as input for the sequenceness analyses. The first set of control analyses examined mean classifier activity across the time period, separately for the current and other world, averaging across the states within a path. Additional checks examined each of the three path states separately. As in the sequenceness analysis, these measures reflect the relationship across the full time period of interest (planning, feedback) for each trial. The second set of control analyses examined pair-wise ‘clustered’ reactivation of states present on the same path, following previous work (Wimmer et al., 2020). Clustered reactivation was calculated for each trial as the zero-lag correlation of classifier evidence between two elements of a given path. Correlation values were z-transformed and then these z-values were averaged across

the potential pairs in a path or world. Both of these trial-by-trial control measures were analyzed using multilevel models

In a more temporally fine-grained analysis, we also examined the relationship between reactivation at each time point (in 10 ms resolution) and variables of interest. For statistical comparison, these results were submitted to cluster permutation analyses as used in the preceding replay onset time-frequency analyses.

Representational similarity

We used Representational Similarity Analysis (RSA) (Diedrichsen and Kriegeskorte, 2017) to investigate the emergence of a representation of task structure from the stimulus localizer (pre-learning) to reward learning, following previous procedure (Nour et al., 2021). We utilized this analysis because we lacked a post-task localizer and because RSA may be more sensitive to effects due to abstract structure than sequenceness analyses in short task periods (versus longer rest durations in previous studies). First, for each trial in each session we subtracted the mean of the 100ms preceding stimulus onset from the pre-processed MEG data (baseline correction); for each of the three stimuli presented in reward learning paths, this correction was performed per-stimulus. For each sensor, we then z-scored over all trials and time points relative to stimulus-onset (t, -100 to +1500 ms). We then regressed the [trial x 1] neural data, $Y(s)_t$ (from time point, t, and sensor, s) onto a session design matrix, X, denoting the stimulus label of each trial (dummy coded) (Luyckx et al., 2019),

$$Y(s)_t = X * \beta(s)_t \tag{17}$$

and used the resulting [stimulus x 1] vector of regression weights, $\beta(s)_t$, as an estimate of the unique activation for each stimulus, in sensor s at time point t. Repeating this procedure over all sensors yielded a [sensor x stimulus] matrix at each time point. We then calculated the Pearson correlation distance between the sensor patterns for each pair of stimuli (columns). This generated a symmetrical [12 x 12] Representational Dissimilarity Matrix (RDM) at each time point (Deuker et al., 2016). We conducted this procedure identically in both stimulus localizer (SL) and reward learning (R), enabling us

to calculate the learning-induced change in representational similarity (similarity change) at each time point, ΔS_t, as

$$\Delta S_t = RDM(SL)_t - RDM(R)_t \quad (18)$$

where entry s_{ij} of ΔS_t quantifies the post-learning similarity increase between evoked signals for stimuli i and j, at time t (Deuker et al., 2016). ΔS_t was then smoothed over time with a 50 ms Gaussian kernel (Luyckx et al., 2019).

Finally, we used a second multiple regression to quantify the variance in ΔS_t that was uniquely explained by an abstracted representation of ordinal position for each participant at each time point. The position predictor was separated into two matrices, capturing position similarity across worlds and position similarity within the same world. Path identity was also included in the model. The multiple regression thus included 3 predictor variables corresponding to each of the hypothesized representational patterns, plus an intercept (see **Fig. S8**). Note that motor response to the stimuli was not a confound; participants responded with a button press whether stimuli were presented on the left or right side of the screen, but the left/right position was assigned randomly per trial per stimulus.

We used nonparametric tests to identify time windows (clusters) with significant positive evidence for each predictor, correcting for multiple comparisons over time (Nour et al., 2021). Specifically, for a given predictor (e.g., position), at each time point post-stimulus onset we conducted a separate one-sample t test on the representational effect (regression weights) over participants, to obtain the evidence for an effect > 0 (i.e., t-value). We computed the sum of t-values within each continuous stretch of time points exhibiting a positive effect at $p < 0.05$. We then repeated this procedure for 1000 permutations, on each occasion applying the same novel shuffling to the rows and columns (stimulus labels) of ΔS_t prior to the second regression. The shuffled order was consistent across time within each permutation to preserve temporal smoothness in visually evoked neural data. We then extracted the maximal sum-of-t value for the group mean effect in each permutation (matching the procedure applied to the unpermuted data), to generate an empirical null distribution for the predictor in question. We

expected any position representation effect to be positive. A suprathreshold positive cluster in the unpermuted data was deemed significant at $P_{FWE} < 0.05$ if its sum-of-t values exceeded the 95th percentile of this empirical null distribution (Eldar et al., 2018; Nour et al., 2021). We had no expectations about the direction of any path identity representation effect, so clusters were judged to be significant at $P_{FWE} < 0.05$, two-tailed, if its sum-of-t values exceeded 97.5% (or fell below 2.5%) of this null distribution.

