Behavioural and computational evidence for memory consolidation biased by reward-prediction errors

Author names: Emma L. Roscow* [1]; Matthew W. Jones[1]; Nathan F. Lepora[2]

Author affiliations: [1] School of Physiology, Pharmacology & Neuroscience, University of Bristol, University Walk, Bristol BS8 1TD, UK; [2] Department of Engineering Mathematics, University of Bristol, Bristol, UK

* corresponding author

Abstract

Neural activity encoding recent experiences is replayed during sleep and rest to promote consolidation of the corresponding memories. However, precisely which features of experience influence replay prioritisation to optimise adaptive behaviour remains unclear. Here, we trained adult male rats on a novel maze-based reinforcement learning task designed to dissociate reward outcomes from reward-prediction errors. Four variations of a reinforcement learning model were fitted to the rats' behaviour over multiple days. Behaviour was best predicted by a model incorporating replay biased by reward-prediction error, compared to the same model with no replay; random replay or reward-biased replay produced poorer predictions of behaviour. This insight disentangles the influences of salience on replay, suggesting that reinforcement learning is tuned by post-learning replay biased by reward-prediction error, not by reward per se. This work therefore provides a behavioural and theoretical toolkit with which to measure and interpret replay in striatal, hippocampal and neocortical circuits.

1 Introduction

To make good decisions, it is typically beneficial to use past experience to guide future behaviour. Actions which have previously produced good outcomes in a similar context can be reinforced to adapt behaviour for maximising benefit. Crucial to this mechanism is the ability for neuronal spiking activity to drive synaptic plasticity, strengthening the synaptic connections between neurons to establish functional networks which encode task-relevant information or drive task-relevant actions. These functional networks are refined during sleep and rest, when many neurons switch to an "offline" state in which they replay activity encoding previous or anticipated upcoming experiences rather than current events or behaviours (Yu et al. 2017). This offline replay, found across cortical, limbic and basal ganglia regions, has been suggested to play a role in decision-making (Pfeiffer and Foster 2013), emotional processing (Cairney et al. 2014), generalising across episodes (Lewis and Durrant 2011), and reinforcement learning (Dupret et al. 2010).

Studies in which replay has been manipulated provide strong evidence for its contributions to memory consolidation. Artificially enhancing replay by presenting odours or sounds during sleep, which had previously been paired with object locations or visual stimuli, leads to better subsequent recall of the paired stimuli (Rasch et al. 2007; Rudoy et al. 2009; Antony et al. 2012; Bendor and Wilson 2012). Disrupting replay events, meanwhile, impairs subsequent spatial memory (Girardeau et al. 2009; Ego-Stengel and Wilson 2010; Jadhav et al. 2012; Michon et al. 2019).

An examination of how replay aids these cognitive processes requires assessment of which activity is replayed with greatest strength or frequency. Activity which is associated with experiences of reward (Foster and Wilson 2006; Lansink et al. 2009; Singer and Frank 2009) or fear (Girardeau et al. 2017; Wu et al. 2017), or with recent experiences (Cheng and Frank 2008), is replayed preferentially. This suggests a replay bias towards the most salient experiences to be processed, consolidated or incorporated into the internal model of the world. However, these salient experiences could also be interpreted as those with the highest prediction error, i.e. the most informative experiences for updating internal models and for reinforcement learning. Tasks which involve learning the locations of rewards often conflate reward with reward-prediction error (RPE), leading to the possibility that apparent replay biases towards reward actually reflect biases towards RPE.

Here we explore the possibility that it is reward prediction errors, rather than reward or salience, which biases replay. We used variations of a reinforcement learning model, Q-learning, to estimate the value of actions encoded in the striatum during a reinforcement learning task, and varied the amount and type of replay in the model to predict behaviour. In the striatum, representations of reward values differ following learning acquired over weeks compared to when acquired over minutes (Wimmer et al. 2018), and, correspondingly, reward-responsive cells are replayed preferentially in the ventral striatum (Lansink et al. 2009). We therefore propose that replay triggers value updates in the striatum, to enhance striatum-dependent reinforcement learning, and moreover that activity encoding events that resulted in high RPE is preferentially replayed.

