Experience replay is associated with efficient non-local learning

Yunzhe Liu,[1,2,3,4,9]* Marcelo G. Mattar,[5,9] Timothy E J Behrens,[4,6] Nathaniel D. Daw,[7,10] Raymond J Dolan [1,3,4,8,10]

1. State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China
2. Chinese Institute for Brain Research, Beijing, China
3. Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK
4. Wellcome Centre for Human Neuroimaging, University College London, London, UK
5. Department of Cognitive Science, University of California, San Diego, CA, USA
6. Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK
7. Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA
8. Department of Psychiatry, Universitätsmedizin Berlin (Campus Charité Mitte), Berlin, Germany
9. Equal contribution
10. Senior author

* Corresponding author. E-mail: yunzhe.liu@bnu.edu.cn

Abstract

To make effective decisions we need to consider the relationship between actions and outcomes. These are often separated by time and space. The mechanisms spanning these gaps remain unknown. One promising hypothesis involves neural replay of non-local experience. Using a task segregating direct from indirect value learning, combined with magnetoencephalography, we examined the role of neural replay in human non-local learning. Following reward receipt, we found significant backward replay of non-local experience, with a 160 msec state-to-state time lag, this was linked to efficient learning of action values. Backward replay, and also behavioural evidence of non-local learning, was more pronounced for experiences of greater benefit for future behaviour. These findings support non-local replay as a neural mechanism for solving complex credit assignment problems during learning.

One Sentence Summary

Non-local reverse replay is associated with model-based reinforcement learning in humans and is rationally prioritised according to utility.

Main Text

Effective decision making incorporates new experience into our existing knowledge of the world. This allows us to infer the likely future consequences of different actions without having to experience them. When you encounter a traffic jam at crossroads, for example, you learn that the route just taken should be avoided, but you might also infer the value in avoiding the alternate paths leading to this same location. Learning from direct experience can be straightforwardly achieved by detecting co-occurrence between actions (like routes taken), and subsequent rewards (*1-3*). However, to propagate that experience to many other distal situations requires additional computation, as in the example of alternate converging roads. We understand little about how this type of indirect learning is achieved in the brain (*4-7*).

In reinforcement learning (RL) theory (*8*), non-local value propagation can be achieved by "model-based" methods. In essence, these leverage a learnt map or model of the environment to simulate, or simply retrieve, potential trajectories (*9, 10*). These covert trajectories can substitute for direct experience and thereby span the gaps between actions and outcomes (*11*), a process known as experience replay.

In neuroscience, a potential neural substrate for this process is the phenomenon of hippocampal "replay". Here, cells in the hippocampus that encode distinct locations in space fire sequentially during rest in a time-compressed manner, recapitulating past or potential future trajectories (*12-14*). In rodents, hippocampal replay has been linked to learning in a number of different types of task (*15-19*), potentially reflecting (but in most cases not specifically isolating) a common mechanism of nonlocal value propagation. Also, hippocampal replay events co-occur with the firing of reward responsive cells in the dopaminergic midbrain (*20*), again suggesting the possibility that sequences can propagate value. More recently, replay was shown to support nonlocal propagation of value in an inferential reasoning task (*21*).

Here we build on this line of work to investigate whether such a replay mechanism specifically supports trial-by-trial reinforcement learning and whether it is preserved in humans. Using methods developed to measure fast neural sequences noninvasively (*22*), replay has now been found in humans during rest (*23-25*), with strong parallels to observations in rodents (*23*). However, a direct connection between replay of this sort and non-local reinforcement learning has yet to be established.

If replay supports non-local value learning, then its statistics should also be relevant for a second unresolved question, namely, given limited available time and resources which of the myriad possible future actions should the brain prioritise during replay? A reward-maximising agent might prioritise replay of whichever past experiences are most likely to improve future choices and thereby earn more reward (*26*). Recent theoretical analysis (*27*) argues that such rational priority of replay can be decomposed into the product of two factors, *need* and *gain*. *Need* captures how frequent a given experience will be encountered again in the future, while *gain* quantifies an expected reward increase from better decisions if that experience is replayed. Consistent with this view, Igata, Ikegaya and Sasaki (*28*) reported that replay preferentially represents salient locations when rats update their behavioural strategies.

Accordingly, we designed a novel decision-making task to measure both the behavioural effect and neural signature of nonlocal learning in humans, while at the same time manipulating *need* and *gain* to test its rational prioritisation.

Task design

Our key hypothesis was that neural replay facilitates non-local learning, and that such replay is prioritised by its utility for future behaviour. To detect human replay, we measured whole-brain activity using magnetoencephalography (MEG) while subjects performed a novel decision-making task. The task explicitly separates learning from direct vs. non-local experience, permitting the measurement of unambiguous neural and behavioural signatures of the latter (**Fig. 1**).

To isolate local and non-local learning, the task comprised three starting states (henceforth called "arms"), each with two alternative choices (**Fig. 1A**). On each trial, subjects are presented with one of the three starting arms and asked to make a choice between two paths within the arm. A choice then leads to a sequence of three stimuli ("path") followed by an end state (**Fig. 1D**). Each end state carries a reward (£1 or 0) with a probability that changes slowly from trial to trial. Crucially, the two end states, reachable from each arm, are shared across all three starting arms. This task structure allows subjects to use reward feedback to inform their choices, in particular their future choices at the other two starting arms (non-local learning). Put more explicitly, local learning in this task is defined as updating action value in the current arm based on the received outcome (£1 or 0), while the non-local learning is defined as value updating in the other paths (from the other two starting arms) that lead to the same end state. This feature allows us to isolate learning about nonlocal options and to compare non-local learning within the same trial, but between paths with different properties (e.g., *gain* and *need*). Note this is rendered possible because there are always two non-local paths per trial that are matched to one another in all respects, including the actual outcome (**Fig. 2A)**. The use of three-stimulus sequences allows unambiguous measurement of extended replay sequences (vs. co-occurrence) as well as their directionality.

In addition to distinguishing learning from local experience (the path just chosen) vs. non-local experience, the task allowed us to test our hypotheses that replay, and learning, should favour the higher priority of the two non-local paths. Priority differed between paths as a function of both *need* and *gain*. Differences in *need* arise out of the fact that each starting arm was encountered with a different, but constant, probability: rare (17%), occasional (33%), and common (50%) respectively (**Fig. 1A)**. These probabilities were learnt prior to the main task. *Gain* is a function of subject's experience of rewards in the main RL task, which in turn depends on the subject's own choices (i.e., gain is not manipulated explicitly or directly, nor is it necessarily independent from need. However, empirically, no significant correlation was found between *need* and *gain*, r = -0.004, p = 0.61). Since rewards were stochastic with fluctuating probability (**Fig. 1B**), the *gain* of propagating information about outcomes to different paths also fluctuated from trial to trial according to their individual reward histories. For instance, a newly encountered reward is more informative if this information promotes the selection of actions that would otherwise not be favoured, whereas the absence of reward is more informative for avoiding actions that would otherwise have been chosen.

Notably, a drifting reward probability creates a continuous learning task. As a result, subjects never know for sure whether either end state (or both) will deliver reward on a particular trial, or which of the two has a higher rewarding probability. Consequently, there is no absolute "correct" or "wrong" choice, only an ongoing adjustment of choice preference in light of experienced rewards and non-rewards, for local, as well as non-local experiences (**Fig. 2A)**.