Multilevel regression models

Multilevel models relating sequenceness to trial-by-trial variables of interest were estimated using R, as further detailed above for the behavioral models. In these models, the independent variables were the relevant and control sequenceness strength values, e.g. current and other world paths sequenceness. All relevant sequenceness measures were included in the same model. To allow for valid comparisons, where analyses compared regression coefficients or computed interactions, the relevant variables were z-scored across trials.

Statistical correction and null effects

For permutation-based significance measures of mean overall replay, see the relevant section above. For trial-by-trial analyses, multilevel models were implemented in R, as described above. Other tests of means or fit parameters utilized t tests except where the noted. All reported tests are two-tailed unless otherwise noted. One-tailed tests were used when we had an *a priori* expected direction of a comparison, including modulation of current versus other world replay or planning versus feedback period replay. For the exploratory analyses across additional replay time lags, significance was determined using Bonferroni correction for the number of comparisons (34, where calculated lags ranged from 10 to 350 ms). Contrasts of multilevel model regression coefficients were estimated using the glht function in the multcomp package; note that all variables entered into regressions were z-scored to allow for valid contrasts. For comparisons of softmax inverse temperature parameters, as expected, the distribution of the inverse temperature parameter was not normal, so we used Wilcoxon signed rank tests. For results of interest, 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), as implemented in the TOSTER library in R (Lakens, 2017). We used bounds of Cohen's d = 0.60, where power to detect a medium- or larger-sized effect in the group of 24 participants was estimated to be 80%.

Data availability

Complete behavioral data are publicly available on the Open Science Framework (https://osf.io/szjxp/). The full MEG dataset will be publicly available on openneuro.org.

Code availability

Example code for sequenceness analyses will be available on GitHub (https://github.com/gewimmer-neuro/multistep_replay).

References

Ambrose, R.E., Pfeiffer, B.E., and Foster, D.J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron 91*, 1124-1136.

Barr, D.J., Levy, R., Scheepers, C., and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang 68*.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron 100*, 490-509.

Bornstein, A.M., and Daw, N.D. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLoS Comput Biol 9*, e1003387.

Brewin, C.R., Gregory, J.D., Lipton, M., and Burgess, N. (2010). Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychol Rev 117*, 210-232.

Buckner, R.L. (2010). The role of the hippocampus in prediction and imagination. *Annu Rev Psychol 61*, 27-48, C21-28.

Buzsaki, G. (2015). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus 25*, 1073-1188.

Carey, A.A., Tanaka, Y., and van der Meer, M.A.A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nat Neurosci 22*, 1450-1459.

Carr, M.F., Jadhav, S.P., and Frank, L.M. (2011). Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nat Neurosci 14*, 147-153.

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

Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal replay of extended experience. *Neuron 63*, 497-507.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron 69*, 1204-1215.

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*, 876-879.

Deuker, L., Bellmund, J.L., Navarro Schroder, T., and Doeller, C.F. (2016). An event map of memory space in the hippocampus. *Elife 5*.

Diba, K., and Buzsaki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nat Neurosci 10*, 1241-1242.

Diedrichsen, J., and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput Biol 13*, e1005508.

Dolan, R.J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron 80*, 312-325.

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. *Nat Neurosci 18*, 767-772.

Ego-Stengel, V., and Wilson, M.A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus 20*, 1-10.

Eldar, E., Lievre, G., Dayan, P., and Dolan, R.J. (2020). The roles of online and offline replay in planning. *Elife 9*.

Eldar, E., Roth, C., Dayan, P., and Dolan, R.J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current biology : CB*.

Feher da Silva, C., and Hare, T.A. (2020). Humans primarily use model-based inference in the two-stage task. *Nat Hum Behav*.

Fernandez-Ruiz, A., Oliva, A., Fermino de Oliveira, E., Rocha-Almeida, F., Tingley, D., and Buzsaki, G. (2019). Long-duration hippocampal sharp wave ripples improve memory. *Science 364*, 1082-1086.

Foster, D.J. (2017). Replay comes of age. *Annu Rev Neurosci 40*, 581-602.

Gillan, C.M., Otto, A.R., Phelps, E.A., and Daw, N.D. (2015). Model-based learning protects against forming habits. *Cogn Affect Behav Neurosci*.