Q-learning (Watkins 1989) has been used successfully to model reinforcement learning, particularly in humans (O'Doherty et al. 2003, Daw et al. 2005) but also in rodents (Kim et al. 2013, Ito and Doya 2009). Q-learning models fit both behavioural outcomes and striatal activity, suggesting that they describe mechanisms of updating values in the striatum in response to RPEs which in turn guide behaviour (Day et al. 2014, Morris et al. 2010,

Pagnoni et al. 2002, Roesch et al. 2007). Temporal-difference-based RPEs, i.e. the difference between expected reward and actual reward which drives the update of Q-values, resemble quite closely the dopaminergic input of ventral tegmental area (VTA) to the striatum (McClure et al. 2003, Roesch et al. 2007, Schultz 2016), which mediates synaptic plasticity in the striatum (Calabresi et al. 2007) and may provide a mechanism of biological implementation of Q-learning. Dyna-Q (Sutton 2014), a variant of Q-learning which incorporates offline temporal-difference updates, has been used to model replay in ways which produce learning qualitatively similar to animal reinforcement learning (Johnson and Redish 2005). RPE-biased replay incorporated into machine learning algorithms show that it can also be very efficient, learning to play Atari games (Andrychowicz et al. 2017) or navigate a simulated environment (Karimpanal and Bouffanais 2017) faster and with more success compared to replay without such a bias.

We trained 6 rats on a stochastic reinforcement learning task which elicited both positive and negative RPE, and fitted Q-learning parameters to each rat's behavioural data. We then included replay events between sessions, to simulate the effect of replay during sleep on reinforcement learning. Four replay policies were compared, prioritising state-action pairs to be updated according to different biases: random replay, replay proportional to expected reward, and two forms of RPE-biased replay. Random replay was included as a control, while reward-biased replay reflects the prevailing view of how replay is prioritised. Fitting the model parameters showed that the two RPE-biased replay policies increased the model's predictive accuracy, while random and reward-biased replay impaired model performance. This suggests that replay between sessions of a probabilistic reinforcement learning task in rats is biased by RPE and not by reward.

2 Results

Animals successfully learned a stochastic reinforcement learning task

Six rats were trained to forage for stochastic sucrose rewards on a three-armed maze, to assess their reinforcement learning on a task where reward outcome and reward-prediction error (RPE) were dissociable. Each arm was assigned as either "high probability", "mid probability" or "low probability", which determined the protocol for reward delivery (fig. 1a). For the first 15 training sessions, the high-probability arm delivered a reward on 75% legitimate entries to the arm, the mid-probability arm on 50%, and the low-probability arm on 25%. A legitimate entry was one in which a different arm had been entered on the previous trial; entering the same arm twice in a row was incorrect and did not result in a reward delivery. For sessions 16-20, the reward probabilities for the high- and low-probability arms were amplified: reward was delivered on 87.5% and 12.5% legitimate entries respectively. For sessions 21-22 the reward probabilities for the high- and low-probability arms were switched, such that the (formerly) high- and low- probability arms delivered reward on 12.5% and 87.5% of legitimate entries respectively. This set-up meant that receiving a reward in a low-probability arm would elicit a higher RPE than the same reward value in a high-probability arm, so reward outcome and RPE could be dissociated.

Over 22 sessions, animals learned to distinguish between the high-, mid- and low-probability arms in their frequency of visits to each arm, indicating successful learning of the reward probabilities. Rats performed 45.1 $\pm$ 2.5 trials per session, eventually showing a significant preference for the high-probability arm and against the low-probability arm, evident by session 6 and stable by session 10. The six animals varied in the degree of their discrimination between the arms (fig. 1b), but on average they distinguished between all arms on 13 out of 22 sessions (fig. 1c; χ^2 test, Bonferroni-corrected), visiting the arms which delivered a higher probability of reward more often, primarily in later sessions. The differences in arm discrimination between animals may be accounted for by the orientation of the maze in the room; for example, animals may have shown a confounding preference for the arm which was closest to the door of the recording room, an effect which was overcome by rotating the arm probabilities between animals. (For rats H and K, the mid-probability arm was closest to the door; for rats I and L, the high-probability arm was closest to the door; and for rats J and M, the low-probability arm was closest to the door.)

To quantify performance on the task, each trial was coded as optimal or suboptimal according to the animal's choice of arm given the arm most recently visited. Because no reward was given for re-entering the same arm visited on the previous trial, the optimal action choice following a visit to the mid- or low-probability arm was to visit the high-probability arm; the optimal action following the high-probability arm was the mid-probability arm. Over sessions, animals increased the proportion of trials on which they behaved optimally, achieving performance significantly above chance from session 3 onwards (fig. 1d, 46 trials optimal out of 106, $p = 0.02$, binomial test, Bonferroni-corrected).