Thus, our main RL task allowed us to investigate how subjects learn efficiently by incorporating new experiences, particularly those derived from a different starting arm, into updated choices. Before the main RL task, subjects were first taught an overall task model

comprising knowledge of the relations among different elements in the task, as well as the different starting probabilities assigned to each arm. To avoid any biased learning of the model, we introduced each component of the task carefully at different times (**Fig. 1C**).

To index neural representations of states in the main RL task, we first showed subjects 18 visual stimuli in random order, a task phase called the *functional localiser*. These stimuli acted as the different states in the main RL task (e.g., *A1, A2, A3* in **Fig. 1A**). We constructed a probabilistic decoding model for each stimulus based on their evoked neural response in this *functional localiser* task. These decoding models are used later to search for sequential reactivation of states in the main RL task. Note that the classifiers are unbiased to task experience and structure, because at this phase of the experiment subjects have no knowledge of the relationship among those stimuli, nor their value.

The experiment proceeded across distinct phases to ensure good knowledge of the task model (i.e., *model construction,* **Fig. 1C**). Consequently, upon completion of the functional localiser phase, subjects learnt how the 18 stimuli formed 6 distinct sequences, i.e., the relationship among the 18 stimuli. We refer to this phase as *sequence learning.* Subjects next learnt a mapping between sequences and end states, i.e., *end state learning* and then learnt which sequence belongs to which starting arm, i.e., *arm learning*. Note that, up to this point, no rewards have been introduced yet; subjects have only learnt the relational structure among arms, end states, and sequences. Following the arm learning phase, subjects learnt the starting probability of each arm, including the fact that these probabilities remain constant throughout the experiment. Subjects also learnt the frequency of each starting arm by experience, i.e., *arm frequency learning*. To ensure subjects had acquired knowledge of the full task structure, we included a quiz after each learning phase. All subjects achieved performance greater than 85% (see Materials and Methods). Upon completion of the entire set of preparatory phases, subjects performed the main RL task.

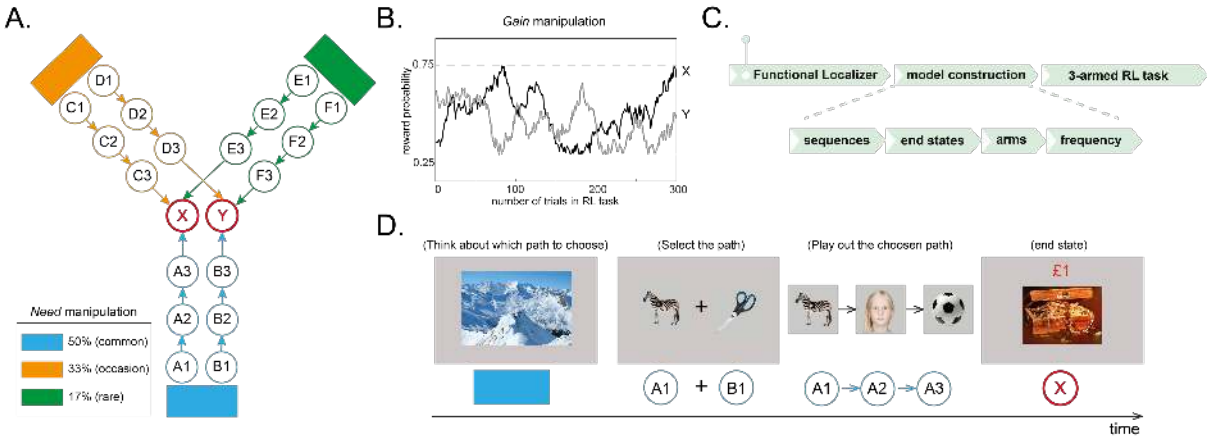

Fig. 1 Experimental design for model-based reinforcement learning task. **(A)** At each trial of the main RL task, subjects were presented with one of the three starting arms according to a fixed probability, and asked to select one from two alternative paths within this arm. This was followed by a transition through the associated path states and ended with an outcome (£1 or 0). The reward probability of the end states (i.e., X and Y) varied slowly and independently over time. A crucial feature of this task is that the end states are shared across all three arms, which enables non-local learning. *Need* is manipulated by the starting probability of each arm, shown as colour codes on the left. *Gain* is manipulated by the fluctuating reward probability of end states, X and Y, respectively. **(B)** An example of such drifting reward schedule. The reward probability of X and Y changes gradually and independently over trials, with gaussian random walk, bounded between 25% and 75%. **(C)** Each phase

of the experiment is shown in order. Subjects learnt the task model before commencing the main RL task. **(D)** An example of task trial in the main RL task. On the top, the text indicates what subjects need to do at a given time point in the trial. On the bottom, the corresponding symbols of task stimuli are shown. All photos shown are from pixabay.com and are in the public domain.

Behavioural evidence of non-local learning and prioritisation

The main RL task required subjects to learn the value of each action at each starting arm, with the aim of maximising reward. Direct, model-free learning allows subjects to favour a previously rewarded action when they encounter the same starting arm again. Consistent with this, when the starting arm is the same, subjects were more likely to repeat the same action if they had been rewarded compared to not rewarded on the last trial (Mixed effects logistic regression, $p = 7.5\times10^{-15}$). We then tested whether subjects transfer the value obtained in the chosen (i.e., local) path to the other non-local paths that lead to the same end state (**Fig. 2A**). Achieving effective non-local learning requires use of a model-based mechanism (such as replay), to propagate local rewards to non-local actions. A path leading to a previously rewarded end state was favoured even when the choice was presented at a different starting arm ($p = 9.5\times10^{-23}$). This effect did not differ significantly between trials whether the starting arm was repeated or not ($p = 0.90$ for the main effect of arm, $p = 0.46$ for the interaction effect between arm and reward, **Fig. 2B**). This is a hallmark of non-local, model-based learning (*4, 29*).

The previous analyses consider choices only as a function of events happening on the immediately preceding trial. To ask more detailed questions about learning, we built a computational model that incorporates longer-run effects of experience on multiple later choices. The model used – a modified Q-learning model – updates the value of each action based on experienced rewards and chooses action on the basis of these values (see Materials and Methods). However, we allow for the possibility that action values leading to the local path are learnt with a potentially different learning rate (α_d) than action values leading to the non-local path (α_n). Upon fitting this model to subject's trial-by-trial choices (*30*), we found that non-local action values were updated to a similar extent as local action values ($\alpha_d = 0.64$, $\alpha_n = 0.60$, diff in learning rate = 0.04, $p = 0.61$). These results confirm that subjects incorporate reward information into non-local actions, again, a hallmark of model-based learning.

We then asked whether the behavioural signature of learning from non-local outcomes was greater for paths with higher priority. We augmented the baseline model with additional free parameters measuring the strength of non-local learning as a function of the two task features that determines priority: *gain* (the informativeness of the current reward for improving choice at a given arm) and *need* (the likelihood that arm will be visited in the future, given by its frequency). This was possible because, in the task, there are always two non-local paths sharing the same end state with the current chosen one, allowing us to compare learning directly across them. We calculated the strength of learning by estimating separate learning rates for the higher and lower priority paths on each trial, in addition to a third learning rate for updating the local (chosen) path ($\alpha_d = 0.63$). Numerically, a higher learning rate was estimated for both higher-gain ($\alpha_h = 0.79$ vs $\alpha_l = 0.37$, **Table. S1**) and higher-need paths ($\alpha_h = 0.61$ vs $\alpha_l = 0.54$, **Table. S1**), a difference significant for gain (credible interval based statistical test, $p = 0.020$, see Materials and Methods for details), but not need ($p = 0.16$, **Table. S1**), indicating a divergence in gain.

