RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Experience replay is associated with efficient nonlocal learning

Yunzhe Liu*†, Marcelo G. Mattar†, Timothy E. J. Behrens, Nathaniel D. Daw‡, Raymond J. Dolan‡

INTRODUCTION: Adaptive decision-making requires assimilation of reward information to guide subsequent choices. However, actions and outcomes are often separated by time and space, rendering this difficult. In reinforcement learning, this problem can be solved using "model-based" inference, in which an agent leverages a learned model of the environment to link local reward to nonlocal actions; this process is known as experience replay.

A potential neural substrate for this process is hippocampal "replay." In rodents, cells in the hippocampus fire in an organized manner that recapitulates past or potential future trajectories during rest. A similar phenomenon has also been observed in humans during a post-task rest period. However, a direct connection between replay and nonlocal (i.e., model-based) learning has yet to be established.

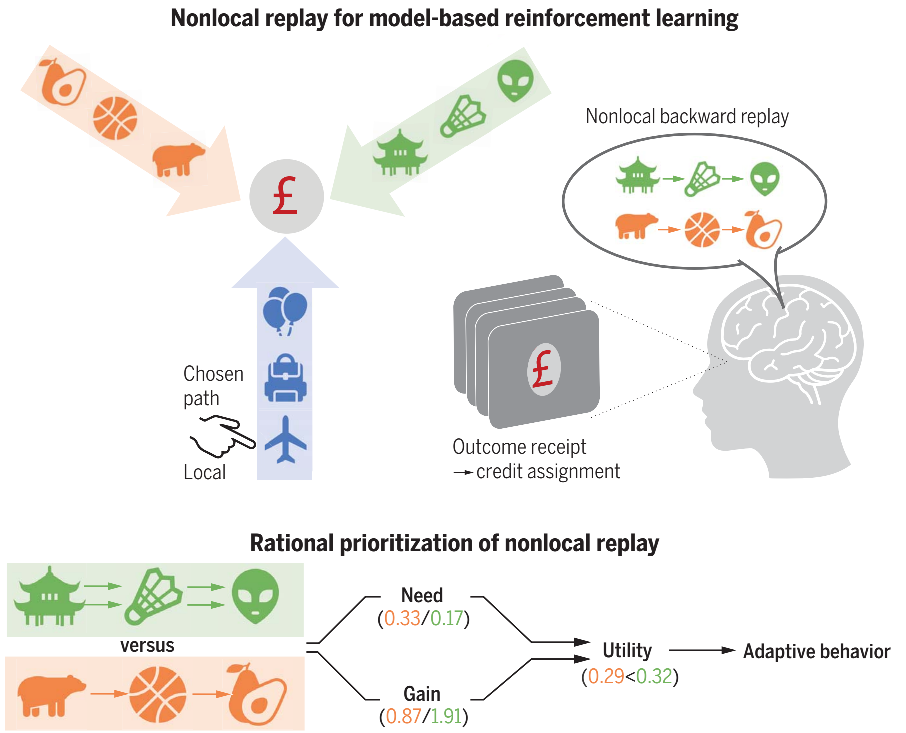
RATIONALE: The question of how to achieve model-based learning in the service of adaptive behavior is central to understanding intelligence in both biological and artificial agents. We addressed this question by exploiting a normative model of replay based on reinforcement learning theory. This model makes specific predictions regarding how replay relates to nonlocal learning and about which information is more or less useful if replayed. To measure replay in humans, we used machine learning techniques in conjunction with magnetoencephalography (MEG) recordings of whole-brain neural activity. These techniques enabled us to ask whether and how neural replay contributes to nondirect learning in humans. In so doing, we address an outstanding question in reinforcement learning theory: how the brain forms links between disjoint actions and goals and uses these to inform better decisions in the future.

RESULTS: We designed a novel decision-making task to separate local from nonlocal learning—that is, learning from direct experience as opposed to indirect learning based on inference. Specifically, we developed a three-arm task, wherein each arm comprised two paths leading to distinct end states. Crucially, the two end states, reachable from each arm, are shared across all three arms. This design feature allows post-choice reward feedback to inform future choices not only in the chosen arm (local learning) but also in either of the other two arms (nonlocal learning).

This work revealed the existence of backward neural replay of nonlocal experiences after reward receipt, with a 160-ms state-to-state time lag. In line with normative theory, such replay predominantly represents the path most useful for future behavior. This backward replay encoded nonlocal experience alone and was physiologically distinct from a faster forward replay (30-ms time lag), which was associated with power increase in a ripple band.

Using computational modeling, we showed that the strength of this backward replay relates to efficient trial-by-trial within-subject learning of the same nonlocal experience, as well as a better overall task performance across subjects. This is consistent with our theoretical predictions and provides strong support for a reinforcement learning–based account of neural replay in decision-making.

CONCLUSION: Backward replay accompanies efficient nonlocal learning in humans and is prioritized according to its utility. These results connect several findings in human and rodent neuroscience and implicate experience replay in model-based reinforcement learning. ■



Experience replay is associated with efficient nonlocal learning. Top left: The key element of the experimental design is a separation of local versus nonlocal learning. The chosen path (indicated by hand) reflects local experience; the other two paths leading to the same outcome state (the red £), but not directly experienced, are the nonlocal paths. Arrow direction indicates the order of actual experience; color indicates the arm identity. There are three arms, and outcomes are shared across the three arms. Top right: Nonlocal backward replay after reward receipt. There are two nonlocal experiences per trial. We found neural replay of these paths after reward receipt, consistent with a credit assignment account in reinforcement learning—an assignment of local reward information to nonlocal actions. Bottom: Consistent with reinforcement learning theory, replay was prioritized according to utility (need × gain) and was related to more efficient nonlocal learning. In the example, this is illustrated as stronger replay (double arrows) for the green path, because of its higher utility (0.32 versus 0.29) relative to the orange path.

The list of author affiliations is available in the full article online.
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
*Corresponding author. Email: yunzhe.liu@bnu.edu.cn

Cite this article as Y. Liu *et al.*, *Science* **372**, eabf1357 (2021). DOI: 10.1126/science.abf1357

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abf1357

RESEARCH ARTICLE

NEUROSCIENCE

Experience replay is associated with efficient nonlocal learning

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

To make effective decisions, people need to consider the relationship between actions and outcomes. These are often separated by time and space. The neural mechanisms by which disjoint actions and outcomes are linked remain unknown. One promising hypothesis involves neural replay of nonlocal experience. Using a task that segregates direct from indirect value learning, combined with magnetoencephalography, we examined the role of neural replay in human nonlocal learning. After receipt of a reward, we found significant backward replay of nonlocal experience, with a 160-millisecond state-to-state time lag, which was linked to efficient learning of action values. Backward replay and behavioral evidence of nonlocal learning were more pronounced for experiences of greater benefit for future behavior. These findings support nonlocal replay as a neural mechanism for solving complex credit assignment problems during learning.