Reward probabilities were changed twice over the course of learning, triggering clear changes in behaviour. In

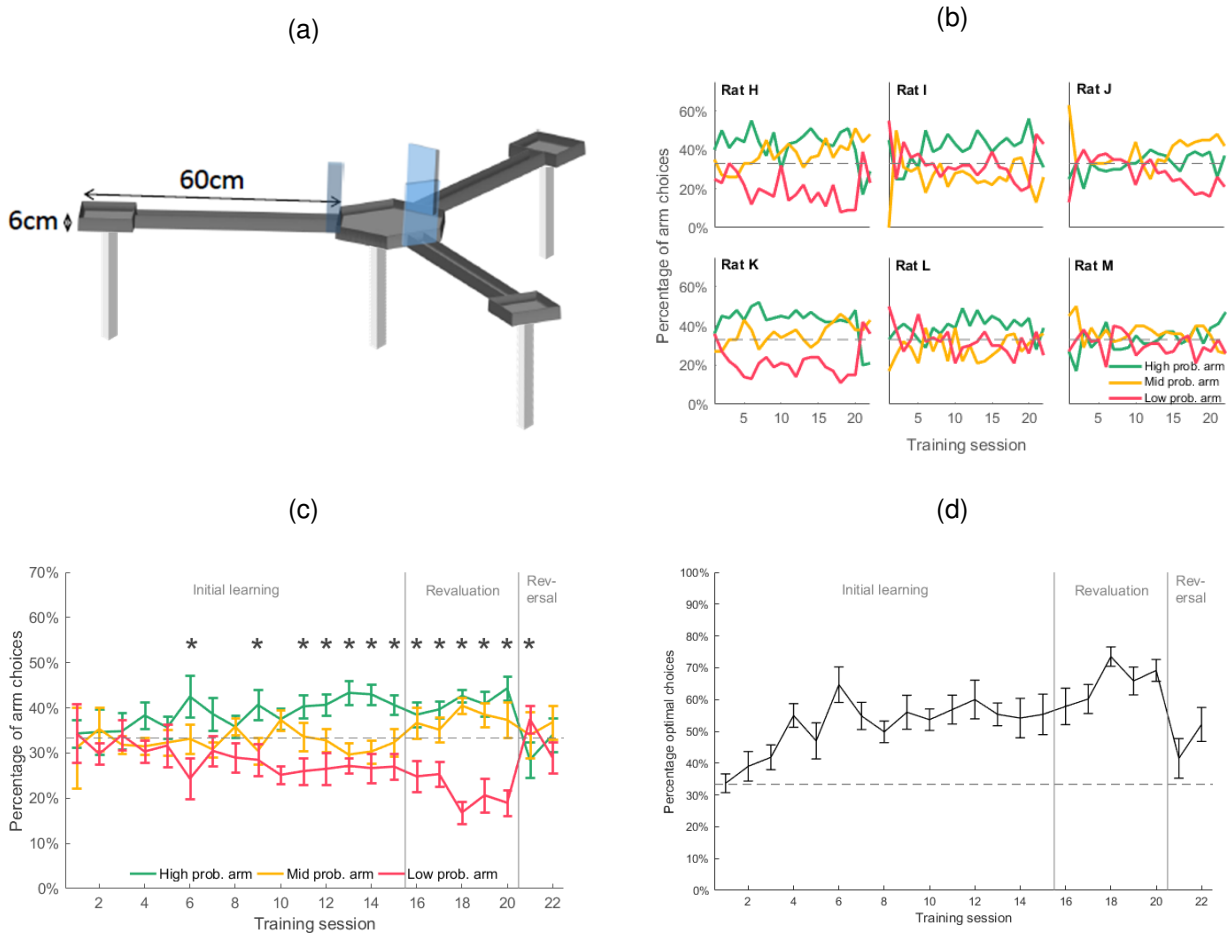


Figure 1: **A.** Illustration of the maze used to train animals. Lick ports located at the end of each arm delivered reward with either high, medium or low probabilities. **B.** Frequency of entry to each arm over all sessions, shown separately for each rat. **C.** Average frequency of entry to each arm. * indicates arm choices statistically different from each other (χ^2 test). **D.** Mean proportion of trials on which the optimal arm was chosen, according to highest probability of reward. Dashed lines represent chance level (33.3%). Error bars represent standard error of the mean (s.e.m.).

the revaluation learning stage (sessions 16-20), the reward probabilities at each arm became more distinct: the high-probability arm delivering an 87.5% probability of reward compared to 75% in the initial learning stage, and the low-probability arm delivering a 12.5% probability of reward compared to 25% in the initial learning stage. This change offered a higher incentive to avoid the low-probability arm and, correspondingly, preference for the high-probability arm over the low-probability arm increased compared to the previous five sessions (fig. 1c; repeated-measures ANOVA, $F = 8.7$, $p = 0.006$). As a result, the rate of optimal performance was also greater in the revaluation stage than the last five sessions of the initial learning stage (fig. 1d; repeated-measures ANOVA, $F = 15.2$, $p = 0.001$).