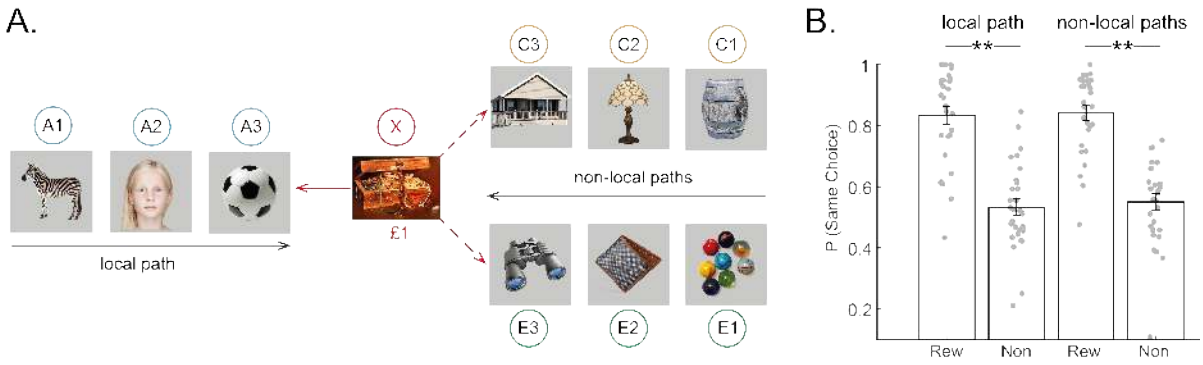

Fig. 2 Behavioural evidence of non-local learning. (A) An illustration of sequences of states for local (left – single path) and non-local experience (right – two non-local paths). The black arrow indicates the direction of actual experience, the red arrow indicates the hypothesized direction of credit (i.e., outcome, £1 or 0) assignment after receiving reward, solid red for the local experience, and dotted red for the two non-local experiences. **(B)** Behavioural results. The difference in performance between reward and no-reward in non-local paths is a defining feature of non-local learning. *Rew/Non* indicates whether subjects were rewarded or not rewarded on the last trial. P (same choice) is the probability that subjects, in the current trial, select the path leading to the same end state as that on the last trial. Error bars show the 95% standard error of the mean, each dot indicating results from each subject. * indicates $p < 0.05$, ** indicates $p < 0.01$.

Neural decoding of the task states

We next asked how the observed non-local learning is achieved in the brain. First, we verified that we could decode all 18 visual stimuli (corresponding to the 18 states, comprised of 6 distinct paths in the main RL task), well above chance. Classifiers were trained based on the evoked neural response of visual stimuli in the *functional localiser* task. In a leave-one-trial-out cross-validation scheme, one trial from each stimulus was omitted to form the testing set, and the remaining trials comprised the training set. We trained a binary classifier for each stimulus, based on their whole-brain neural response at a single time bin from post stimuli onset. This avoids potential timing confound for later sequence detection (*22, 31*). We obtained a peak cross-validation decoding accuracy of 47 ± 3 % (vs. chance level, 1/18 ≈ 6%), around 200 ms post stimulus onset (**Fig. 3,** see also **Fig. S1**, and Materials and Methods), consistent with previous findings (*23, 24*). Note that the mapping between the 18 visual stimuli and the corresponding state index was fixed within subject but was randomised across subjects. This randomisation ensures that any systematic difference among stimuli (e.g., stimulus preference or stimulus decodability), even if consistent across subjects, could not contribute to a difference in state decoding at the group level. We also verified, in simulation, that a decoding accuracy of 47% is sufficient to allow reliable detection of sequences (**Fig. S2,** see Materials and Methods for details). This showed that the sensitivity in detecting a ground-truth sequence strength was about 80% of that possible with perfect decoding accuracy, providing evidence of our ability to detect reliable sequences with a similar level of decoding accuracy in the real data.

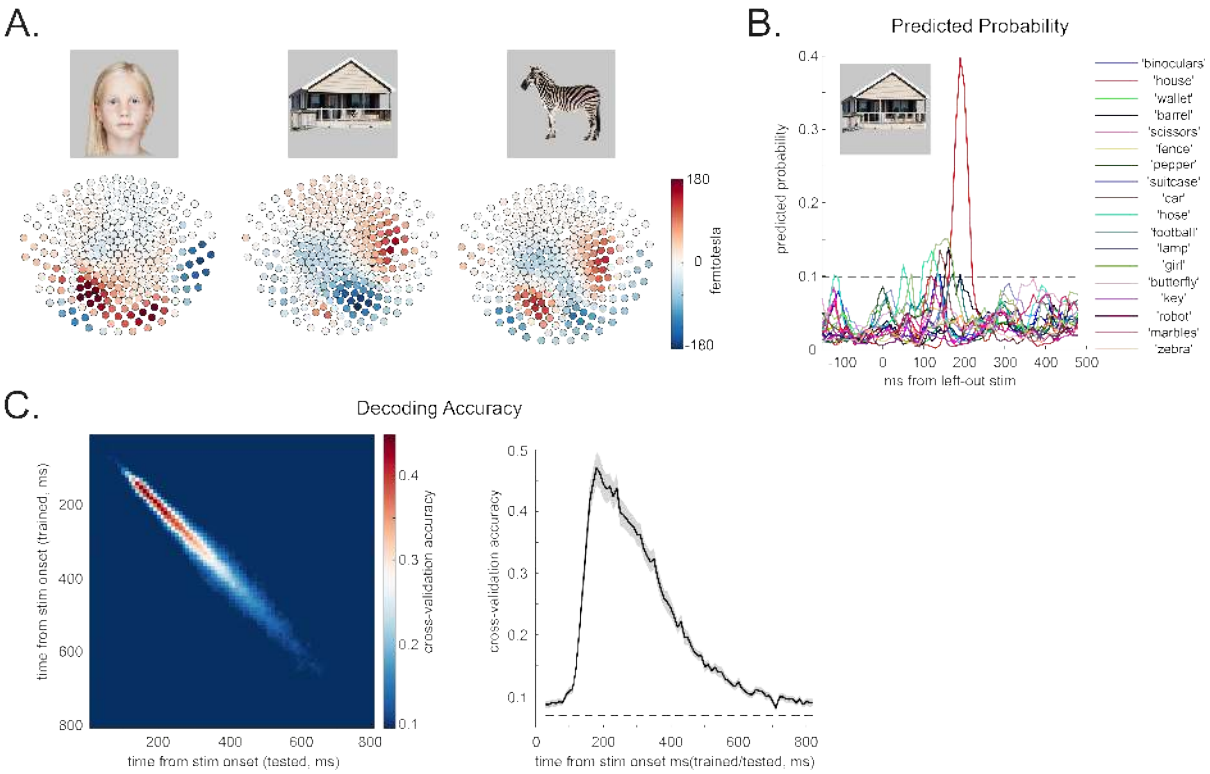

Fig. 3 Multivariate stimuli decoding. (A) Examples of multivariate whole-brain neural activity for classifier training, e.g., girl, house, and zebra. **(B)** Example of "house" classifier performance (red) plotted against all other 17 stimuli classifiers, when "house" picture was presented. The mapping between visual stimuli and their index states was randomised across subjects. **(C)** Mean decoding result for all subjects. The temporal generalisation plot is on the left panel, with Y axis indicating the time bins (10 ms each), the classifiers were trained on, and the X axis indicating the test time of classifiers. On the right panel, we plot the diagonal pattern of the temporal generalisation, namely the decoding accuracy obtained at the same time we trained the classifiers on. The dotted line is the permutation threshold. The mean performance for each individual state is shown in Supplementary Figure 1, data for each subject is shown in Supplementary Figure 3A.