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 a 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 (e.g., 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*), nonlocal value propagation can be achieved by "model-based" methods. In essence, these methods leverage a learned 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 nonlocal reinforcement learning has yet to be established.

If replay supports nonlocal value learning, then its statistics should also be relevant for a second unresolved question: Given limited available time and resources, which of the myriad possible future actions should the brain prioritize during replay? A reward-maximizing agent might prioritize 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 frequently a given experience will be encountered again in the future, whereas gain quantifies an expected reward increase from better decisions if that experience is replayed. Consistent with this view, Igata *et al*. (*28*) reported that replay preferentially represents salient locations when rats update their behavioral strategies.

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

Task design

Our key hypothesis was that neural replay facilitates nonlocal learning, and that such replay is prioritized by its utility for future behavior. To detect human replay, we measured whole-brain activity by magnetoencephalography (MEG) while subjects performed a novel decision-making task. Learning from direct experience and learning from nonlocal experience are explicitly separated in the task, permitting the measurement of unambiguous neural and behavioral signatures of the latter (Fig. 1).

To isolate local and nonlocal learning, we constructed the experiment as follows: The task consists of 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 are asked to make a choice between two paths within the arm. A choice then leads to a sequence of three stimuli ("paths") 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. 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 (nonlocal learning). Put more explicitly, local learning in this task is defined as updating action value in the current arm on the basis of received outcome (£1 or £0), whereas nonlocal 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 nonlocal learning between paths with different properties (e.g., gain and need) within the same trial. This is possible because there are always two nonlocal paths per trial that are matched to one another in all respects, including the actual outcome (Fig. 2A). The use of three-stimulus

[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.
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
***Corresponding author. E-mail: yunzhe.liu@bnu.edu.cn**

sequences allows unambiguous measurement of extended replay sequences (versus co-occurrence) as well as their directionality.

In addition to distinguishing learning from local experience (the path just chosen) versus nonlocal experience, the task allowed us to test our hypotheses that replay, and learning, should favor the higher priority of the two nonlocal paths. Priority differed between paths as a function of both need and gain. Differences in need were created because each starting arm was encountered with a different but constant probability: rare (17%), occasional (33%), and common (50%), respectively (Fig. 1A). These probabilities were learned 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). Because 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 favored, whereas the absence of reward is more informative for avoiding actions that would otherwise have been chosen.

A drifting reward probability creates a continuous learning task. As a result, subjects

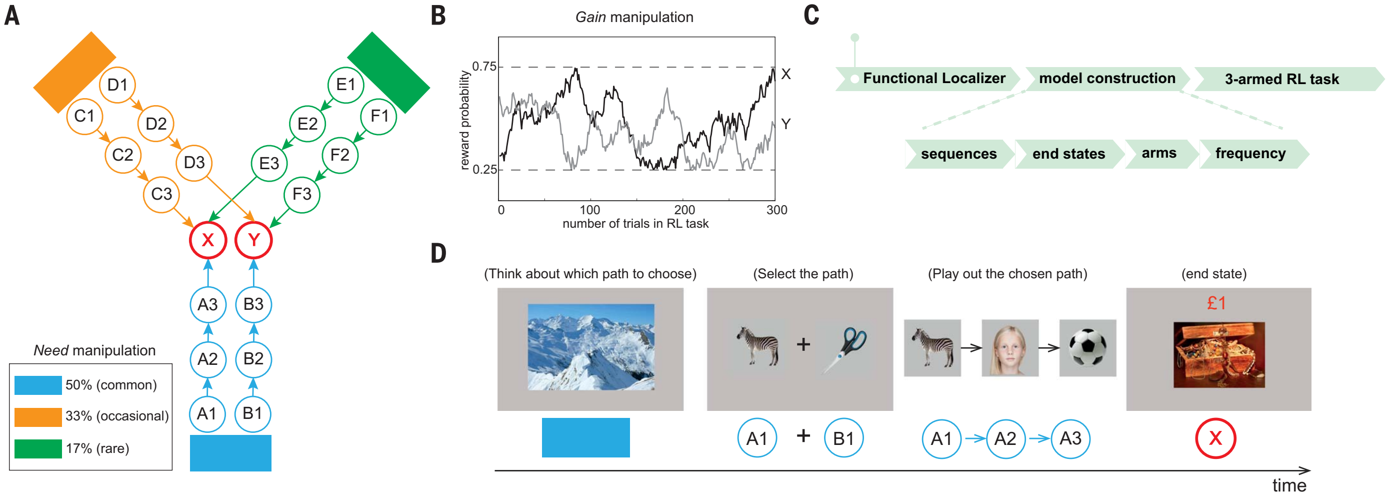


Fig. 1. Experimental design for model-based reinforcement learning task. (**A**) At each trial of the main RL task, subjects were presented with one of the three starting arms according to a fixed probability, and were asked to select one of two alternative paths within this arm. This was followed by 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 nonlocal learning. Need is manipulated by the starting probability of each arm, shown as color codes at lower left. Gain is manipulated by the fluctuating reward probability of the end states X and Y, respectively. (**B**) An example of such a 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 learned the task model before commencing the main RL task. (**D**) An example of a 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. (Photos shown in this and other figures are from pixabay.com and are in the public domain.)