The definition of optimal behaviour was the same in the initial and revaluation learning stages, because the arms did not change. However, optimal behaviour required a different behavioural policy in the reversal learning stage (sessions 21-22) when the high- and low-probability arms were switched. As expected, optimal performance correspondingly dipped when reward probabilities were reversed in sessions 21 to 22 as this new behavioural policy was learned. The frequency of optimal arm choices was lower for the reversal learning stage than the last two sessions of the revaluation stage (repeated-measures ANOVA post-hoc test, $p = 0.17$), although it did not differ significantly from the last two sessions of the initial learning stage (repeated-measures ANOVA

post-hoc test, $p > 0.05$). These behavioural data demonstrate that reward probabilities successfully influenced learning and behaviour in the task, and that animals were capable of showing flexibility in response to changing reward. We therefore went on to test whether reinforcement learning algorithms were able to recapitulate rat behaviour and whether instantiating between-session ("offline") replay of different task features improved model performance.

Q-learning modelled animal behaviour

We trained a Q-learning algorithm with no replay to generate probabilities of each action for each trial, based on Q-values estimated from the animals' previous experience (fig. 2). Q-learning is a reinforcement learning algorithm in which an agent selects actions in its environment and observes the outcome, recording at each time step t its starting state s_t, selected action a_t, resulting reward r_t, and resulting state s_{t+1}. The agent builds up a matrix Q of Q-value estimates for every state-action pair:

$$\begin{bmatrix} Q_{s_1,a_1} & Q_{s_1,a_2} & \cdots & Q_{s_1,a_A} \\ Q_{s_2,a_1} & Q_{s_2,a_2} & \cdots & Q_{s_2,a_A} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s_S,a_1} & Q_{s_S,a_2} & \cdots & Q_{s_S,a_A} \end{bmatrix} \tag{1}$$

corresponding to the future discounted expected reward, i.e. the temporal difference between the current state and the reward state. These Q-value estimates are used to guide actions to maximise reward. At each time step t, the Q-value for the state-action pair observed is updated by:

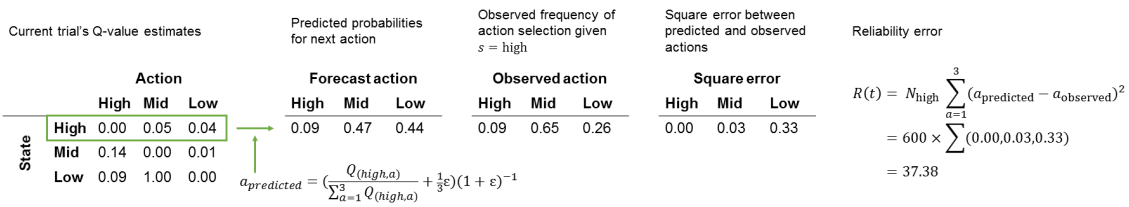
$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot (r_t + \gamma \cdot \max Q(s_{t+1}, a)) \tag{2}$$

where $\alpha \in (0, 1)$ is a learning rate parameter which determines the degree to which new information overrides old information, and $\gamma \in (0, 1)$ is a discount parameter which determines the importance of long-term gains.

In this task, entries into a chosen arm (and arrival at the goal location at the end of the arm) were modelled as actions, while the arm entered on the previous trial, on which reward probabilities were contingent, were modelled as states. Each trial therefore gave rise to one state-action transition out of nine possible state-action pairs.

For each trial, a matrix of Q-values for all state-action pairs was updated based on experience and used to calculate predicted action probabilities, which were compared to the observed frequencies of state-action pairs to produce a vector of errors for the three available actions. A reliability error was calculated from the summed square of the error vector, weighted by the prevalence of the state. This produced a measure of how reliably the Q-value estimates predicted behaviour (fig. 2; see Materials and Methods).