Overall sequential reactivations of experiences during reward receipt

Having developed a set of stimulus classifiers, we next searched for their sequential reactivation in the main RL task. We applied the decoding models of the 18 stimuli (consisting of 6 paths) to the time of reward receipt, the period when new reward information is received, and learning occurs (**Fig. 4A,** see also **Fig. S3B** for representative MEG traces). Note that this period is analogous to the time when rodents consume a reward and backward replay sequences are observed (*32*) (but also see Discussion for connections to rodent sequences). We operationally refer to any reactivation of sequences here as *replay*.

We first look for spontaneous sequential replay of all stimulus reactivations whose orderings are consistent with the task. We refer to sequences that express the same direction as experience (e.g., $A1 \rightarrow A2 \rightarrow A3$) as forward replay, and sequences in the opposite direction (e.g., $A3 \rightarrow A2 \rightarrow A1$) as backward replay. Utilizing a recent methodological advance in MEG decoding of replay (*22*), we first assessed replay strength of all possible pairwise transitions at different speeds (i.e., state-to-state time lags), in both forward and backward directions (see Materials and Methods). Then, we obtained the sequence strength for each path by averaging their

corresponding pairwise transitions (e.g., $A1 \rightarrow A2$ and $A2 \rightarrow A3$, for the $A1 \rightarrow A2 \rightarrow A3$ path). We used a conservative non-parametric permutation test to determine the significant time lags, while controlling for multiple comparisons for all computed time lags. The same sequence analysis procedure has been validated in our previous work (*23, 24*).

Overall, we found evidence for two types of replay after reward receipt. First, we found significant forward replay encompassing a 20-30 ms state-to-state time lag. Second, we found backward replay encompassing a 130-170 ms state-to-state lag (**Fig. 4B,** see also **Fig. S3C** for individual sequence plots; **Fig. S4** for group level effects in linear mixed models). As in our previous work (*23, 24*), we then identified the time lags of interest based on a contrast between forward and backward sequences involving the same states (e.g., $A1 \rightarrow A2$ vs. $A2 \rightarrow A1$). These reflect the time lags at which forward replay is significantly stronger than backward replay, and vice-versa. We found that the forward sequence peaked at 30 ms lag, while the backward sequence peaked at 160 ms lag (**Fig. 4C**). Consequently, for all subsequent analyses, we focus exclusively on forward replay with 30 ms lag and backward replay with 160 ms lag. This focus allows us to investigate the finer-grained properties of replay at lags where it is known to be present, while avoiding further multiple comparisons over lags.

Recall that we tested subjects' knowledge of all 6 paths both prior to and during the main RL task. At either time, subjects' knowledge was not different across the 6 paths (before the main RL task, F (5,168) = 1.49, p = 0.20; during the main RL task, F (5,168) = 1.39, p = 0.23). Within subject, replay strength of the 6 paths also cannot be predicted by their corresponding structural knowledge during the RL task (96.2 ± 0.5% correct on average) for either 30 ms replay (p = 0.67), or 160 ms lag replay (p = 0.82). We also verified that differences in decoding accuracy across states did not predict sequence strength for either 30 ms lag forward replay (p = 0.30), or 160 ms lag backward replay (p = 0.56). These findings suggest that the (small) differences in structural knowledge, or state decoding abilities, do not contribute to the measured sequence strength.