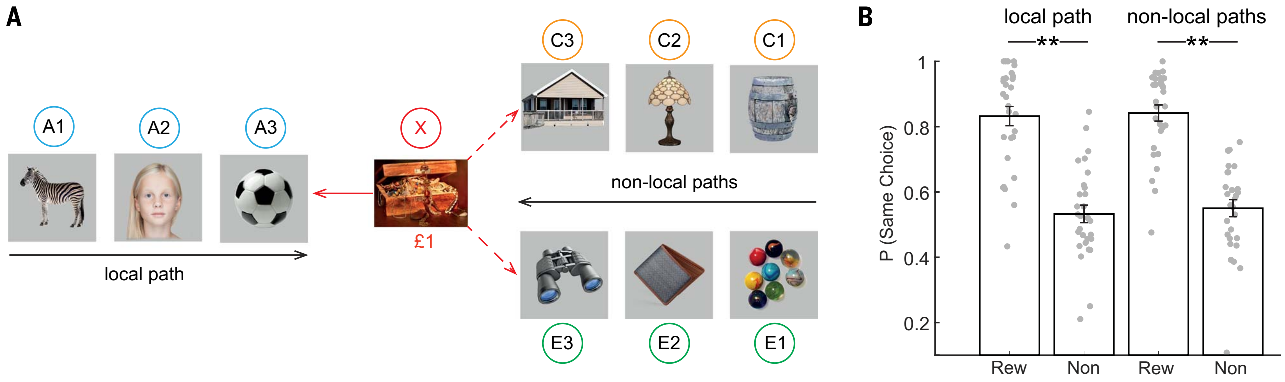


Fig. 2. Behavioral evidence of nonlocal learning. (**A**) An illustration of sequences of states for local experience (left; single path) and nonlocal experience (right; two nonlocal paths). Black arrows indicate the direction of actual experience; red arrows indicate the hypothesized direction of credit (i.e., outcome, £1 or £0) assignment after receiving reward (solid arrow, local experience; dashed arrows, the two nonlocal experiences). (**B**) Behavioral results. The difference in performance between reward and no-reward in nonlocal paths is a defining feature of nonlocal 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 95% SEM; each dot indicates results from each subject. *P < 0.05, **P < 0.01.

never know for sure whether either end state (or both) will deliver a 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 nonrewards, for local as well as nonlocal 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 localizer. These stimuli acted as the different states in the main RL task (e.g., A1, A2, and A3 in Fig. 1A). We constructed a probabilistic decoding model for each stimulus based on its evoked neural response in this functional localizer 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 with respect 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 localizer phase, subjects learned how the 18 stimuli formed six distinct sequences (i.e., the relationship among the 18 stimuli). We refer to this phase as sequence learning. Subjects next learned a mapping between sequences and end states (i.e., end-state learning) and then learned which sequence belongs to which starting arm (i.e., arm learning). Note that up to this point, no rewards have been introduced; subjects have only learned the relational structure among arms, end states, and sequences. After the arm-learning phase, subjects learned the starting probability of each arm, including the fact that these probabilities remain constant throughout the experiment. Subjects also learned the frequency of each starting arm by experience (i.e., arm frequency learning). To ensure that subjects had acquired knowledge of the full task structure, we included a quiz after each learning phase. All subjects achieved performance greater than 85% ([29]). Upon completion of the entire set of preparatory phases, subjects performed the main RL task.

Behavioral evidence of nonlocal learning and prioritization

The main RL task required subjects to learn the value of each action at each starting arm, with the aim of maximizing reward. Direct, model-free learning allows subjects to favor a previously rewarded action when they encounter the same starting arm again. Consistent with this, when the starting arm was the same, subjects were more likely to repeat the same action if they had been rewarded, versus not rewarded, on the last trial (mixed-effects logistic regression, $P = 7.5 \times 10^{-15}$). We then tested whether subjects transferred the value obtained in the chosen (i.e., local) path to the other nonlocal paths that led to the same end state (Fig. 2A). Achieving effective nonlocal learning requires the use of a model-based mechanism (such as replay) to propagate local rewards to nonlocal actions. A path leading to a previously rewarded end state was favored even when the choice was presented at a different starting arm ($P = 9.5 \times 10^{-23}$). 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 nonlocal, model-based learning ([4], [30]).

The previous analyses considered 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 we used, a modified Q-learning model, updates the value of each action on the basis of experienced rewards and chooses further action on the basis of these values ([29]). However, we allow for the possibility that action values leading to the local path are learned with a potentially different learning rate (α_d) relative to action values leading to the nonlocal path (α_n). Upon fitting this model to subjects' trial-by-trial choices ([31]), we found that nonlocal action values were updated to a similar extent as local action values ($\alpha_d = 0.64$, $\alpha_n = 0.60$, difference in learning rate = 0.04, $P = 0.61$). These results confirm that subjects incorporate reward information into nonlocal actions—again a hallmark of model-based learning.

We then asked whether the behavioral signature of learning from nonlocal outcomes was greater for paths with higher priority. We augmented the baseline model with additional free parameters measuring the strength of nonlocal learning as a function of the two task features that determine priority: gain (the informativeness of the current reward for improving choice at a given arm) and need (the likelihood that this arm will be visited in the future, given by its frequency). This was possible because in the task, there are always two nonlocal paths sharing the same end state with the current chosen one, allowing us to compare learning directly across the three paths. 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 paths ($\alpha_h = 0.79$ versus $\alpha_l = 0.37$, table S1) and higher-need paths ($\alpha_h = 0.61$ versus $\alpha_l = 0.54$, table S1). This difference was significant for gain (credible interval–based statistical test, $P = 0.020$) ([29]) but not need ($P = 0.16$; table S1), indicating a divergence in gain.

Neural decoding of the task states

We next asked how the observed nonlocal learning is achieved in the brain. First, we verified that we could decode all 18 visual stimuli (corresponding to the 18 states that constitute six distinct paths in the main RL task), well above chance. Classifiers were trained according to the evoked neural response of visual stimuli in the functional localizer 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 interval after stimulus onset. This avoids potential timing confound for later sequence detection ([22], [32]). We obtained a peak cross-validation decoding accuracy of 47 ± 3% (versus chance level, 1/18 ≈ 6%), ~200 ms after stimulus onset (Fig. 3 and fig. S1) ([29]), consistent with previous findings ([23], [24]). Note that the mapping between the 18 visual stimuli and the corresponding state index was fixed within each subject but was randomized across subjects. This randomization 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) ([29]). 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.