(a)



(b)

Figure 2: Example of model prediction for one trial, t = 100, in which rat H had most recently visited the high-probability arm (s = high) and chose the mid-probability arm (a = mid). **A.** The far left table shows the Q-learning model's estimate of the Q-values based on rat H's experience to date. Other tables show the predicted action probabilities calculated from the Q-values, the ground-truth of observed action frequencies over all visits to this state, and the mean square error between them. Far right shows how the error for this trial is calculated. **B.** A cartoon illustration of the same trial.

Observed action frequency correlated well with predicted action probabilities (fig. 3a), indicating a good baseline model for reinforcement learning. Predicted action probabilities were binned in 100 percentile-bins for each animal, and for each bin the average frequency of these actions occurring was compared to the average predicted probability, resulting in a strong correlation ($r = 0.92$, $p = 7.8e^{-08}$, Pearson's correlation). This result was consistent across animals (correlations ranging from $r = 0.86$ to $r = 0.96$).

The error between predicted action probability and observed action frequency spanned a large range, which was greatest in the earlier training sessions and diminished towards 0 for later training sessions as Q-values were learned (fig. 3b; early trials in blue have larger errors).

Reliability errors spanned a different range for each animal (fig. 3c), so all further analysis was performed on reliability errors normalised by the mean reliability error for each animal. On this measure, normalised reliability errors were similarly highest in early training sessions, when behaviour is least optimal and most unpredictable. Following this, reliability errors became consistently low for most sessions (fig. 3d), confirming a consistent fit with behaviour which captured the learning process over multiple sessions and changes in reward probabilities.

As described in Materials and Methods, the reliability error was used as the cost function to optimise three parameters in the Q-learning algorithm for each animal: a learning rate α, a discount factor γ, and an exploration factor ϵ. The resulting optimised parameter values are shown in table 1. A perturbation analysis was performed