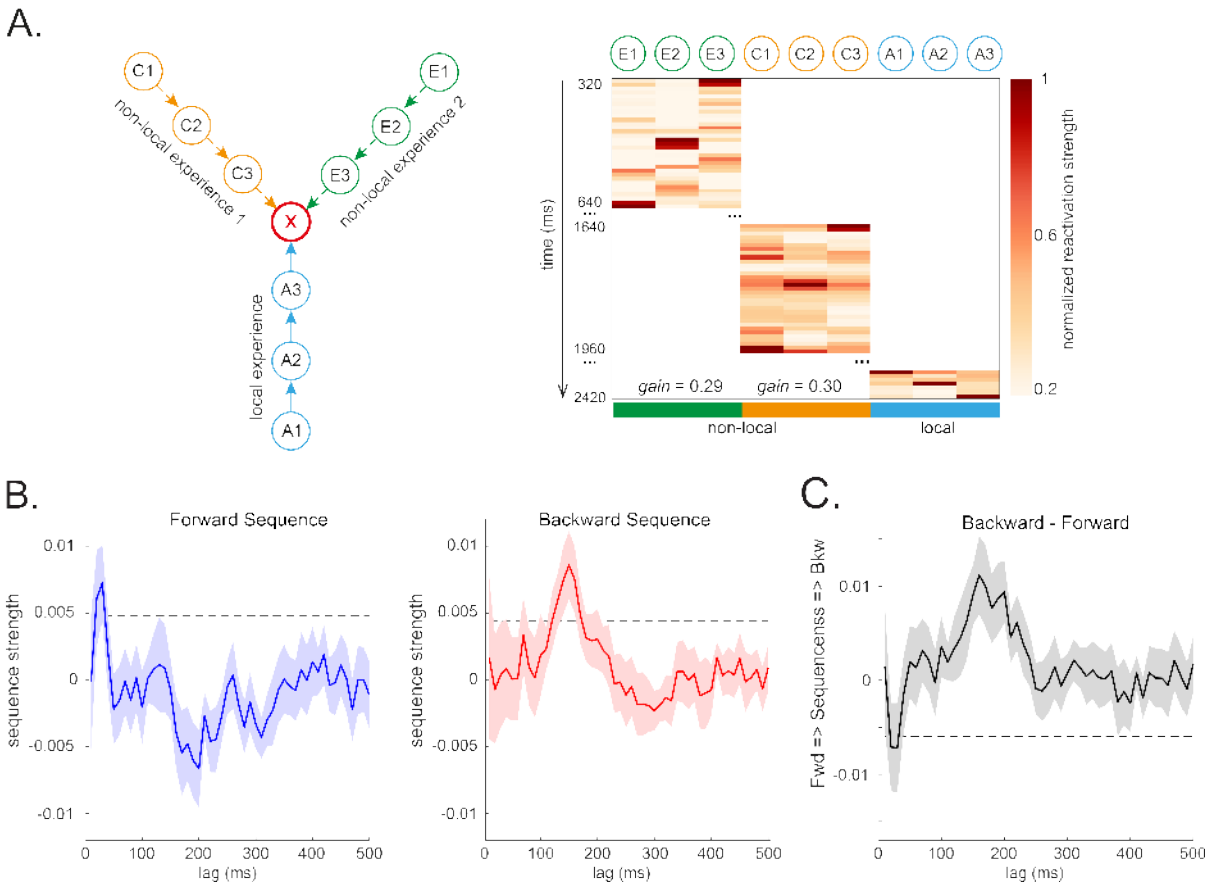

Fig. 4 Sequential replay of experiences during reward receipt. (A) An illustrative exemplar trial in the main RL task is shown (subject 14, trial 107). On the left panel, subject selected an A1->A2->A3 path, which renders A1->A2->A3 as the local experience, and C1->C2->C3 and E1->E2->E3 as two non-local experiences on this trial. On the right panel, the state decoding matrix during outcome receipt time (e.g., getting £1 in X) is shown, along with the *gain* estimate for the two non-local paths. A backward 160 ms lag sequences for both C1->C2->C3 and E1->E2->E3 path, and a forward 30 ms lag sequence for A1->A2->A3, are depicted. For visualization purpose, the reactivation strength of each state is max-normalised. Each time bin is 10 ms. **(B)** Sequence analysis at outcome receipt time shows two distinct signatures, one forward sequence (blue) with a 20-30 ms state-to-state time lag (left panel), and a backward sequence (red), with a 130-170 ms time lag (right panel). The X axis is the time lags. The Y axis is the evidence of sequence strength. **(C)** Contrast between backward and forward sequences in the computed time lags (i.e., speed). In this contrast, a forward sequence peaked at 30 ms time lag, and a backward sequence peaked at 160 ms time lag. Consequently, these time-points were selected for all later analyses. The dotted line is the permutation threshold after controlled for multiple comparisons.

Two types of replay: functional and physiological differences

The forward replay with 30 ms state-to-state time lag accords with previous work measuring replay in humans during post-task rest (*23*), though our results now extend those findings to a context that includes learning. The 160 ms backward replay has not been reported previously (although see Wimmer, Liu, Vehar, Behrens and Dolan (*24*) for memory replay at a similar speed). This replay pattern is intriguing as its direction is consistent with theoretical proposals for solving credit assignment by backpropagating reward information (*27*), and is also consistent with empirical results (*12, 23, 32*).

If this 160 ms backward replay supports non-local updating, we would expect it to also represent the contents of non-local paths. In line with this prediction, the 160 ms backward replay significantly represented non-local paths (one sample t test, t (28) = 2.92, p = 0.007), and to a significantly greater degree than local ones (paired t test, t (28) = 2.21, p = 0.03, **Fig. 5B**). The 30 ms forward replay showed an opposite pattern (interaction between replay types and representational content, F (1,28) = 7.37, p = 0.01). It did not represent non-local paths (one sample t test, t (28) = -0.09, p = 0.93), but likely the local one, i.e., the path just taken (t (28) = 1.42, p = 0.08, **Fig. 5A**).

We also tested whether these distinct replay signatures differ in terms of their underlying physiological properties. Fast human replay (e.g., with 40 ms time lag) during rest is associated with an increased ripple frequency power (*23*), akin to sharp wave ripple replay in rodents (*33-35*). In line with these results, we found that the initiation of a 30 ms forward replay was associated with a ripple frequency power increase (one sample t test, t (28) = 3.98, $p = 4.3\times10^{-4}$), but this power increase was not seen for the 160 ms backward replay (t (28) = 0.64, p = 0.53). A significant difference was also evident in the ripple power between the two types of replay (paired t test, t (28) = 3.03, p = 0.0052, **Fig. 5C**, see also **Fig. S5**). Mindful of well-known caveats regarding source localising MEG signals (*36*), whole-brain beamforming results indicated that while both replay types are associated with activation in visual cortex and medial temporal lobe, the 30 ms forward replay has higher hippocampal activation, and intriguingly, also higher activation in a region that encompassed the ventral tegmental area (VTA), compared to the 160 ms backward replay. Conversely, the 160 ms backward replay has greater cortical engagement (**Fig. S6**).

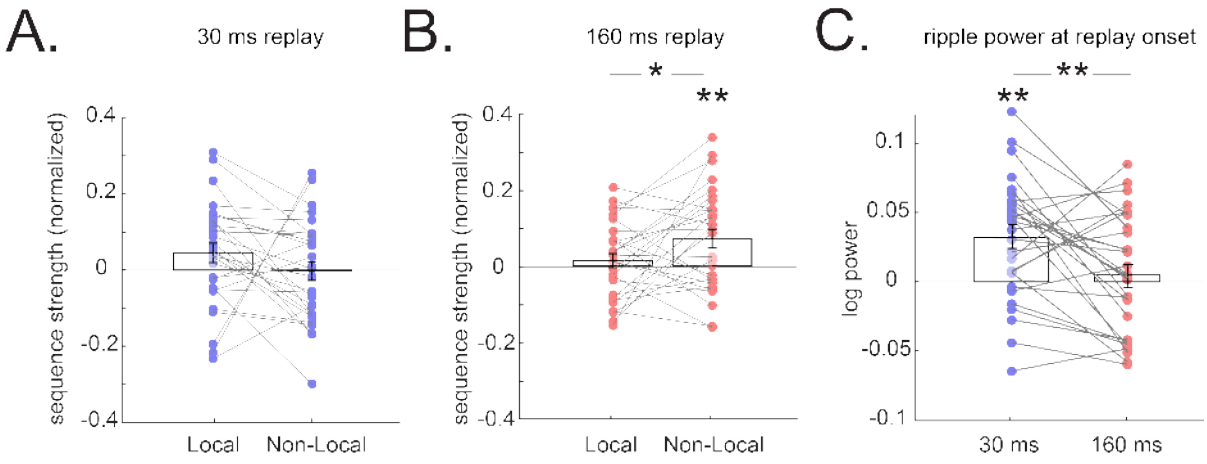

Fig. 5 Representational and physiological differences between the two types of replay. (A) A 30 ms forward sequence is likely to encode local experience, but not non-local. **(B)** A 160 ms backward replay encodes non-local as opposed to local experience. **(C)** The initialization of 30 ms forward sequence is associated with a power increase in a ripple frequency band (80-180 Hz), but this

is not the case for 160 ms backward sequence. These frequency power signatures are significantly different. The grey line connects results from the same subject. Error bars show the 95% standard error of the mean, each dot indicating results from each subject. * indicates $p < 0.05$, ** indicates $p < 0.01$.

Non-local replay accompanies efficient non-local learning

Having identified neural candidates for learning, we tested whether non-local replay (i.e., the 160 ms backward replay) is associated with non-local learning and, if so, whether such replay is competitively prioritised between the two non-local paths in accord with theoretical accounts (*27*). We again posed these questions in terms of RL-based computational models of trial-by-trial choice behaviour (see Materials and Methods).

First, in asking whether replay accompanies non-local learning, we augmented a baseline Q-learning model with a term measuring the effect of trial-by-trial neural replay on value learning. Having first separated learning rates for local and non-local paths (as before, these are paths leading to the same end state), we tested whether the baseline learning rate for each non-local path was significantly increased on trials when that path exhibited significant neural replay, vs. when it did not. We found higher nonlocal learning rate in the presence vs. absence of significant 160 ms backward replay (see supplementary material for detail, $\alpha_{replay} = 0.70$; $\alpha_{no-replay} = 0.61$; difference in learning rates = 0.09; $p = 0.023$, **Table. S2**). This was not the case when the same analysis was repeated for the 30 ms forward replay (difference in learning rates = 0.01, $p = 0.457$, **Table. S2**), and neither of the two replays was linked to local learning (with vs. without replay, $p = 0.60$ for 160 ms replay; $p = 0.88$ for 30 ms replay, **Table. S3**).

