Distinct replay signatures for planning and memory maintenance

Authors: G. Elliott Wimmer[1,2,*], Yunzhe Liu [3,4,1,2], Daniel McNamee[1,2], 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.

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

Abstract

Theories of neural replay propose that it supports a range of different functions, most prominently planning and memory maintenance. Here, we test the hypothesis that distinct replay signatures relate to planning and memory maintenance. Our reward learning task required human participants to utilize structure knowledge for 'model-based' evaluation, while maintaining knowledge for two independent and randomly alternating task environments. Using magnetoencephalography (MEG) and multivariate analysis, we found neural evidence for compressed forward replay during planning and backward replay following reward feedback. Prospective replay strength was enhanced for the current environment when the benefits of a model-based planning strategy were higher. Following reward receipt, backward replay for the alternative, distal environment was enhanced as a function of decreasing recency of experience for that environment. Consistent with a memory maintenance role, stronger maintenance-related replay was associated with a modulation of subsequent choices. These findings identify distinct replay signatures consistent with key theoretical proposals on planning and memory maintenance functions, with their relative strength modulated by on-going computational and task demands.

Introduction

Humans have a remarkable ability to process information that extends beyond what is immediately perceptible, facilitating simulation of prospective plans and retrieval of past memories. It is hypothesized that hippocampal replay contributes to 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 has now been identified in humans using the high temporal resolution of MEG (Kurth-Nelson et al., 2016; Liu 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 evaluating and planning future actions (Buckner, 2010; Dolan and Dayan, 2013; Pezzulo et al., 2017; Pezzulo et al., 2019) as well as in supporting memory maintenance (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). However, the contribution of awake replay to planning and memory maintenance 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). 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). Here we address the contribution of replay to both roles in a single task context.

Replay of trajectories leading to a goal is proposed to underpin a form of 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).

This contrasts with model-free evaluation which depends solely on repetition of previously rewarded actions. Interestingly, brain lesion studies highlight a necessary role for the hippocampus in model-based behavior (Miller et al., 2017; Vikbladh et al., 2019), suggesting a potential role for hippocampal replay. Thus far, there is no evidence for a link between planning replay and the demands for model-based control (Kurth-Nelson et al., 2016; Eldar et al., 2020; Liu et al., 2021b; Wise et al., 2021). Some 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, given the inconsistency in such findings (Gillespie et al., 2021), it remains possible 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).

Replay is also considered critical for memory maintenance, where replay of previous experiences can strengthen memory, preventing 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 in rodents that involve disrupting hippocampal activity indicate 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 is conjectured that human rest-period replay serves a similar function (Liu et al., 2019; Schuck and Niv, 2019; Nour et al., 2021). Studies of replay in rodents navigating single environments provide initial, but inconclusive, evidence of a link between lesser experience of an environment and replay (Gupta et al., 2010; Carey et al., 2019). Here, we address this question in the context of multiple environments where participants need to maintain memory of a distal environment while at the same time learning in the local one.

To address our questions 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), but now include two independent randomly alternating environments. We used recently-developed MEG analytic methods to 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). During pre-choice planning, we were first interested in whether path replay would increase 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). Following receipt of reward feedback, when there are minimal cognitive demands in relation to the current environment, we hypothesized that replay of the alternative environment would support memory maintenance (Olafsdottir et al., 2018).

Results

Behavior

In the experiment, participants navigated two separate, independent environments (worlds) to earn reward points (**Fig. 1a**). Each world contained two path options leading to separate streams of reward feedback, which drifted over time. Importantly, within a given world, each trial could start in one of two equivalent start states leading to those paths. As demonstrated in previous research on model-based learning (Doll et al., 2015; Kool et al., 2016; Kool et al., 2017; Patzelt et al., 2019), this feature allows us to characterize how well participants use structure knowledge to guide model-based behavior (**Fig. 1c**). To perform well in the task, an agent should generalize what is learned in one start state and apply it to trials starting in the alternative, equivalent, start state. Only a model-based agent can exploit such structure knowledge to allow generalization of reward feedback across equivalent start states. This flexibility is proposed to involve looking ahead to values associated with the terminal states, possibly using prospective neural reactivation or replay (Doll et al., 2015). Conversely, a model-free agent can only update the values of the current start state; thus, reward following one start state will not inform choices in the alternative start state.

Based upon this logic, the task 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 this paradigm that use probabilistic state transitions (Daw et al., 2011; Miller et al., 2017; Vikbladh et al., 2019). One feature of the deterministic task is that trials are divided between those where model-based behavior is more beneficial (different start state) versus less beneficial (same start state) (Doll et al., 2015), allowing us to test for an association with changes in neural replay. Furthermore, a deterministic transition structure and a lack of branching paths increases our ability to detect evidence of sequential neural reactivation (Liu et al., 2021b). Finally, by including multiple randomly alternating worlds, we decrease the predictability of an upcoming trial, promoting the deployment of planning-related processes at choice onset instead of at outcome feedback (Konovalov and Krajbich, 2016; Liu et al., 2021b).