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

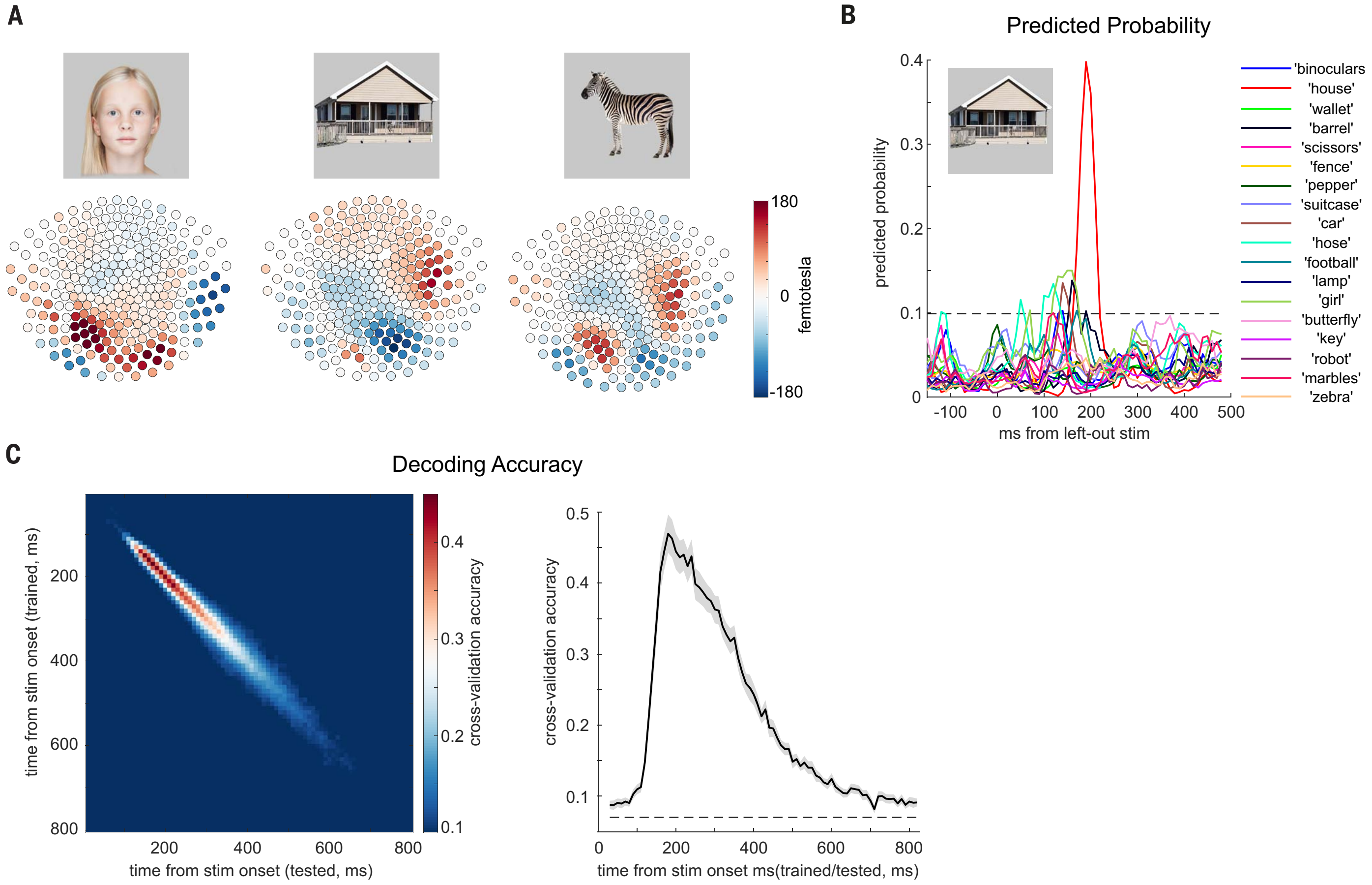


Fig. 3. Multivariate stimuli decoding. (**A**) Examples of multivariate whole-brain neural activity for classifier training (e.g., girl, house, and zebra). (**B**) Example of "house" classifier performance (red) when the "house" picture was presented, plotted against all 17 other stimuli classifiers. The mapping between visual stimuli and their index states was randomized across subjects. (**C**) Mean decoding result for all subjects. Left: Temporal generalization plot; time bins are 10 ms each. Right: Diagonal pattern of the temporal generalization; the dashed line is the permutation threshold. The gray shaded area represents the 95% confidence interval. The mean performance for each individual state is shown in fig. S1. Data for each subject are shown in fig. S3A.

six 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 (33) (see below for connections to rodent sequences). We operationally refer to any reactivation of sequences here as replay.

We first looked for spontaneous sequential replay of all stimulus reactivations whose orderings were consistent with the task. We refer to sequences that express the same direction as experience (e.g., A1 → A2 → A3) as forward replay, and sequences in the opposite direction (e.g., A3 → A2 → A1) as backward replay. Using a recent methodological advance in MEG decoding of replay (22), we first assessed the replay strength of all possible pairwise transitions at different speeds (i.e., state-to-state time lags) in both forward and backward directions (29). Then, we obtained the sequence strength for each path by averaging their corresponding pairwise transitions (e.g., A1 → A2 and A2 → A3, for the A1 → A2 → A3 path). We used a conservative nonparametric 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- to 30-ms state-to-state time lag. Second, we found backward replay encompassing a 130- to 170-ms state-to-state lag (Fig. 4B; see also fig. S3C for individual sequence plots and 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 → A2 versus A2 → 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, whereas 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 six paths both before and during the main RL task. At either time, subjects' knowledge was not different across the six 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 subjects, replay strength of the six 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 lag 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