Gillespie, A.K., Astudillo Maya, D.A., Denovellis, E.L., Liu, D.F., Kastner, D.B., Coulter, M.E., Roumis, D.K., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. *Neuron*.

Girardeau, G., Benchenane, K., Wiener, S.I., Buzsaki, G., and Zugaro, M.B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nat Neurosci 12*, 1222-1223.

Gomperts, S.N., Kloosterman, F., and Wilson, M.A. (2015). Vta neurons coordinate with the hippocampal reactivation of spatial experience. *Elife 4*.

Gupta, A.S., van der Meer, M.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal replay is not a simple function of experience. *Neuron 65*, 695-705.

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., and Woolrich, M. (2021). Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron 109*, 882-893 e887.

Huys, Q.J.M., Browning, M., Paulus, M.P., and Frank, M.J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology 46*, 3-19.

Jadhav, S.P., Kemere, C., German, P.W., and Frank, L.M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science 336*, 1454-1458.

Johnson, A., and Redish, A.D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *J Neurosci 27*, 12176-12189.

Jones, E.A., Gillespie, A.K., Yoon, S.Y., Frank, L.M., and Huang, Y. (2019). Early hippocampal sharp-wave ripple deficits predict later learning and memory impairments in an alzheimer's disease mouse model. *Cell Rep 29*, 2123-2133 e2124.

Joo, H.R., and Frank, L.M. (2018). The hippocampal sharp wave-ripple in memory retrieval for immediate use and consolidation. *Nat Rev Neurosci 19*, 744-757.

Kali, S., and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat Neurosci 7*, 286-294.

Konovalov, A., and Krajbich, I. (2016). Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nat Commun 7*, 12438.

Kool, W., Cushman, F.A., and Gershman, S.J. (2016). When does model-based control pay off? *PLoS Comput Biol 12*, e1005090.

Kool, W., Gershman, S.J., and Cushman, F.A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol Sci 28*, 1321-1333.

Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn Sci 20*, 512-534.

Kumaran, D., Summerfield, J.J., Hassabis, D., and Maguire, E.A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron 63*, 889-901.

Kurth-Nelson, Z., Economides, M., Dolan, R.J., and 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 meta-analyses. *Soc Psychol Personal Sci 8*, 355-362.

Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav 84*, 555-579.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron 64*, 431-439.

Liu, Y., Dolan, R.J., Higgins, C., Penagos, H., Woolrich, M.W., Olafsdottir, H.F., Barry, C., Kurth-Nelson, Z., and Behrens, T.E. (2021a). Temporally delayed linear modelling (tdlm) measures replay in both animals and humans. *Elife 10*.

Liu, Y., Dolan, R.J., Kurth-Nelson, Z., and Behrens, T. (2019). Human replay spontaneously reorganises experience. *Cell 178*, 640-652 e614.

Liu, Y., Mattar, M.G., Behrens, T.E.J., Daw, N.D., and Dolan, R.J. (2021b). Experience replay is associated with efficient nonlocal learning. *Science 372*.

Lopez-Persem, A., Bastin, J., Petton, M., Abitbol, R., Lehongre, K., Adam, C., Navarro, V., Rheims, S., Kahane, P., Domenech, P., and Pessiglione, M. (2020). Four core properties of the human brain valuation system demonstrated in intracranial signals. *Nat Neurosci 23*, 664-675.

Luyckx, F., Nili, H., Spitzer, B., and Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. *Elife 8*.

Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat Neurosci 21*, 1609-1617.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev 102*, 419-457.

McCloskey, M., and Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychol Learn Motiv 24*, 109-165.

McNamee, D.M., Stachenfeld, K.L., Botvinick, M., and Gershman, S.J. (2021). Flexible modulation of sequence generation in the entorhinal-hippocampal system. *Nat Neurosci*.

Michelmann, S., Staresina, B.P., Bowman, H., and Hanslmayr, S. (2019). Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nat Hum Behav 3*, 143-154.

Miller, K.J., Botvinick, M.M., and Brody, C.D. (2017). Dorsal hippocampus contributes to model-based planning. *Nat Neurosci 20*, 1269-1276.

Nour, M.M., Liu, Y., Arumuham, A., Kurth-Nelson, Z., and Dolan, R.J. (2021). Impaired neural replay of inferred relationships in schizophrenia. *Cell*.

O'Neill, G.C., Barry, D.N., Tierney, T.M., Mellor, S., Maguire, E.A., and Barnes, G.R. (2021). Testing covariance models for meg source reconstruction of hippocampal activity. *Sci Rep 11*, 17615.

Olafsdottir, H.F., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife 4*, e06063.