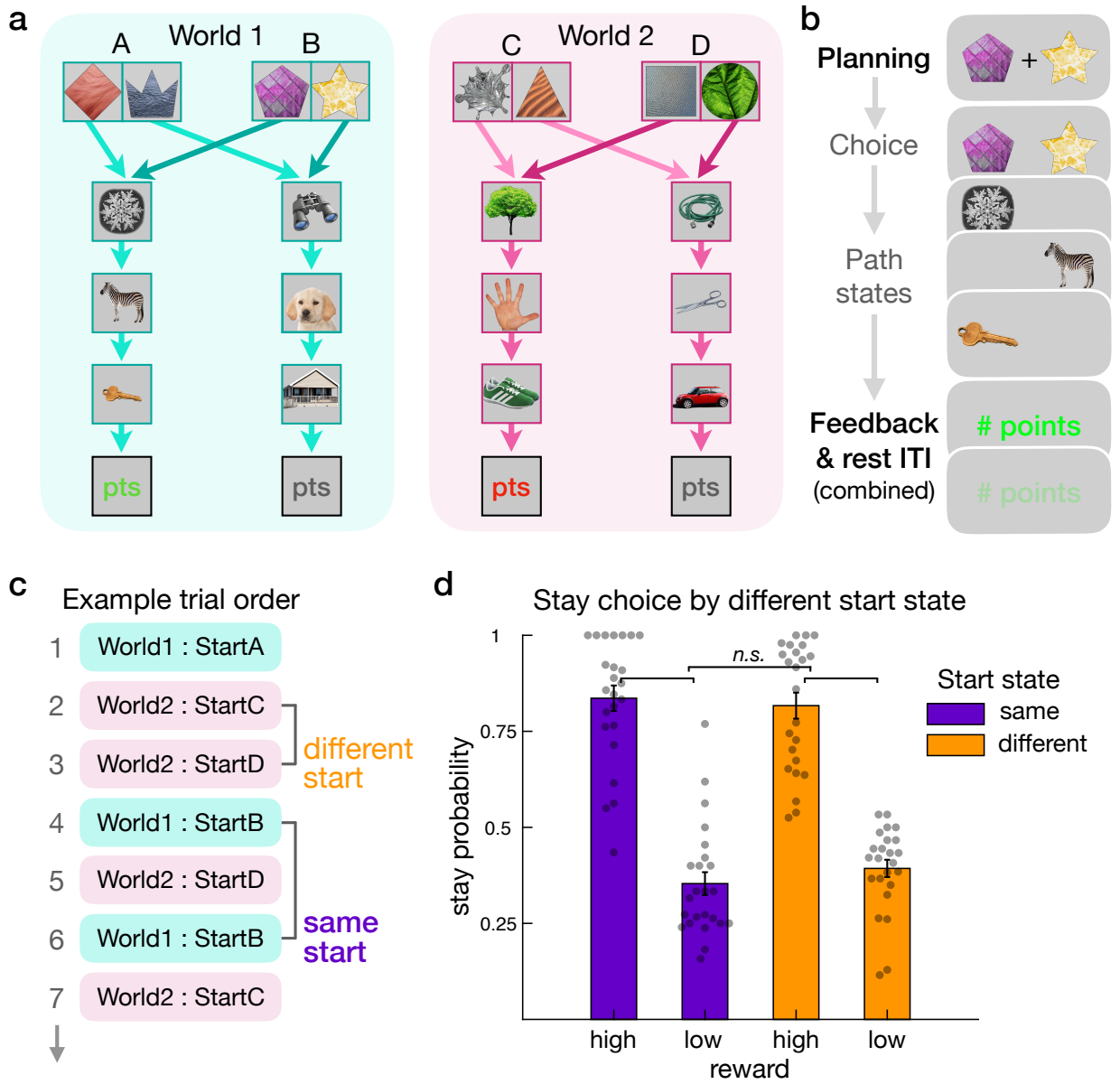

Figure 1. Two-environment reward learning task and generalization behavior. (**a**) Task schematic showing two alternative 'worlds' and their two equivalent start states. Trials in each world start at one of two equivalent pairs of shape options. The shape options then lead deterministically to the same paths and reward outcomes (0-9 points). Following a training session to learn this general structure, on the day of imaging, participants learned two worlds populated with new images. Participants' memory for the path sequences was tested by occasional probe trials, where performance rapidly increased (**Supp. Results**). (**b**) Key trial periods in the reward learning phase. Replay

was measured during a pre-choice planning period. A choice period followed, signaled by the disappearance of the central cross, during which a response was entered. 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 after processing of feedback. 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 different and a same start state, 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 rate of stay choices (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, suggesting behavior was model-based. For display purposes, point feedback was binarized 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 exploration period without reward feedback, participants engaged in 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 previous 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, using structure knowledge to look ahead and evaluate expected terminal rewards. Alternatively, if updating the equivalent start state occurs after feedback (prior to choice) this can be conceptualized as 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 participants' behaviour reflects an exploitation of structure knowledge for a task environment, a hallmark of a model-based contribution to behaviour (Daw et al., 2011; Doll et al., 2015). Qualitatively, we saw that receipt of high versus low reward enhanced the likelihood of choice repetition (**Fig. 1d**) while a high

degree of model-based behavior was indicated by similar 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). This showed a robust overall effect of previous reward on stay choices ($\beta = 0.536$ [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, indicating the absence of a significant model-free contribution (interaction between reward and same start state, $\beta = 0.0313$ [-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 also employed a two-one-sided test (TOST) equivalence procedure that enables us to reject the presence of a medium- or larger-sized effect (Lakens, 2017). Indeed, we can reject a medium- or larger-sized model-free interaction effect (TOST equivalence test $p = 0.027$). Nevertheless, reaction times were slower for choices where generalization was more beneficial (different versus same start state trials $\beta = 0.009$ [0.0017 0.0163]; $z = 2.423$, $p = 0.0155$). In a complimentary 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. As above, this lack of an interaction is an essential signature of model-based learning (Daw et al., 2011; Doll et al., 2015). Overall, we could 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 was 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 setting to next examine the underling neural processes.

Sequenceness identification

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 from 140-210 ms post stimulus onset. Based on this, and consistent 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). This classifier generalized from the localizer to the presentation of path objects during reward learning, showing significant across-phase classification ($t_{(23)} = 7.361$, $p < 10^{-7}$; **Fig. S3**).