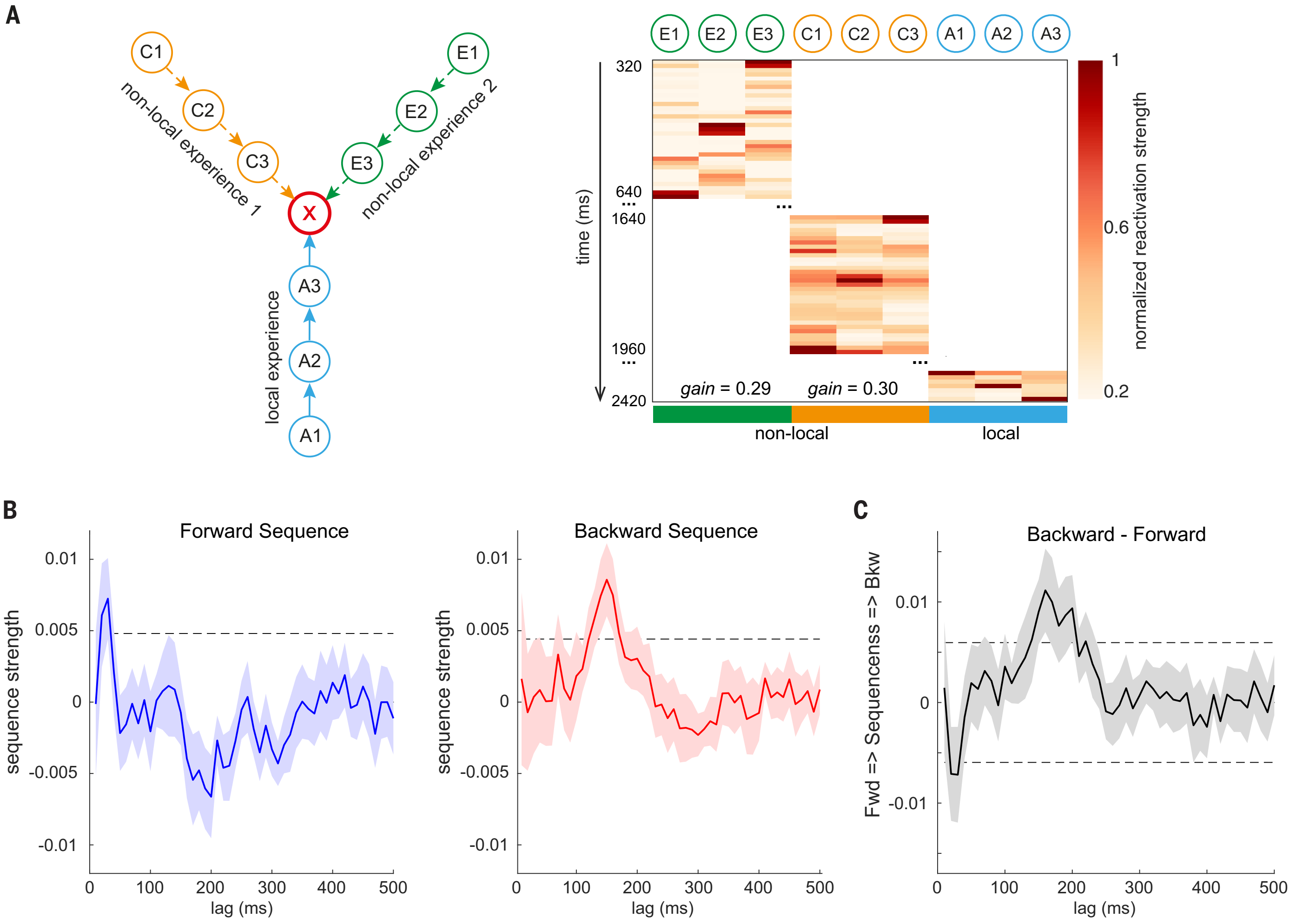


Fig. 4. Sequential replay of experiences during reward receipt. (**A**) An illustrative example of a trial in the main RL task (subject 14, trial 107). Left: The subject selected an A1 → A2 → A3 path, which renders A1 → A2 → A3 as the local experience and C1 → C2 → C3 and E1 → E2 → E3 as the two nonlocal experiences on this trial. Right: The state decoding matrix during outcome receipt time (e.g., getting £1 in X), along with the gain estimate for the two nonlocal paths. A backward 160-ms lag sequence for both C1 → C2 → C3 and E1 → E2 → E3, and a forward 30-ms lag sequence for A1 → A2 → A3, are depicted. For visualization purposes, the reactivation strength of each state is max-normalized. Each time bin is 10 ms. (**B**) Sequence analysis at outcome receipt time shows two distinct signatures: one forward sequence (blue) with a 20- to 30-ms state-to-state time lag (left) and a backward sequence (red) with a 130- to 170-ms time lag (right). (**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 dashed line is the permutation threshold after controlling for multiple comparisons in both (B) and (C). The two dashed lines in (C) represent the thresholds for forward > backward (negative value) and backward > forward (positive value), respectively. The shaded area in both (B) and (C) represents the 95% confidence interval.

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.

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]), although our results now extend those findings to a context that includes learning. The 160-ms backward replay has not been reported previously [but see ([24]) for memory replay at a similar speed]. This replay pattern is intriguing because 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], [33]).

If this 160-ms backward replay supports nonlocal updating, we would expect it to also represent the contents of nonlocal paths. In line with this prediction, the 160-ms backward replay significantly represented nonlocal paths [one-sample t test, $t(28) = 2.92$, $P = 0.007$]; moreover, it represented nonlocal paths 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 nonlocal paths [one-sample t test, $t(28) = -0.09$, $P = 0.93$] but likely represented the local one—that is, 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]–[36]). 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,

$t(28) = 3.98$, $P = 4.3 \times 10^{-4}$], 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-localizing MEG signals (37), we used beamforming technique, as an exploratory analysis, to look for the neural source of replays at the whole-brain level. The results indicated that although both replay types are associated with activation in the visual cortex and medial temporal lobe, the 30-ms forward replay displayed higher hippocampal activation—and, intriguingly, also higher activation in a region that encompasses the ventral tegmental area (VTA)—relative to the 160-ms backward replay. Conversely, the 160-ms backward replay displayed greater cortical engagement (fig. S6).

Nonlocal replay accompanies efficient nonlocal learning

Having identified neural candidates for learning, we tested whether nonlocal replay (i.e., the 160-ms backward replay) is associated with nonlocal learning and, if so, whether such replay is competitively prioritized between the two nonlocal paths in accord with theoretical accounts (27). We again posed these questions in terms of RL-based computational models of trial-by-trial choice behavior (29).

First, in asking whether replay accompanies nonlocal 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 nonlocal paths (as before, these are paths leading to the same end state), we tested whether the baseline learning rate for each nonlocal path was significantly increased on trials when that path exhibited significant neural replay versus when it did not. We found a higher nonlocal learning rate in the presence versus absence of significant 160-ms backward replay (29) [$\alpha_{\text{replay}} = 0.70$, $\alpha_{\text{no-replay}} = 0.61$; 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 versus without replay, $P = 0.60$ for 160-ms replay, $P = 0.88$ for 30-ms replay; table S3).

We next asked whether replay is prioritized to favor the more useful nonlocal experience. Recall that each trial has one local and two nonlocal paths. Thus, on each trial, we can classify the two nonlocal paths as high versus low priority. This priority can be computed on the basis of either the need (17%, 33%, and 50% for paths in the rare, occasional, and common arms, respectively), the gain (estimated per arm, per trial, and per subject from the behavioral model), or their product (need × gain) (fig. S7). According to RL theory (27), the need × gain interaction (i.e., utility) should determine the actual priority for replay. Indeed, we found that the strength of the 160-ms backward replay was significantly stronger for a high- versus low-utility (need × gain) path [paired t test, $t(28) = 3.30$, $P = 0.003$; Fig. 6A]. Such prioritization was absent in a high- versus low-need or high- versus low-gain comparison (fig. S7), nor did it exist for 30-ms replay [$t(28) = -0.34$, $P = 0.74$; Fig. 6A]. These prioritization results cannot be explained by differences in the actual frequency of encountering a given path (which was determined by subjects' own choices): We found no link between the 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 versus no-reward) paralleled these results (fig. S8).