We next asked whether replay is prioritised to favour the more useful non-local experience. Recall that each trial has one local and two non-local paths. Thus, on each trial, we can classify the two non-local paths as high vs. low priority. This priority can be computed based on either the *need* (17%, 33%, 50%, for paths in rare, occasion and common arm, respectively), the *gain* (estimated per-arm, -trial, and -subject from behavioural model), or their product (*need* * *gain*) – i.e., utility (**Fig. S7**). According to RL theory (*27*), *need* should interact with *gain* (i.e., utility) in determining the actual priority for replay. Indeed, we found that the strength of the 160 ms backward replay was significantly stronger for a high vs. low utility (*need* * *gain*) path (**Fig. 6A,** paired *t* test**,** $t(28) = 3.30$, $p = 0.003$). Such prioritisation was absent in a high vs. low need or gain comparison (**Fig. S7**), nor did it exist for 30 ms replay (**Fig. 6A,** $t(28) = -0.34$, $p = 0.74$). These prioritisation results cannot be explained by differences in the actual frequency paths were encountered (which was determined by subjects' own choices): We found no link between replay strength of a specific path, and its frequency of occurrence in the RL task ($p = 0.59$ for 30 ms replay, $p = 0.54$ for 160 ms replay). Model-agnostic analyses (e.g., reward vs. no-reward) parallel these results (**Fig. S8**).

Finally, given that the 160 ms lag replay was associated with better non-local learning trial-by trial, *within-subject* (which is our main hypothesis), we conjectured that stronger 160 ms replay might also be positively associated with better task performance *across subjects*. This indeed was the case: a significant positive correlation, across subjects, was evident between average 160 ms lag replay strength and average reward earned per trial (robust correlation, $r = 0.41$, $p = 0.03$, **Fig. 6B**). This was not true for the 30 ms lag replay ($r = -0.29$, $p = 0.13$). We also tested whether a 40 ms backward replay (albeit non-significant on its own) may be related to value learning, given its reported involvement in a previous study (*23*). We found no evidence that a 40 ms lag replay was associated with learning for either local ($p = 0.28$) or non-local experience within-subject ($p = 0.32$), nor that it was linked to task performance across subjects ($p = 0.18$).

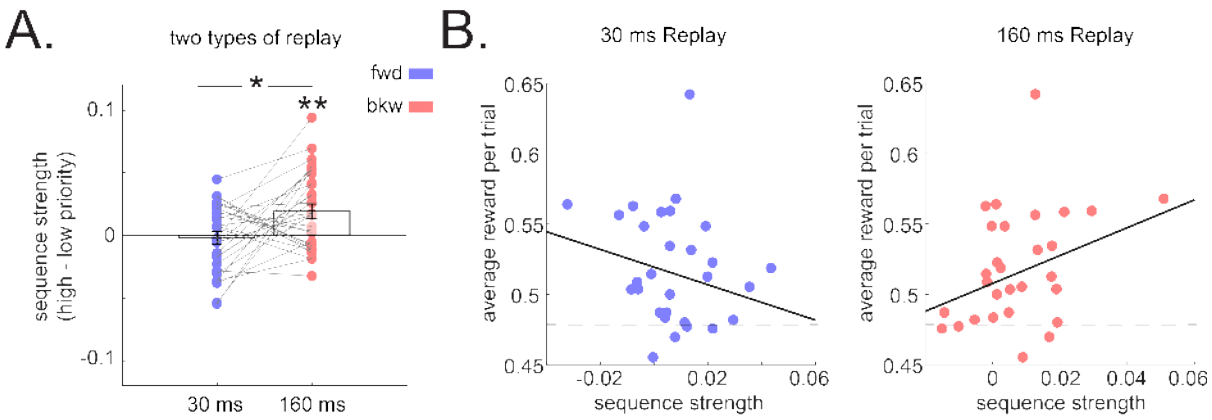

Fig. 6 Prioritisation of non-local replay. (A) 160ms backward sequence is replayed to a greater degree in the higher priority non-local path compared to lower priority one. The 30 ms forward replay does not differentiate between the two non-local paths. Error bars show the 95% standard error of the mean, with dots indicating results from each subject. * indicates $p < 0.05$, ** indicates $p < 0.01$. Grey line connects results from the same subject. **(B)** Sequence strength of 30 ms lag replay does not correlate with task performances (left panel). By contrast there is a significant positive correlation between the 160 ms lag replay and task performances across subjects (right panel). Each dot indicates result from one subject. The solid line reflects the best robust linear fit. The dotted line indicates the chance level of reward rate per trial with random choices.

Discussion

In the current study, we disassociated between two types of replay as a function of local vs. non-local learning. As a result, we established a connection between neural replay and learning via non-local credit assignment as expressed in behaviour.

Replay of non-local experiences was associated with more effective learning of action values, evidenced by enhanced assimilation of reward information on subsequent choices. In other words, replay connects actions and outcomes across intervening states and offers a neural mechanism for model-based reinforcement learning. Furthermore, the content of this replay, and separately the strength of updating as expressed behaviourally, were prioritised according to their utility for future behaviour (*27*).

These findings corroborate a long-standing hypothesis about the role of awake replay on model-based planning and credit assignment. This hypothesis was based primarily on rodent studies reporting replay patterns that would be appropriate for this function (*12, 32*). These results also extend on our previous fMRI results in humans linking non-local reactivation (without assessing sequences) to planning (*4, 5, 37*). In the current study, by exploiting the temporal resolution of MEG and the use of three-stimulus sequences, we could distinguish sequential replay from mere reactivation of isolated states. Notably, there were no significant effects related to reactivation of individual states alone (**Fig. S9**).

The 160 ms backward replay supporting non-local learning is distinct from the 40 ms replay reported in previous studies (*23, 38*). Unlike the latter, the 160 ms replay is not associated with a ripple frequency power increase (*23*). This raises an intriguing possibility that the 160 ms replay, which has a state-to-state transition frequency of around 6 Hz, might be processing states on consecutive theta cycles, which may have connections to rodent theta sequences (*39-*

42). However, theta sequences generally occur during ongoing behaviour in rodents and are in a forward direction, akin to a "look ahead" signal (but see (*43*) for backward theta sequence), while the 160 ms sequence we identify is backward in direction and occurs at the end of a trial.

It is interesting to note that the 160 ms backward replay alone is associated with value learning, while in Liu, Dolan, Kurth-Nelson and Behrens (*23*), we observed a faster replay (30-50 ms lag) shifting from forward to backward after a pairing with reward. This 160 ms lag replay might reflect a stronger task-engagement, or a more conscious computation compared to the 40 ms lag replay reported previously. This is plausible given there is no substantial *gain* (because reward contingency was fixed) in the previous study (*23*), and therefore replay was not required to promote learning. On the other hand, we can speculate that the faster 40ms lag replay previously observed may be similar to the 30ms lag replay observed here, which might reflect a stereotyped recapitulation of recent experience. This interpretation is consistent with previous findings (*38*) where the fast 40 ms lag sequences in a sequential planning task were shown to represent all possible transitions, instead of a specific planning trajectory.