We next used the trained classifiers to identify 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. These analyses focused on two periods, namely pre-choice planning and post-feedback rest (**Fig. 1b**). In a next step we then 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. This sequence detection method quantifies the average predictivity of state *i* to state *j* within a period, which reflects both the frequency and fidelity of replay events, an approach validated in our previous work (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021a; Liu et al., 2021b).

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). To increase power in this analysis, 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'. During planning, we found significant forward sequenceness with a state-to-state time lag of 70 and 80 ms indicating 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**). We found no significant evidence for backward sequenceness during planning.

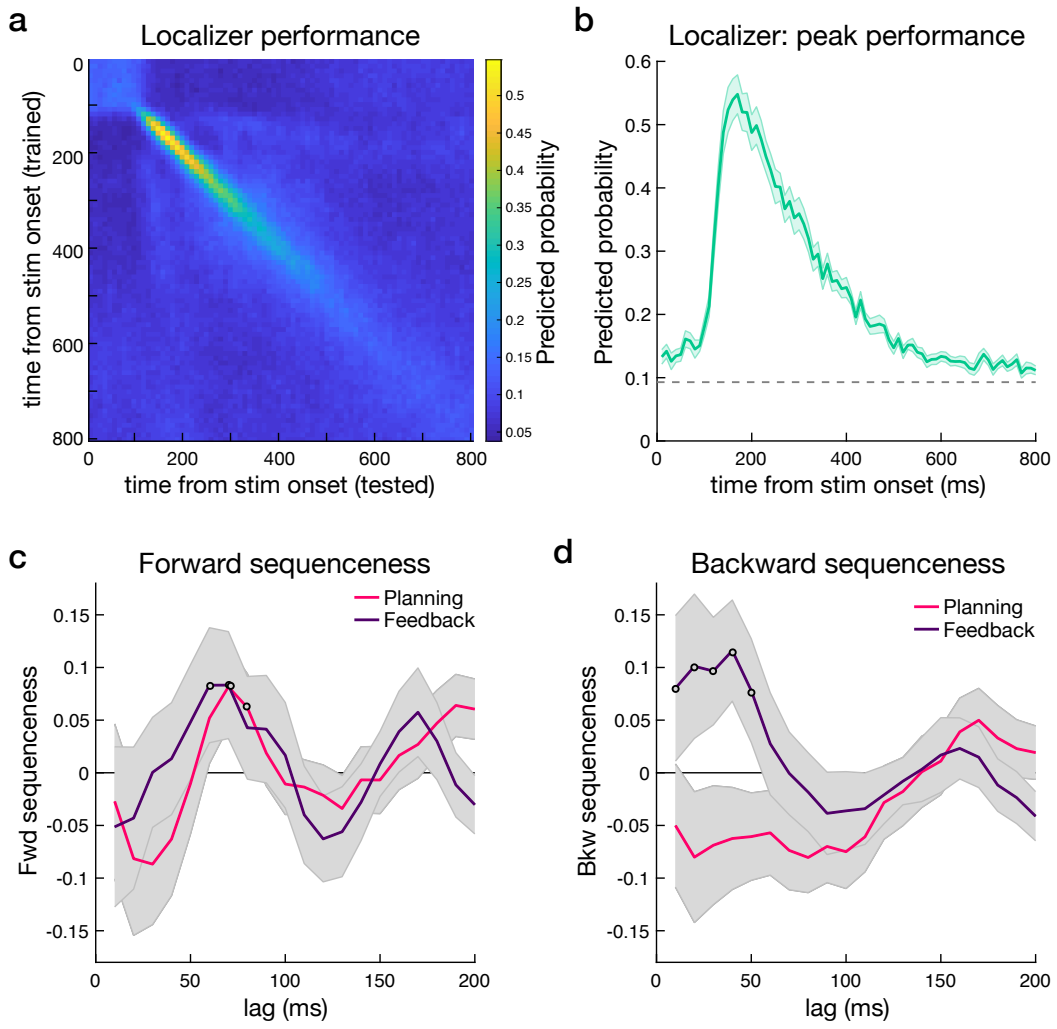

Figure 2. Training of state localizer and sequenceness time lag identification. (**a**) Classifier performance for path state stimuli presented during a pre-task localizer phase, training and testing at all time points. This revealed good discrimination between the 12 path stimuli used in the learning 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, evident at a common state-to-state lag of 70 ms in both trial periods. (**d**) Backward sequenceness, for all learned paths during planning and feedback periods, evident at state-to-state lags that spanned 10-50 ms in feedback period alone. See **Fig. S4** for example sequenceness events. Shaded error margins represent SEM. Open dots in panels c and d indicate time points exceeding a permutation significance threshold.

We then examined replay following feedback, the period when the displayed reward points faded from the screen, leading to a brief inter-trial interval rest (**Fig. 1b**; **Fig. S1a**). This corresponds to a time when replay is often 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 the backward direction, and 60 to 70 ms in the forward direction (**Fig. 2**). To limit effects of initial feedback processing, and in line with a convention in rodent studies (Olafsdottir et al., 2017), our analysis was focused on the latter 3.5 s of the 5 s post-feedback period. Note that qualitatively similar results were found when using the full feedback period. We found that backward replay in the feedback period was significantly stronger than backward replay expressed in the planning period ($t_{(23)}$ = 2.638 [0.039 0.318], p = 0.0147). Notably, relatively stronger backward replay after feedback also aligns with findings in rodents, where backward replay has been associated with reward receipt, while forward replay has been associated with upcoming behavior (Diba and Buzsaki, 2007).

Our subsequent analyses of forward sequenceness focused on the 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 also by similar effects in our previous work (Kurth-Nelson et al., 2016; Liu et al., 2019).