Olafsdottir, H.F., Bush, D., and Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current biology : CB 28*, R37-R50.

Olafsdottir, H.F., Carpenter, F., and Barry, C. (2017). Task demands predict a dynamic switch in the content of awake hippocampal replay. *Neuron 96*, 925-935 e926.

Otto, A.R., Gershman, S.J., Markman, A.B., and Daw, N.D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol Sci 24*, 751-761.

Papale, A.E., Zielinski, M.C., Frank, L.M., Jadhav, S.P., and Redish, A.D. (2016). Interplay between hippocampal sharp-wave-ripple events and vicarious trial and error behaviors in decision making. *Neuron 92*, 975-982.

Patzelt, E.H., Kool, W., Millner, A.J., and Gershman, S.J. (2019). Incentives boost model-based control across a range of severity on several psychiatric constructs. *Biol Psychiatry 85*, 425-433.

Pezzulo, G., Donnarumma, F., Maisto, D., and Stoianov, I. (2019). Planning at decision time and in the background during spatial navigation. *Curr Opin Beh Sci 29*, 69-76.

Pezzulo, G., Kemere, C., and van der Meer, M.A.A. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Ann N Y Acad Sci 1396*, 144-165.

Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature 497*, 74-79.

Prince, S.M., Paulson, A.L., Jeong, N., Zhang, L., Amigues, S., and Singer, A.C. (2021). Alzheimer's pathology causes impaired inhibitory connections and reactivation of spatial codes during spatial navigation. *Cell Rep 35*, 109008.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychol Rev 97*, 285-308.

Schapiro, A.C., McDevitt, E.A., Rogers, T.T., Mednick, S.C., and Norman, K.A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat Commun 9*, 3920.

Schuck, N.W., and Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science 364*.

Schuirmann, D.J. (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.

Shin, J.D., Tang, W., and Jadhav, S.P. (2019). Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making. *Neuron 104*, 1110-1125 e1117.

Singer, A.C., Carr, M.F., Karlsson, M.P., and Frank, L.M. (2013). Hippocampal swr activity predicts correct decisions during the initial learning of an alternation task. *Neuron 77*, 1163-1173.

Sosa, M., Joo, H.R., and Frank, L.M. (2020). Dorsal and ventral hippocampal sharp-wave ripples activate distinct nucleus accumbens networks. *Neuron 105*, 725-741 e728.

Suh, J., Foster, D.J., Davoudi, H., Wilson, M.A., and Tonegawa, S. (2013). Impaired hippocampal ripple-associated replay in a mouse model of schizophrenia. *Neuron 80*, 484-493.

Sun, C., Yang, W., Martin, J., and Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nat Neurosci 23*, 651-663.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: An introduction (Cambridge: MIT Press).

van de Ven, G.M., Siegelmann, H.T., and Tolias, A.S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nat Commun 11*, 4069.

van de Ven, G.M., Trouche, S., McNamara, C.G., Allen, K., and Dupret, D. (2016). Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron 92*, 968-974.

Van Veen, B.D., Van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on biomedical engineering 44*, 867-880.

Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., and Daw, N.D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron 102*, 683-693 e684.

Wikenheiser, A.M., and Redish, A.D. (2015). Hippocampal theta sequences reflect current goals. *Nat Neurosci 18*, 289-294.

Wilson, M.A., and McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science 265*, 676-679.

Wimmer, G.E., Braun, E.K., Daw, N.D., and Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J Neurosci 34*, 14901-14912.

Wimmer, G.E., Daw, N.D., and Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *Eur J Neurosci 35*, 1092-1104.

Wimmer, G.E., Li, J.K., Gorgolewski, K.J., and Poldrack, R.A. (2018). Reward learning over weeks versus minutes increases the neural representation of value in the human brain. *J Neurosci 38*, 7649-7666.

Wimmer, G.E., Liu, Y., Vehar, N., Behrens, T.J., and Dolan, R.D. (2020). Episodic memory retrieval success is related to rapid replay of episode content. *Nat Neurosci*.

Wise, T., Liu, Y.Z., Chowdhury, F., and Dolan, R.J. (2021). Model-based aversive learning in humans is supported by preferential task state reactivation. *Science Advances 7*.

Wu, C.T., Haggerty, D., Kemere, C., and Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nat Neurosci 20*, 571-580.

Xu, H., Baracskay, P., O'Neill, J., and Csicsvari, J. (2019). Assembly responses of hippocampal ca1 place cells predict learned behavior in goal-directed spatial tasks on the radial eight-arm maze. *Neuron 101*, 119-132 e114.