The backward direction, representational contents and timing of this reverse replay are well suited to solve the non-local credit assignment problem, where an outcome at the end of a path impacts on decisions made at the (alternative) beginning. Theoretical work has focused more often on forward replay (or mental stimulation) of potential trajectories assumed to occur at choice time. Such patterns – more reminiscent of "planning" in the colloquial sense – also occur in rodents and could also, in principle, solve the current task. More generally, in the same framework they can be viewed as another means by which replay serves to connect actions and outcomes (*27*). We found no evidence that forward replay at choice time related to credit assignment (**Fig. S10**, also see **Table. S4** – for related modelling results). Such a process may play a role in other circumstances or in other task implementations, for instance in games like chess where particular choice situations are unlikely to have been anticipated ahead of time.

Together, our results connect several findings in human and rodent neuroscience, reveal that non-local backward replay serves as a neural mechanism for model-based reinforcement learning.

Materials and methods summary

Full materials and methods information are in the supplementary materials.

Participants

29 subjects (mean 23 ± 0.41 years, 17 females) were included for all analyses. All participants provided informed consent. They were all healthy with no history of psychiatric or neurological disorders. The number of subjects collected (30 + 1 pilot) were determined based on a prior power analysis where a one-sample *t* test requires approximately 27 people to find an effect different from 0 of size $d = 0.5$ (with alpha = 0.05, power = 0.80). Data from one subject were excluded due to contamination of metal on the MEG signal; pilot data was also excluded from formal analysis, leaving 29 subjects in total.

Stimuli and task design

In the current task, there were 3 starting arms, 2 end states, and 18 intermediate states (consisting of 6 paths). All of them are indexed by distinct pictures. The mapping between stimuli and states was fixed within subject but randomised across subjects. The task was run in following order: A) *functional localiser* – to obtain neural representations of the 18 stimuli (i.e., 6 paths); B) Model construction I: *sequence learning* – of the transition structures among 6 paths; C) Model construction II: *end state learning* – connections between 6 paths and 2 end states; D) Model construction III: *arm learning* –connections between 3 starting arms and 6 paths; E) Model construction IV: *arm frequency learning* – occurrence probability (i.e., *need*) of each starting arm in the main RL task (*44*). F) Main RL task – value learning, separating local and non-local experiences. In addition, *need* and *gain* were manipulated separately. *Need* was defined by the occurrence probability of the 3 starting arms (learnt in frequency learning, fixed across the experiment). *Gain* was manipulated by a drifting reward probability of each end state (with binary outcome, £1 or 0), they follow independent Gaussian random walk across trials, bounded between 25% and 75% (*27, 29, 45*).

MEG data acquisition and preprocessing

The MEG data was collected while subjects sat upright, performing the task (with exception to *frequency learning*). The data was recorded at 1200 samples/second using a whole-head 275-channel axial gradiometer system (CTF Omega, VSM MedTech). The task was divided into multiple scanning sessions, with each session less than 10 mins. Subjects were asked to remain still during the scanning session but were able to take a rest between sessions. At the start of each scanning session, participants were asked to move back to where they were, and their head positions were registered.

In preprocessing, the raw MEG data was first high passed at 0.5 Hz, and down sampled to 100 Hz for later analyses (with exception to temporal frequency analysis, for which the data were down sampled to 400 Hz, thereby preserve the ability to look for power change in high frequency, up to 200 Hz). Following that, excessively noisy segments and sensors were identified and removed, the resulting MEG data were then submitted to independent component analysis (ICA). The ICA was used for de-noising purposes alone. In each scanning session, up to 10 independent components (150 in total) can be excluded if they were clearly noise based on properties like spatial topography, time course, kurtosis of the time course and

frequency spectrum. At the end, all analyses were performed on the filtered, cleaned MEG signal at whole-brain sensor level (except for source localisation).

Behavioural analysis and modelling

Choice behaviour in the main RL task was analyzed as a function of reward (£1 or 0) at last trial and starting arm (same vs. different compared to last trial) at current trial. The choice at current trial was binarized based on whether it led to the same, or different, end state as that of the last trial. Linear mixed model was used to assess the group level effect while treating subjects as random effects, thereby accounting for trial-by-trial, subject-by- subject variations.

Modelling analyses were performed based on a modified Q learning algorithm (*4, 29*). In particular, learning rate were modelled separately for local and non-local experiences. Further extensions of the model separated learning for the two non-local experiences, based on priority (*need* or *gain*). The key comparison here was the learning rate difference between high vs. low priority paths.

Neural decoding analysis

Classifiers for the 18 intermediate states (i.e., 6 paths) in the main RL task were trained based on the evoked visual response at 200 ms post-stimulus onset (whole-brain sensor pattern) in the *functional localiser* task (*23, 24*). Importantly, during the functional localiser task, participants did not know either the mapping or occurrence probability of the stimuli and its corresponding states, and those stimuli were presented in a random order with equal occurrence. Thus, the classifiers were unbiased by the task structure.

Classifiers for the 2 end states were trained during the quiz question during *end state learning*; classifiers for the 3 starting arms were trained during the quiz question during the *arm learning*. In all those quiz questions, the picture for either end state or starting arm was presented in the centre of the screen, and subjects were asked to think about its associated paths. The training procedure and parameters were chosen to be identical as for the 18 intermediate states. Those classifiers were also unbiased by the occurrence probability (i.e., *need*) which was only learnt afterwards.

All classifiers were later used to examine for reactivation or sequences (i.e., sequential reactivation) in the main RL task. The decoding was performed both at the end (after reward receipt) and start (when the starting arm picture was presented) of a RL trial, to probe for credit assignment (value learning) and choice-related neural signatures respectively.

Neural sequence analysis

Sequence analysis was performed on the time series of decoded states at either the end or the start of a trial in the RL task. This analysis focused on the sequential reactivation of the 18 intermediate states, which consist of 6 distinct paths. The reactivation of starting arms states or end states were not considered in the sequence analysis to avoid potential visual confound. Sequence strength of a pair-wise state to state transition (e.g., state $i \rightarrow j$) measures the extent to which the representation of state i statistically predicts subsequent representation of some other state j at a particular time-lag (i.e., speed of replay), in a multiple regression model. This is an average measure of statistical predictiveness, where both the number and strength of replay events contribute to the current measure, which we call "sequence strength". This approach is motivated, in large part, by the fact that neural representations (of different states)

are only noisily and probabilistically decoded. The detailed approach, including related simulations, are described in Liu, Dolan, Penagos-Vargas, Kurth-Nelson and Behrens (*22*). The same human replay detection procedure has been applied successfully in previous empirical work (*23, 24*).

Sequence-behavioural modelling

We built a novel Q-learning model to formally test the effect of replay on learning. This separately models the learning rate for local vs. non-local experience. Crucially, replay of a specific path (local or non-local) is allowed to influence learning of the same path by having an additional free parameter associated with the existence of its replay (1 or 0, based on whether it is significant in permutation test). The key comparison is the learning rate difference between α_{replay} for paths with significant replay and $\alpha_{no-replay}$ for paths without significant replay.