Replay during planning and model-based generalization

To examine a link between replay and task experience, we extracted measures of sequenceness for each period on each trial. We first tested a link between planning and relative sequence strength. We leveraged the fact that in our task model-based generalization was more beneficial on one, but not the other, trial type (**Fig. 1d**) (Doll et al., 2015). A planning account of replay predicts that, if participants use this period to decide, when the start state changes the benefit of model-based generalization should be reflected in a boosting of replay compared 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 our prediction, we found that sequenceness for current world paths was significantly stronger when generalization was more beneficial than when not (multilevel regression $\beta = 0.1411$ [0.0625 0.2197]; $z = 3.519$, $p = 0.0004$; **Fig. 3a**). There was no relationship between sequenceness for other world paths and generalization ($\beta = -0.0201$ [-0.0979 0.0577]; $z = -0.506$, $p = 0.613$; other world TOST equivalence test $p = 0.012$), while the relationship between generalization and sequenceness was significantly stronger for current versus other world paths (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). We then examined whether stronger forward sequenceness related to reduced reaction time slowing for generalization trials. Stronger current world sequenceness was numerically, but not significantly, related to faster reaction times on different versus same start state trials (interaction $\beta = -0.3956$ [-

0.7989 0.0078]; z = -1.922, p = 0.0546). Finally, we found that pre-choice replay was more related to a planning account, involving a simple increase on trials when generalization was more beneficial, compared to an alternative memory-inspired account which would predict a graded modulation as a function of recency in experience of a given start state (**Supp. Results**).

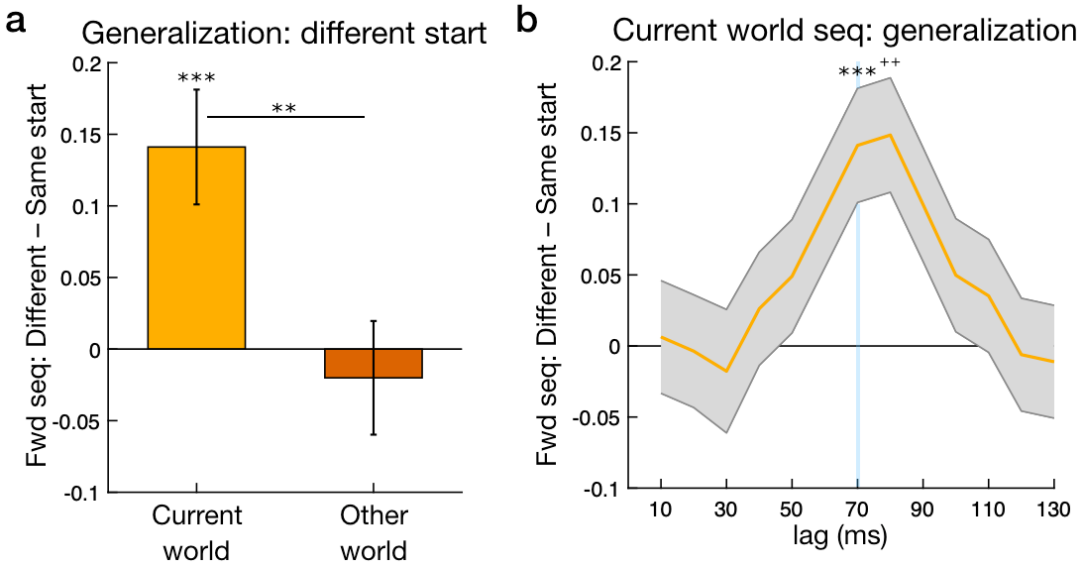

Figure 3. Planning period replay relationship with model-based generalization. (**a**) Stronger forward replay when there is a greater benefit in deploying model-based knowledge: on trials where the start state is different from the previous trial. (**b**) Timecourse of regression coefficients for variables of interest, showing effects at state-to-state lags from 10 to 130 ms, highlighting the 70 ms time lag of interest (light blue line). Y-axes represent sequenceness regression coefficient for binary different versus same start state. Seq = sequenceness. ** p<0.01; *** p<0.001; + p<0.01, corrected for multiple comparisons.

We next examined whether replay related to the dynamics of changing option value across the task. If replay is involved in evaluation, we would not necessarily expect a modulation of replay by value. However, it is possible that replay would be biased by option value when values are more immediately retrievable. Relatedly, we note many previous imaging and electrophysiology studies in humans reported 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).

We examined the relationship between replay and mean state value, the average model-predicted value across the two options. We found that current world replay strength was significantly correlated with mean state value (β = 0.0479 [0.0130 0.0829]; t = 2.688, p = 0.0072). However, this relationship was only found on trials with 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 selective to current world forward replay during planning (**Supp. Results** and **Table S3**). Thus, replay was positively related to option value in conditions where the start state options could have been directly reinforced on the preceding trial, where we speculate value inference is less demanding.

Backward replay prioritization at feedback and memory maintenance

At feedback we hypothesized that replay would reflect maintenance of memory representations of the distal environment. Computationally, a memory maintenance mechanism and related neural signature would be driven by the rarity of recent experience in a given world (McCloskey and Cohen, 1989). As described above, we found significant backward replay in the feedback period alone (**Fig. 2d**), with a 40 ms time lag previously linked to replay during rest in humans (Liu et al., 2019; Nour et al., 2021). In control analyses, we found no relationship between backward replay and reward feedback or reward prediction error (**Supp. Results** and **Table S3**).