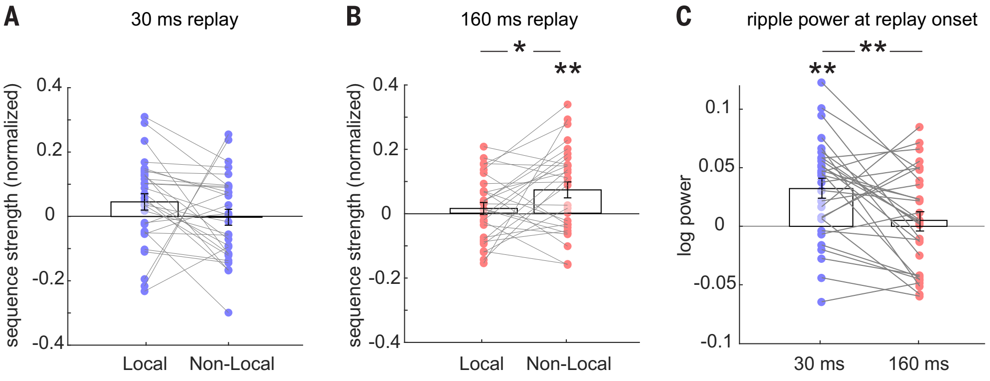


Fig. 5. Representational and physiological differences between the two types of replay. (**A**) A 30-ms forward sequence is likely to encode local experience but not nonlocal experience. (**B**) A 160-ms backward replay encodes nonlocal as opposed to local experience. (**C**) The initialization of a 30-ms forward sequence is associated with a power increase in a ripple frequency band (80 to 180 Hz), but this is not the case for a 160-ms backward sequence. These frequency power signatures are significantly different. The gray lines connect results from the same subject. Error bars show 95% SEM; each dot indicates results from each subject. *$P < 0.05$, **$P < 0.01$.

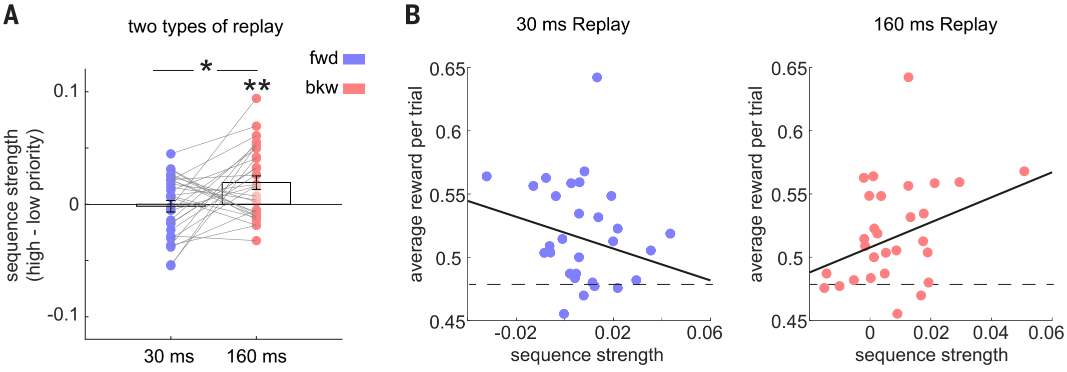


Fig. 6. Prioritization of nonlocal replay. (**A**) The 160-ms backward sequence is replayed to a greater degree in the higher-priority nonlocal path than in the lower-priority one. The 30-ms forward replay does not differentiate between the two nonlocal paths. Gray lines connect results from the same subject. Error bars show 95% SEM; each dot indicates results from each subject. *$P < 0.05$, **$P < 0.01$. (**B**) Left: Sequence strength of the 30-ms lag replay does not correlate with task performance. Right: By contrast, there is a significant positive correlation between the 160-ms lag replay and task performance across subjects. Each dot indicates a result from one subject. The solid line reflects the best robust linear fit; the dashed line indicates the chance level of reward rate per trial with random choices.

Finally, given that the 160-ms lag replay was associated with better trial-by-trial, within-subject nonlocal learning (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 nonsignificant 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 within-subject learning for either local ($P = 0.28$) or nonlocal experience ($P = 0.32$), nor that it was linked to task performance across subjects ($P = 0.18$).

Discussion

In the current study, we dissociated two types of replay as a function of local versus nonlocal learning. As a result, we established a connection between neural replay and learning through nonlocal credit assignment as expressed in behavior.

Replay of nonlocal experiences was associated with more effective learning of action values, as 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 RL. Furthermore, the content of this replay, and separately the strength of updating as expressed behaviorally, were prioritized according to their utility for future behavior (27).

These findings corroborate a long-standing hypothesis about the role of awake replay in 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, 33). These results also extend our previous functional magnetic resonance imaging results in humans, which linked nonlocal reactivation (without assessing sequences) to planning (4, 5, 38). 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 nonlocal learning is distinct from the 40-ms replay reported in previous studies (23, 39). Unlike the latter, the 160-ms replay is not associated with a ripple frequency power increase (23). This raises the intriguing possibility that the 160-ms replay, which has a state-to-state transition frequency of ~6 Hz, might be processing states on consecutive theta cycles, which may have connections to rodent theta sequences (40–43). However, theta sequences generally occur during ongoing behavior in rodents and are in a forward direction, akin to a "look ahead" signal [but see (44) for backward theta sequence], whereas the 160-ms sequence we identify is backward in direction and occurs at the end of a trial.

Although the 160-ms backward replay alone is associated with value learning in the present results, in Liu *et al.* (23) we observed a faster replay (30- to 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 relative to the 40-ms lag replay reported previously. This is plausible given that there was 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 40-ms lag replay previously observed may be similar to the 30-ms lag replay observed here, which might reflect a stereotyped recapitulation of recent experience. This interpretation is consistent with previous findings (39) in which the fast 40-ms lag sequences in a sequential planning task were shown to represent all possible transitions, rather than a specific planning trajectory.