ACKNOWLEDGMENTS

We thank Rani Moran and Evan Russek for discussions on study design and analysis. **Funding:** We acknowledge funding from the Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning to Y.L., a Wellcome Trust Investigator Award (098362/Z/12/Z) to R.J.D. This work was carried out whilst R.J.D. was in receipt of a Lundbeck Visiting Professorship (R290-2018-2804) to the Danish Research Centre for Magnetic Resonance. T.B is supported by a Wellcome Trust Senior Research Fellowship (104765/Z/14/Z), and a Principal Research Fellowship (219525/Z/19/Z), together with a James S. McDonnell Foundation Award (JSMF220020372). N.D and M.M are funded by the US National Science Foundation grant IIS-1822571, part of the CRCNS program. **Author Contributions:** Y.L, M.M, T.B, N.D & R.J.D contributed to conception and design of the study; Y.L contributed to data acquisition with help from M.M; Y.L, M.M & N.D contributed to data analysis. Y.L wrote the manuscript with assistance from all authors. **Competing interests**: None. **Data and materials availability**: Data and code are available at https://github.com/YunzheLiu/RLReplay

SUPPLEMENTARY MATERIALS

Materials and Methods

Supplementary text

Figs. S1 to S10

Table. S1 to S4

References (46–54)

REFERENCES AND NOTES

1. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience* **8**, 1704-1711 (2005).
2. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).
3. J. O'Doherty *et al.*, Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science* **304**, 452-454 (2004).
4. B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, N. D. Daw, Model-based choices involve prospective neural activity. *Nature neuroscience* **18**, 767 (2015).
5. G. E. Wimmer, D. Shohamy, Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270-273 (2012).
6. P. A. Lewis, S. J. Durrant, Overlapping memory replay during sleep builds cognitive schemata. *Trends in cognitive sciences* **15**, 343-351 (2011).
7. K. Doya, What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks* **12**, 961-974 (1999).
8. R. S. Sutton, A. G. Barto, *Reinforcement learning: An introduction*. (MIT press, 2018).
9. D. Silver *et al.*, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484 (2016).
10. K. Doya, Reinforcement learning in continuous time and space. *Neural computation* **12**, 219-245 (2000).
11. R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting. **2**, 160-163 (1991).
12. D. J. Foster, M. A. Wilson, Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680 (2006).
13. W. E. Skaggs, B. L. McNaughton, Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870-1873 (1996).
14. M. A. Wilson, B. L. McNaughton, Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676-679 (1994).
15. G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, M. B. Zugaro, Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience* **12**, 1222 (2009).
16. G. De Lavilléon, M. M. Lacroix, L. Rondi-Reig, K. Benchenane, Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature neuroscience* **18**, 493-495 (2015).
17. I. Gridchyn, P. Schoenenberger, J. O'Neill, J. Csicsvari, Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron*, (2020).
18. A. C. Singer, M. F. Carr, M. P. Karlsson, L. M. Frank, Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. *Neuron* **77**, 1163-1173 (2013).
19. H. F. Ólafsdóttir, C. Barry, A. B. Saleem, D. Hassabis, H. J. Spiers, Hippocampal place cells construct reward related sequences through unexplored space. *eLife* **4**, e06063 (2015).
20. S. N. Gomperts, F. Kloosterman, M. A. Wilson, VTA neurons coordinate with the hippocampal reactivation of spatial experience. *eLife* **4**, e05360 (2015).
21. H. C. Barron *et al.*, Neuronal computation underlying inferential reasoning in humans and mice. *Cell* **183**, 228-243. e221 (2020).

22. Y. Liu, R. Dolan, H. L. Penagos-Vargas, Z. Kurth-Nelson, T. E. Behrens, Measuring Sequences of Representations with Temporally Delayed Linear Modelling. *bioRxiv*, (2020).
23. Y. Liu, R. J. Dolan, Z. Kurth-Nelson, T. E. J. Behrens, Human replay spontaneously reorganizes experience. *Cell* **178**, 640-652 (2019).
24. G. E. Wimmer, Y. Liu, N. Vehar, T. E. J. Behrens, R. J. Dolan, Episodic memory retrieval success is associated with rapid replay of episode content. *Nature Neuroscience*, (2020).
25. N. W. Schuck, Y. Niv, Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, eaaw5181 (2019).
26. A. W. Moore, C. G. Atkeson, Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning* **13**, 103-130 (1993).
27. M. G. Mattar, N. D. Daw, Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience* **21**, 1609 (2018).
28. H. Igata, Y. Ikegaya, T. Sasaki, Prioritized experience replays on a hippocampal predictive map for learning. *Proceedings of the National Academy of Sciences* **118**, e2011266118 (2021).
29. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204-1215 (2011).
30. C. J. Watkins, P. Dayan, Q-learning. *Machine learning* **8**, 279-292 (1992).
31. D. Vidaurre, N. E. Myers, M. Stokes, A. C. Nobre, M. W. Woolrich, Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *Cerebral Cortex* **29**, 863-874 (2019).
32. R. E. Ambrose, B. E. Pfeiffer, D. J. Foster, Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron* **91**, 1124-1136 (2016).
33. M. F. Carr, S. P. Jadhav, L. M. Frank, Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience* **14**, 147 (2011).
34. K. Diba, G. Buzsáki, Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience* **10**, 1241 (2007).
35. S. P. Jadhav, C. Kemere, P. W. German, L. M. Frank, Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454-1458 (2012).
36. J. Mattout, C. Phillips, W. D. Penny, M. D. Rugg, K. J. Friston, MEG source localization under multiple constraints: an extended Bayesian framework. *NeuroImage* **30**, 753-767 (2006).
37. I. Momennejad, A. R. Otto, N. D. Daw, K. A. Norman, Offline replay supports planning in human reinforcement learning. *Elife* **7**, e32548 (2018).
38. Z. Kurth-Nelson, M. Economides, Raymond J. Dolan, P. Dayan, Fast Sequences of Non-spatial State Representations in Humans. *Neuron* **91**, 194-204 (2016).
39. G. Buzsáki, Theta oscillations in the hippocampus. *Neuron* **33**, 325-340 (2002).
40. M. Mehta, A. Lee, M. Wilson, Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417**, 741 (2002).
41. B. E. Pfeiffer, D. J. Foster, Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74 (2013).
42. K. Kay *et al.*, Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552-567. e525 (2020).
43. M. Wang, D. J. Foster, B. E. Pfeiffer, Alternating sequences of future and past behavior encoded within hippocampal theta oscillations. *Science* **370**, 247 (2020).

44. E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology* **13**, e1005768 (2017).
45. O. M. Vikbladh *et al.*, Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683-693. e684 (2019).
46. A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences. *Statistical science* **7**, 457-472 (1992).
47. A. Gelman *et al.*, *Bayesian data analysis*. (CRC press, 2013).
48. Z. Kurth-Nelson, G. Barnes, D. Sejdinovic, R. Dolan, P. Dayan, Temporal structure in associative retrieval. *eLife* **4**, e04919 (2015).
49. G. Agarwal *et al.*, Spatially distributed local fields in the hippocampus encode rat position. *Science* **344**, 626-630 (2014).
50. N. A. Herweg, E. A. Solomon, M. J. Kahana, Theta oscillations in human memory. *Trends in Cognitive Sciences* **24**, 208-227 (2020).
51. D. Bush, N. Burgess, Advantages and detection of phase coding in the absence of rhythmicity. *Hippocampus* **30**, 745-762 (2020).
52. S. E. Qasim, I. Fried, J. Jacobs, Phase precession in the human hippocampus and entorhinal cortex. *bioRxiv*, (2020).
53. T. Eliav *et al.*, Nonoscillatory phase coding and synchronization in the bat hippocampal formation. *Cell* **175**, 1119-1130. e1115 (2018).
54. U. Hauser, E. Eldar, R. J. Dolan, Separate mesocortical and mesolimbic pathways encode effort and reward learning signals. *Proceedings of the National Academy of Sciences* **114**, (2017).