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 β = 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 β = -0.0101 [-0.0307 0.0509]; t = 0.486, p = 0.627; TOST p = 0.011), while the relationship between rare experience and replay was significantly stronger for other world versus current world paths (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 confirmed this experience-replay relationship in a model-free analysis, 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 understand this putative memory maintenance signal, we examined a link between the rare experience replay effect and planning forward replay. We found an inverse relationship between the strength of the modulation of backward replay by rarity and planning replay across participants (current world forward, world change trials r = -0.521, p = 0.009). This relationship indicates that in individuals with a stronger replay of more rare experiences, subsequent forward replay during planning is decreased, potentially due to better memory maintenance.

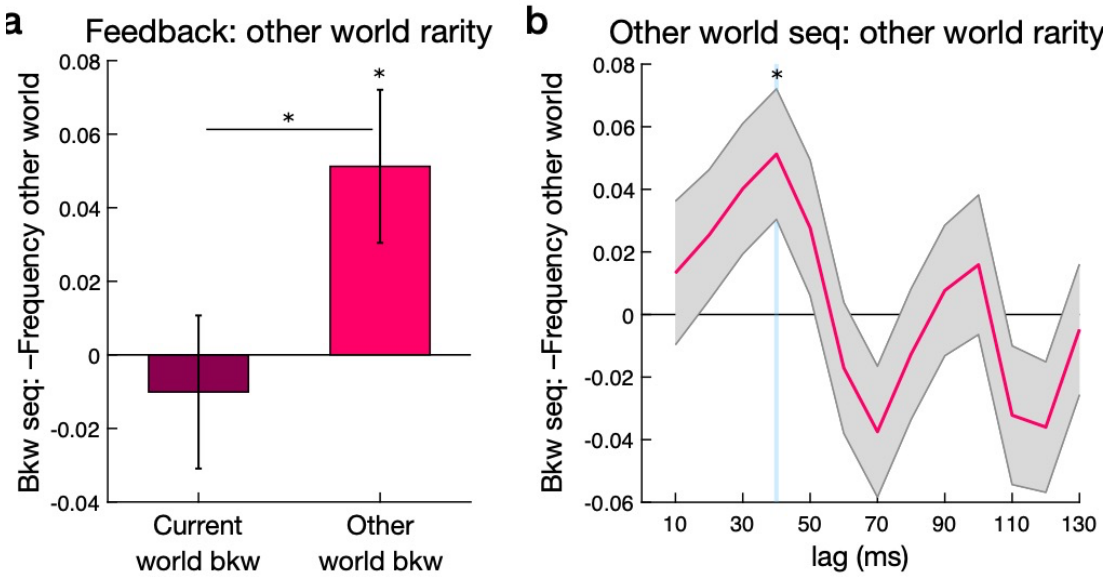

Figure 4. Feedback period backward replay relationship to the rarity of recent other world experience. (**a**) Lower recent experience, or rarity, of the other world correlated with greater backward replay of this other world. (**b**) Timecourse of regression coefficients for the rarity effect of interest, showing effects at state-to-state lags from 10 to 130 ms, highlighting the 40 time lag of interest (light blue line). Y-axes represent sequenceness regression coefficient for rarity of other world. Seq = sequenceness; * p < 0.05.

Next, we sought to link this rare experience replay signature to choice behavior. If feedback replay on a preceding trial supports memory for structure and value in the other world, we might expect the strength of this replay signal to positively influence choice in the other world. In our model, we tested this by allowing replay-related memory maintenance to decrease the uncertainty (or noisiness) of choices. Thus, we augmented the best fit model-based reinforcement learning model to allow variation in uncertainty, captured by two additional softmax inverse temperature parameters, by the presence of high versus low replay for the preceding other world (**Methods**). A higher inverse temperature parameter can reflect lower uncertainty such that choices are more strongly guided by estimated prospective values. We indeed found a significantly higher inverse temperature when choices were preceded by high replay versus low 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 was related to rarity of experience, we found no modulation of inverse temperature by rarity, consistent with backward replay underlying the observed effect. Together, these results indicate that backward replay is positively modulated by the rarity of experience and that the strength of this effect influences uncertainty in subsequent choices.

We then compared the links we identified with backward replay at feedback and forward replay at planning. We found no correlation between planning period forward replay and rarity of recent experience (current world 70 ms lag, $p = 0.815$), while the feedback period effect was significantly stronger than the planning period (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 selectively relating to the expected signature of memory maintenance.

At feedback, we also identified significant forward replay with a 70 ms time lag (**Fig. 2c**). We found no correlations of 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 also found no relationship with variables of interest at feedback or during planning (**Supp. Results**).