The backward direction, representational contents, and timing of this reverse replay are well suited to solve the nonlocal credit assignment problem, where an outcome at the end of a path affects 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 is related to credit assignment (fig. S10; see table S4 for related modeling results). Such a process may play a role in other circumstances or in other task implementations—for instance, in games such as 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, and they reveal that nonlocal backward replay serves as a neural mechanism for model-based reinforcement learning.

Methods summary

See (29) for full materials and methods information.

Participants

All analyses included the full group of 29 subjects (mean age 23 ± 0.41 years; 17 females). 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) was determined on the basis of a prior power analysis where a one-sample t test required 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 because of contamination of metal on the MEG signal; pilot data were also excluded from formal analysis, leaving 29 subjects in total.

Stimuli and task design

In the current task, there were three starting arms, two end states, and 18 intermediate states (constituting six paths). All of them were indexed by distinct pictures. The mapping between stimuli and states was fixed within subject but randomized across subjects. The task was run in the following order: (i) functional localizer [to obtain neural representations of the 18 stimuli (i.e., six paths)]; (ii) model construction I: sequence learning (of the transition structures among six paths); (iii) model construction II: end-state learning (connections between six paths and two end states); (iv) model construction III: arm learning (connections between three starting arms and six paths); (v) model construction IV: arm frequency learning [occurrence probability (i.e., need) of each starting arm in the main RL task (45)]; (vi) main RL task (value learning, separating local and nonlocal experiences). In addition, need and gain were manipulated separately. Need was defined by the occurrence probability of the three starting arms (learned 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) following an independent Gaussian random walk across trials, bounded between 25% and 75% (30, 29, 46).

MEG data acquisition and preprocessing

The MEG data were collected while subjects sat upright, performing the task (with the exception of frequency learning). The data were recorded at 1200 samples/s 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 min. 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 were first high-pass filtered at 0.5 Hz and then downsampled to 100 Hz for later analyses (with the exception of temporal frequency analysis, for which the data were downsampled to 400 Hz, thereby preserving the ability to look for power change in high frequency, up to 200 Hz). After identification and removal of excessively noisy segments and sensors, the resulting MEG data were submitted to independent component analysis (ICA). The ICA was used for denoising purposes alone. In each scanning session, up to 10 independent components (150 in total) could be excluded if they were clearly noise as assessed by such properties as 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 localization).

Behavioral analysis and modeling

Choice behavior in the main RL task was analyzed as a function of reward (£1 or £0) at last trial and starting arm (same versus different compared to last trial) at current trial. The choice at current trial was binarized according to whether or not it led to the same end state as that of the last trial. A 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.

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

Neural decoding analysis

Classifiers for the 18 intermediate states (i.e., six paths) in the main RL task were trained on the basis of evoked visual response at 200 ms after stimulus onset (whole-brain sensor pattern) in the functional localizer task (*23*, *24*). During the functional localizer task, participants did not know either the mapping or occurrence probability of the stimuli and their 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 two end states were trained during the quiz question during end-state learning; classifiers for the three starting arms were trained during the quiz question during the arm learning. In all those quiz questions, the picture for either the end state or the starting arm was presented in the center 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 learned afterward.

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 at the 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 constitute six distinct paths. The reactivations of starting arm states or end states were not considered in the sequence analysis to avoid potential visual confound. Sequence strength of a pairwise 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 appropriate because neural representations (of different states) are only noisily and probabilistically decoded. The detailed approach and related simulations are described in (*22*). The same human replay detection procedure has been applied successfully in previous empirical work (*23*, *24*).

Sequence-behavioral modeling

We built a novel Q-learning model to formally test the effect of replay on learning. This separately models the learning rate for local versus nonlocal experience. Crucially, replay of a specific path (local or nonlocal) 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 a permutation test). The key comparison is the learning rate difference between α_{replay} for paths with significant replay and $\alpha_{\text{no-replay}}$ for paths without significant replay.

REFERENCES AND NOTES

1. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005). doi: 10.1038/nn1560; pmid: 16286932
2. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997). doi: 10.1126/science.275.5306.1593; pmid: 9054347
3. J. O'Doherty *et al.*, Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004). doi: 10.1126/science.1094285; pmid: 15087550
4. B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, N. D. Daw, Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015). doi: 10.1038/nn.3981; pmid: 25799041
5. G. E. Wimmer, D. Shohamy, Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012). doi: 10.1126/science.1223252; pmid: 23066083
6. P. A. Lewis, S. J. Durrant, Overlapping memory replay during sleep builds cognitive schemata. *Trends Cognit. Sci.* **15**, 343–351 (2011). doi: 10.1016/j.tics.2011.06.004; pmid: 21764357
7. K. Doya, What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974 (1999). doi: 10.1016/S0893-6080(99)00046-5; pmid: 12662639
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**, 484–489 (2016). doi: 10.1038/nature16961; pmid: 26819042
10. K. Doya, Reinforcement learning in continuous time and space. *Neural Comput.* **12**, 219–245 (2000). doi: 10.1162/089976600300015961; pmid: 10636940
11. R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull.* **2**, 160–163 (1991). doi: 10.1145/122344.122377
12. D. J. Foster, M. A. Wilson, Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006). doi: 10.1038/nature04587; pmid: 16474382
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). doi: 10.1126/science.271.5257.1870; pmid: 8596957
14. M. A. Wilson, B. L. McNaughton, Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994). doi: 10.1126/science.8036517; pmid: 8036517
15. G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, M. B. Zugaro, Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009). doi: 10.1038/nn.2384; pmid: 19749750
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. *Nat. Neurosci.* **18**, 493–495 (2015). doi: 10.1038/nn.3970; pmid: 25751533
17. I. Gridchyn, P. Schoenenberger, J. O'Neill, J. Csicsvari, Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron* **106**, 291–300.e6 (2020). doi: 10.1016/j.neuron.2020.01.021; pmid: 32070475
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). doi: 10.1016/j.neuron.2013.01.027; pmid: 23522050
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). doi: 10.7554/eLife.06063; pmid: 26112828
20. S. N. Gomperts, F. Kloosterman, M. A. Wilson, VTA neurons coordinate with the hippocampal reactivation of spatial experience. *eLife* **4**, e05360 (2015). doi: 10.7554/eLife.05360; pmid: 26465113
21. H. C. Barron *et al.*, Neuronal computation underlying inferential reasoning in humans and mice. *Cell* **183**, 228–243.e21 (2020). doi: 10.1016/j.cell.2020.08.035; pmid: 32946810
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 066407 [preprint]. 2 May 2020.
23. Y. Liu, R. J. Dolan, Z. Kurth-Nelson, T. E. J. Behrens, Human replay spontaneously reorganizes experience. *Cell* **178**, 640–652.e14 (2019). doi: 10.1016/j.cell.2019.06.012; pmid: 31280961
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. *Nat. Neurosci.* **23**, 1025–1033 (2020). doi: 10.1038/s41593-020-0649-z; pmid: 32514135
25. N. W. Schuck, Y. Niv, Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, eaaw5181 (2019). doi: 10.1126/science.aaw5181; pmid: 31249030