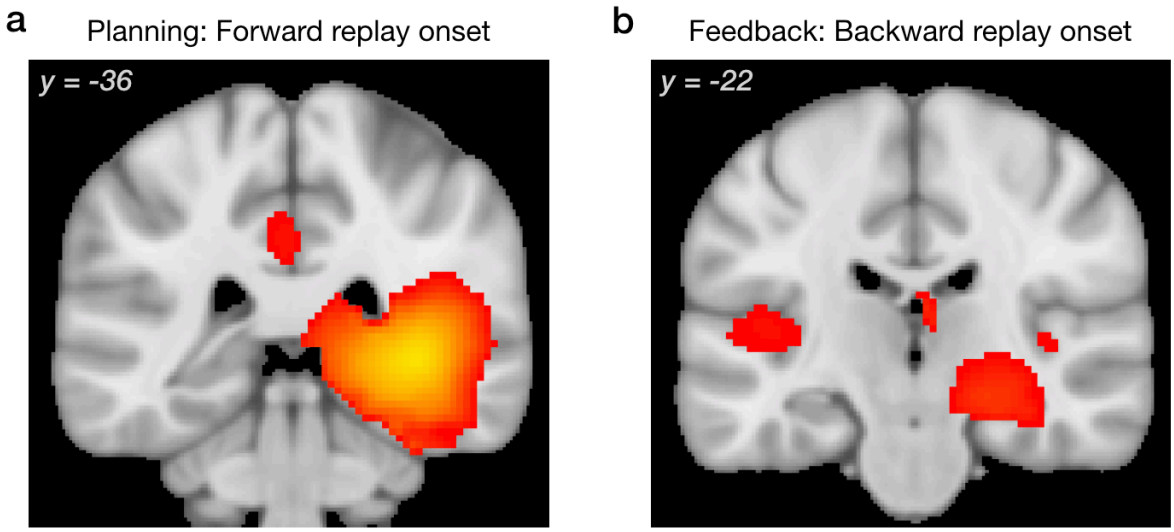

Figure 5. Replay onset beamforming analyses. (**a**) In the planning period, beamforming analyses revealed power increases associated with replay onset in the right MTL, including the hippocampus. The y coordinate refers to the Montreal Neurological Institute (MNI) atlas. (**b**) After reward feedback, power increases associated with replay onset were also found in the right MTL, including the hippocampus. See also **Fig. S6**. For display, statistical maps were thresholded at $P < 0.01$ uncorrected; clusters significant at $P < 0.05$, whole-brain corrected using non-parametric permutation test. For unthresholded statistical maps, see https://neurovault.org/collections/11163/.

Replay onset beamforming analyses

To explore the spatial source of sequenceness events we conducted beamforming source localization analyses to test whether replay onset is associated with increased power increases in the hippocampus (Liu et al., 2019; Wimmer et al., 2020; Liu et al., 2021b; Nour et al., 2021). Candidate replay onsets were identified by locating sequential reactivation events at the time lag of interest, applying a stringent threshold

to these events, and conducting broad-band beamforming analyses (1–45-Hz power). In both the planning and feedback periods, we identified power increases associated with replay onset in a region that encompassed the hippocampus (**Fig. 5**), as part of clusters with peaks in the visual cortex (planning forward replay MNI coordinates [x, y, z] 32, -60, -16; $p < 0.05$ whole-brain permutation-based cluster correction; feedback backward replay, 18, -94, -14; **Fig. S6**; **Table S4**). Using an ROI analysis, during the planning period we found significant power increases at replay onset in the right hippocampus (30, -36, -4; $p < 0.05$ ROI permutation-based cluster correction). In the feedback period, we found bilateral power increases in the hippocampus (right 26, -22, -14 and 26, -36, 0; left -26, -36, -6; $p < 0.05$, corrected). These results support the interpretation that sequential reactivation is linked to enhanced hippocampal activity.

Discussion

Here we tested two prominent functions of neural replay, planning and memory maintenance. During planning, forward replay was linked to the relative benefit derived from model-based, goal-directed, behavior. By contrast, after outcome feedback, backward replay of alternative world paths positively related to the rarity of the other world in recent experience. In line with a role for this signal in memory maintenance, stronger replay at reward feedback related to a subsequent decrease in choice uncertainty in the alternative world. A direct comparison of the planning and feedback period effects revealed a double dissociation, with forward replay selectively related to model-based generalization during planning, and backward replay selectively related to memory maintenance after feedback.

By manipulating the benefit derived from model-based generalization of reward value across trials, we identified a relationship between planning replay and the use of model-based inference. Building on previous work which identified a link between future state reactivation and individual variability in model-based behavior (Doll et al., 2015; Wise et al., 2021) here we show, on a trial-by-trial basis, that replay was increased when making a choice in a different but functionally equivalent state as the preceding trial,

potentially supporting value inference. Source localization results associated planning replay with increased activity in the hippocampus. Planning replay is thus a potential neural mechanism underlying a necessary role for the hippocampus in model-based inference, as demonstrated in lesion studies (Miller et al., 2017; Vikbladh et al., 2019).

Model-based generalization can be accomplished by 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. Our design included two features aimed to enhance our ability to illuminate planning mechanisms. First, unlike the majority of experiments which repeat the same environment on every trial (e.g. Daw et al., 2011; Konovalov and Krajbich, 2016), we employed two separate environments, intermixed unpredictably across trials, thereby limiting utility in planning for the next choice immediately after feedback. Furthermore, in our design, the two alternative start states in a world converge upon shared paths at the first step (Doll et al., 2015; Kool et al., 2016), potentially increasing the degree of planning period inference required to clearly differentiate between trajectories. These features distinguish our current task from a recent report focusing on feedback replay signals (Liu et al., 2021b), which suggested that generalization was accomplished by updating of different non-local start states after feedback. Here, in 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) demonstrate a link between neural replay and model-based inference at different time periods, with our planning results indicating that inference can be accomplished on-demand when facing less certain situations.

To successfully generalize reward feedback and perform well in our task, participants needed a model-based strategy. However, similar to the 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 and identified new links between replay and

model-based planning. We speculate that in many environments outside the lab, replay-assisted planning of trajectories is advantageous, such that is may become a default strategy employed even where it is not strictly advantageous. It will be important for future experiments to study these neural mechanisms in environments where sequential step-by-step choices are required, although we note a previous study in this domain did not identify connections to participants' behavior (Kurth-Nelson et al., 2016).

In contrast to planning, the time following feedback entails minimal cognitive demands in the current environment, rendering it more likely that activity at this time-point might support maintenance of weaker memories (Olafsdottir et al., 2018). Our results, selective to the feedback period, are consistent with this prediction. 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 the reduction in experience (Carey et al., 2019). Our results provide a clear quantification of the link between replay and rare experience by parametrically varying experience in two distinct environments. Such a neural replay mechanism can act to reinforce memories that are becoming weaker, effectively serving as internally-generated interleaved training (Kumaran et al., 2016).

The two separate functions of planning and memory maintenance that we identify may trade-off. One proposal 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, one study has reported that replay content temporally proximal to active navigation was task-directed, while replay content during reward consumption was undirected, potentially related to maintaining memory of distant locations (Olafsdottir et al., 2017). We did not formally manipulate task demands such as cognitive load during planning, but our results provide potential pointers for future targeted studies that might do so (Otto et al., 2013). Computationally, our results suggest the expression of replay may reflect an arbitration of resources as a function of need for planning versus memory (Kali and Dayan, 2004). In relation to such a tradeoff,

we found that the strength of the maintenance-related replay of rare experiences at feedback was related across participants to a lower expression of planning period 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).

Our results provide evidence that prospective replay is enhanced when model-based behavior is needed, while replay consistent with memory maintenance is observed when demands are low. Both planning and memory maintenance functions are necessary for adaptive behavior, as a stable memory of the one's environment can aid successful decision making (Behrens et al., 2018; Joo and Frank, 2018; Olafsdottir et al., 2018). By identifying distinct replay signatures for planning and memory maintenance, our results have relevance for better targeting and understanding disruptions to these functions in psychiatric disorders (Brewin et al., 2010; Suh et al., 2013; Huys et al., 2021; Nour et al., 2021).

Acknowledgments: The authors thank Rani Moran for helpful discussions, 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 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).

Author contributions: G.E.W., Y.L., and D.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.M and R.D. contributed to data interpretation. G.E.W. wrote the paper with Y.L., D.M., and R.J.D.

Competing interests: The authors declare no competing interests.

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.

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.

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

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*.

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.

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.

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.

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.

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*.

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*.

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.

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

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.

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.

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.

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. The MEG sessions were 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), leaving 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). No participants were excluded based on poor MEG localizer decoding performance.

Experimental task

We designed our reward learning experiment to investigate the potential role of replay in prospective planning and memory maintenance (**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 common a ‘two-step’ task (Daw et al., 2011). This version of the task utilizes deterministic transitions between states, with the equivalence between top-level start states providing the ability to dissociate model-based and model-free behavior. 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). Conversely, probabilistic transitions would have decreased our ability to detect

clear evidence of sequential reactivation, given the increase in number of transitions to evaluate and the variability in estimating those probabilities across participants, among other factors. This reasoning also led to using a task with a non-branching path structure (i.e. where there were not choices at each state). We also constrained choice points and branching paths in order to maintain our ability to sufficiently decode visual states using current 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 could take advantage of a 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 maintenance (**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 preceding feedback (Collins and Frank, 2012; Wimmer et al., 2018).

We ensured that participants understood the structure of the task via three features. First, we provided extensive instructions to minimize misunderstanding, as poor understanding of a related paradigm has been reported to result in apparent model-free behavior (Feher da Silva and Hare, 2020). Second, we employed an initial non-reinforced structure exploration learning phase. Third, participants learned the maze structure with different stimuli in an initial session, followed by a break before the

MEG session, allowing for robust learning of the structure prior to scanning and potentially allowing for memory consolidation. Fourth, across all sessions, accurate knowledge of task structure itself was incentivized by periodic memory test trials that allowed participants to harvest additional payment. 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 the same 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 worlds 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 blocks, 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 the paths (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.

Trial lists for the reward learning phase were generated with counterbalanced reward point drifts, shape assignment to paths, and stimulus assignments to states. 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. World order was pseudorandomized with predominant alternation across trials and a maximum repetition of the same world of 3 trials. 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. Stimuli were assigned to states based on four counterbalance lists. 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.

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-participant 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 subject's 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 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:

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

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:

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

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]):

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

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.

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

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:

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

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):

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

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}:

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

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

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

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 maintenance

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 maintenance, we estimated an additional reinforcement learning model augmented by neural replay. Enhanced memory maintenance 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 maintenance-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 maintenance 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 the preceding trial (e.g. from world 1 to world 2, or vice versa). On world change trials, 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 maintenance 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):

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

Rarity of experience

To examine potential memory maintenance 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.

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 (CTF Omega, VSM MedTech) recorded data continuously at 1200 samples per second, utilizing 272 channels (3 channels of the original 275 channels are excluded because of excessive noise in routine testing). Three head position indicator coils were used to locate the position of 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 monitor participant's eye movements and blinks. 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.

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. 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 exhibited high kurtosis (>20), a repeated pattern in the time course and consistent spatial topographies. Mains interference had 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. The first 4 trials were excluded from sequenceness analyses to allow for familiarization with the reward learning task phase. 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, following a related previous report in rodents (Olafsdottir et al., 2017), we focused on the time period following initial reward processing, as we expected that the demands of feedback processing would preclude the engagement of any memory maintenance signal. Thus, the analyses included 3.5 s of data, excluding the first 1.5 s following reward. Note that results remain qualitatively the same without this exclusion.

MEG data decoding and cross-validation

Lasso-regularized logistic regression models were trained for each stimulus. 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.

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.

Sequenceness measure

The decoding model described above allowed us to measure spontaneous reactivation of task-related representations during memory retrieval. We next used Temporally Delayed Linear Modelling 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).

First, we applied each of the 12 stimulus decoding models to the planning period MEG data and, separately, the feedback period MEG data. This yielded twelve timeseries of reactivation probabilities for each period in each trial, each with length N, where N is the number of time samples included in the analysis window.