26. A. W. Moore, C. G. Atkeson, Prioritized sweeping: Reinforcement learning with less data and less time. *Mach. Learn.* **13**, 103–130 (1993). doi: 10.1007/BF00993104
27. M. G. Mattar, N. D. Daw, Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018). doi: 10.1038/s41593-018-0232-z; pmid: 30349103
28. H. Igata, Y. Ikegaya, T. Sasaki, Prioritized experience replays on a hippocampal predictive map for learning. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2011266118 (2021). doi: 10.1073/pnas.2011266118; pmid: 33443144
29. See supplementary materials.
30. 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). doi: 10.1016/j.neuron.2011.02.027; pmid: 21435563
31. C. J. Watkins, P. Dayan, Q-learning. *Mach. Learn.* **8**, 279–292 (1992). doi: 10.1007/BF00992698
32. 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. *Cereb. Cortex* **29**, 863–874 (2019). doi: 10.1093/cercor/bhy290; pmid: 30535141
33. 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). doi: 10.1016/j.neuron.2016.07.047; pmid: 27568518
34. M. F. Carr, S. P. Jadhav, L. M. Frank, Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* **14**, 147–153 (2011). doi: 10.1038/nn.2732; pmid: 21270783
35. K. Diba, G. Buzsáki, Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242 (2007). doi: 10.1038/nn1961; pmid: 17828259
36. S. P. Jadhav, C. Kemere, P. W. German, L. M. Frank, Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454–1458 (2012). doi: 10.1126/science.1217230; pmid: 22555434
37. 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). doi: 10.1016/j.neuroimage.2005.10.037; pmid: 16368248
38. I. Momennejad, A. R. Otto, N. D. Daw, K. A. Norman, Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018). doi: 10.7554/eLife.32548; pmid: 30547886
39. Z. Kurth-Nelson, M. Economides, R. J. Dolan, P. Dayan, Fast Sequences of Non-spatial State Representations in Humans. *Neuron* **91**, 194–204 (2016). doi: 10.1016/j.neuron.2016.05.028; pmid: 27321922
40. G. Buzsáki, Theta oscillations in the hippocampus. *Neuron* **33**, 325–340 (2002). doi: 10.1016/S0896-6273(02)00586-X; pmid: 11832222
41. M. R. Mehta, A. K. Lee, M. A. Wilson, Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417**, 741–746 (2002). doi: 10.1038/nature00807; pmid: 12066185
42. B. E. Pfeiffer, D. J. Foster, Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013). doi: 10.1038/nature12112; pmid: 23594744
43. K. Kay *et al*., Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552–567.e25 (2020). doi: 10.1016/j.cell.2020.01.014; pmid: 32004462
44. M. Wang, D. J. Foster, B. E. Pfeiffer, Alternating sequences of future and past behavior encoded within hippocampal theta oscillations. *Science* **370**, 247–250 (2020). doi: 10.1126/science.abb4151; pmid: 33033222
45. 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 Comput. Biol.* **13**, e1005768 (2017). doi: 10.1371/journal.pcbi.1005768; pmid: 28945743
46. O. M. Vikbladh *et al*., Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683–693.e4 (2019). doi: 10.1016/j.neuron.2019.02.014; pmid: 30871859
47. Y. Liu, M. Mattar, T. Behrens, N. Daw, R. Dolan, Data from "Experience replay is associated with efficient nonlocal learning." Zenodo (2021); DOI: 10.5281/zenodo.4597119.

ACKNOWLEDGMENTS

We thank R. Moran and E. Russek for discussions on study design and analysis. **Funding:** Supported by the Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning (Y.L.); a Wellcome Trust Investigator Award (098362/Z/12/Z) (R.J.D.); a Lundbeck Visiting Professorship (R290-2018-2804) to the Danish Research Centre for Magnetic Resonance (R.J.D.); 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) (T.E.J.B.); NSF grant IIS-1822571, part of the CRCNS program (N.D.D. and M.G.M.); and a Newton International Fellowship of the Royal Society (grant 181426) (M.G.M.). **Author contributions:** Y.L., M.G.M., T.E.J.B., N.D.D., and R.J.D. contributed to conception and design of the study; Y.L. contributed to data acquisition with help from M.G.M.; Y.L., M.G.M., and N.D.D. contributed to data analysis; and 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 (*47*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/372/6544/eabf1357/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S4
References (*48*–*56*)

20 October 2020; accepted 15 April 2021
10.1126/science.abf1357

Science

Experience replay is associated with efficient nonlocal learning

Yunzhe Liu, Marcelo G. Mattar, Timothy E. J. Behrens, Nathaniel D. Daw and Raymond J. Dolan

Science **372** (6544), eabf1357.
DOI: 10.1126/science.abf1357

Replay supports planning

Learning from direct experience is easy––we can always use trial and error––but how do we learn from nondirect (nonlocal) experiences? For this, we need additional mechanisms that bridge time and space. In rodents, hippocampal replay is hypothesized to promote this function. Liu *et al.* measured high-temporal-resolution brain signals using human magnetoencephalography combined with a new model-based, visually oriented, multipath reinforcement memory task. This task was designed to differentiate local versus nonlocal learning episodes within the subject. They found that reverse sequential replay in the human medial temporal lobe supports nonlocal reinforcement learning and is the underlying mechanism for solving complex credit assignment problems such as value learning.

Science, abf1357, this issue p. eabf1357

ARTICLE TOOLS	http://science.sciencemag.org/content/372/6544/eabf1357
SUPPLEMENTARY MATERIALS	http://science.sciencemag.org/content/suppl/2021/05/19/372.6544.eabf1357.DC1
REFERENCES	This article cites 54 articles, 11 of which you can access for free http://science.sciencemag.org/content/372/6544/eabf1357#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.