We then used a linear model to ask whether particular sequences of stimulus activations appeared above chance in these timeseries. For each stimulus *i*, at each possible time lag Δt, we estimated a separate linear model:

$$\mathrm{Y}_i = \mathrm{X}(\Delta t) * \beta_i(\Delta t)$$

The predictors $\mathrm{X}(\Delta t)$ were time-lagged copies of the reactivation timeseries. The model predicted Y_i, the reactivation of stimulus *i*. The linear model had N rows, with each row a time sample. We estimated $\beta_i(\Delta t)$, a vector of coefficients that described the degree to which stimulus *i*'s reactivation was predicted by activation of each other stimulus at time lag Δt. By repeating this procedure for each stimulus *i*, we obtained $\beta_i(\Delta t)$, a 12x12 matrix that can be viewed as an empirical transition matrix between the twelve stimuli at lag Δt.

Specifically:

$$\mathrm{Y}_i = \sum_{j=1}^{s} \mathrm{X}_j(\Delta t)\beta_{ij}(\Delta t)$$

Where $\mathrm{X}_j(\Delta t)$ are time-lagged copies of Y_j, *s* is the number of states, and therefore:

$$\mathrm{Y}_i(t) = \sum_{j=1}^{s} \mathrm{Y}_j(t - \Delta t)\beta_{ij}(\Delta t)$$

The matrix $\beta_i(\Delta t)$ is obtained by solving the following set of equations for each stimulus *i*, up to state *s*.

$$\mathrm{Y}_{i=1}(t) = \sum_{j=1}^{s} \mathrm{Y}_j(t - \Delta t)\beta_{ij}(\Delta t)$$

$$\mathrm{Y}_{i=2}(t) = \sum_{j=1}^{s} \mathrm{Y}_j(t - \Delta t)\beta_{ij}(\Delta t)$$

$$\mathrm{Y}_{i=s}(t) = \sum_{j=1}^{s} \mathrm{Y}_j(t - \Delta t)\beta_{ij}(\Delta t)$$

We next asked whether the $\beta_i(\Delta t)$ was consistent with a specified 12x12 transition matrix by taking the Frobenius inner product between these two matrices (the sum of element-wise products of the two matrices). Note that evidence of replay in all 4 paths were estimated at the same time, and thereby controlled for common variance. This resulted in a single number $Z_{\Delta t}$, which pertained to lag Δt. We calculated sequenceness from time lag 10 ms to 600 ms, but analyses focused only on lags up to 350 ms. Per-trial results were z-scored across lags. In our regression analyses testing for a relationship between sequenceness and trial-by-trial behavior and reward learning task variables, we entered either the forward ($Z_{f\Delta t}$) or backward ($Z_{b\Delta t}$) sequenceness measures. 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). 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 (l1 = 0.002); this parameter matches what we found previously with this approach (Wimmer et al., 2020).

The transition matrix was defined as the valid stimulus order in the four paths across the two worlds. To localize a time lag of interest for subsequent analyses, we utilized permutation tests using shuffled transition matrices. Shuffled transition matrices

were randomly generated with the constraint that they did not overlap with the true transition matrices. Further, each transition matrix was required to have eight entries (two per path), matching the structure of the true transition matrix. 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 (i.e. a mid-path stimulus was paired with both a preceding and a subsequent stimulus). These permutations were run 100 times. The significance threshold was determined by taking the maximum sequenceness value within each permutation. Then across the resulting values from all permutations, we applied at 95% threshold to derive the 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 analyses such as beamforming, 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 an episode (Liu et al., 2019; Wimmer et al., 2020). As described in the Results, a different direction and lag was used for the planning and reward feedback periods. In this analysis, we first found the stimulus-to-stimulus time lag Δt at which there was maximum evidence for sequenceness (as described above), time shifted the reactivation matrix *X* up to this time lag Δt, obtaining $\mathrm{X}(\Delta t)$. We then multiplied *X* by the transition matrix *P*, corresponding to the unscrambled sequences: $X \times P$. Next, we element-wise multiplied $\mathrm{X}(\Delta t)$ by $X \times P$. The resulting matrix had a column for each stimulus, and a row for each time point in the cue period for each trial. We then summed over columns to obtain a long vector *R*, with each element indicating the strength of replay at a given moment in time (across trials). Finally, within-participant, 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.

Specifically:

$$Proj = X(\Delta t)$$

Matrix *Proj* is obtained by time shifting the reactivation matrix *X* to time lag Δt.

$$Orig = X \times P$$

Matrix *Orig* is obtained by matrix multiplication between reactivation matrix *X* and transition matrix *P*.

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

Vector *R* is obtained by elementwise multiplication between matrix *Orig* and *Proj*, and then summing over columns.

MEG Source Reconstruction

Source reconstruction 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 data in broadband power across all frequencies.

For the replay onsets analysis, the baseline activity was 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.

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 $p < 0.05$ ($t = 2.81$); 5000 permutations).

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 (± 500 ms) identified using the preceding method. A cluster-based permutation analysis (OSL) then tested whether there was evidence across participants for (mean) power increase at replay onset compared to a pre-onset baseline, in the theta and high-frequency region of interest. The clustering algorithm used an initial threshold of $p < 0.001$ (a t-threshold of 3.75) 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 was input to 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).

In a more temporally fine-grained analysis, we 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.

Multilevel regression models

We conducted all pre-processing of the MEG data for multilevel modelling in Matlab. As further detailed above for the behavioral models, multilevel models were implemented in R. In MEG analyses, the independent variables included in models were the relevant sequenceness measures: 1) current and other world paths, 2) chosen path, non-chosen path, and other world paths, or 3) current world path 1 and path 2. For each analysis, all relevant sequenceness measures were included in the same model. When appropriate, variables in multilevel models were mean-centered and z-scored.

Statistical correction and null effects

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 regression coefficients were estimated using the glht function in the multcomp package; here, all variables entered into regressions were z-scored to allow for valid contrasts. For comparisons of softmax inverse temperature parameter, 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 can be found on GitHub (https://github.com/gewimmer-neuro/multistep_replay).