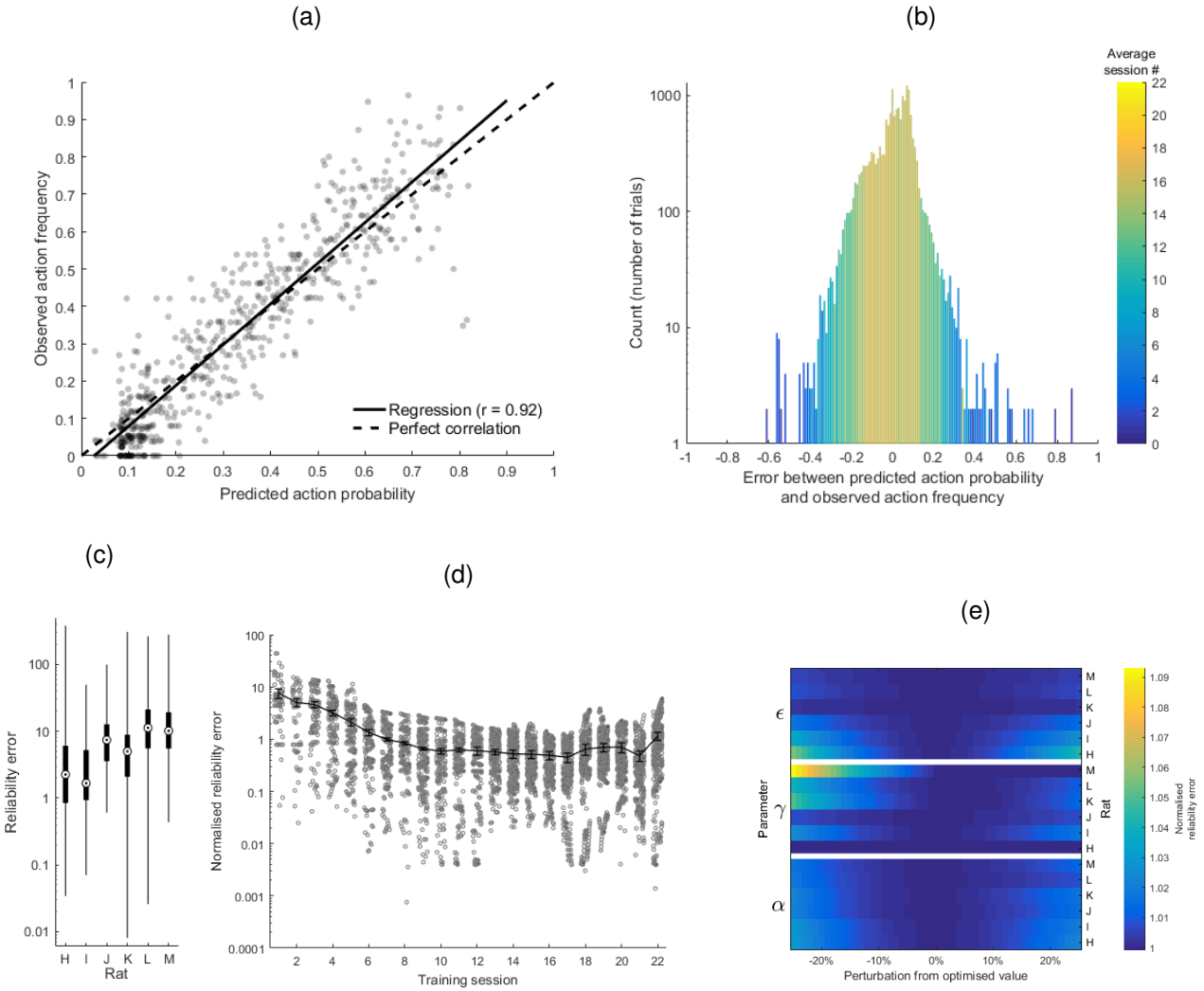


Figure 3: **A.** Reliability diagram (trials pooled across all animals). Observed frequency indicates how often an action was chosen by the animal, averaged over similar predicted action probabilities. Colour scale indicates the average session in which the predicted action probabilities occurred, for each bin; blue data points indicate predicted actions generally early in learning while yellow data points indicate predicted actions generally late in learning. Solid line represents regression (r = 0.92, p=7.8e^{-244}); dashed line indicates perfect correlation. **B.** Histogram of residuals of the data in A. Colour scale indicates on average what session the residuals within each bin occurred in. **C.** Range of reliability errors (calculated from residuals) for each animal. A reliability error of 0 reflects perfect modelling of action choices. Boxes represent 25th and 75th percentiles, circles represent median. **D.** Reliability errors for each trial grouped into training sessions, normalised to the average reliability error for each animal (shown in table 1). Data points show normalised reliability error for all trials; solid line represents mean for all animals. Error bars represent s.e.m. **E.** Change in reliability error, normalised to the optimised reliability error for each animal, with varying perturbations to the optimised parameter values. The optimised values for learning rate α, discount factor γ and exploration factor ϵ were individually perturbed by 1%-25% above and below the optimised value and the Q-learning algorithm was trained on behavioural data according to the perturbed parameter values 1,000 times to obtain an average.

to verify that the Q-learning results were sufficiently insensitive to perturbations to the optimised parameter values. At the optimised values, the average normalised reliability error over all trials was, by definition, 1. Perturbing these values by up to 25% in either direction increased the normalised reliability error by less than 0.05 in most cases (fig. 3e) and less than 0.1 in all cases, indicating that reliability errors were not overly sensitive to small changes in parameter values.

	α	γ	ϵ	**Reliability error**
Rat H	0.009470	3.340e-09	0.3451	8.688
Rat I	0.01399	0.2972	0.4035	4.355
Rat J	0.02591	0.5153	0.3173	10.08
Rat K	0.06887	1.000	0.09363	10.66
Rat L	0.6522	1.000	0.3117	18.72
Rat M	0.1345	1.000	0.3137	16.92

Table 1: Optimised parameter values for Q-learning algorithm trained on each animal's behavioural data. α is the learning rate, γ is the discount factor, and ϵ is the exploration factor.

In summary, the Q-learning algorithm proved able to recapitulate rat behaviour over the course of training and adaptation to new task conditions. The model was robust across a range of parameter values and established a sound basis on which to quantify the effects of mimicking replay by updating Q values between sessions.

Adding RPE-biased replay to the Q-learning model improved prediction accuracy, whereas reward-biased and random replay both reduced accuracy

Against the baseline of no-replay, a variant of the Q-learning algorithm with replay was trained on the same data, with a specified number of samples chosen from all the trials experienced so far to be replayed between each session. Q-learning parameters were optimised for a fixed $(1 \leq n \leq 100)$ number of replay events between each session, for each replay policy. All trials experienced by the animal were stored in a memory buffer, and for each replay event a state-action pair was chosen according to the replay policy and a sample trial from this state-action pair was used to update its Q-value. With a random replay policy, all state-action pairs that had been experienced were sampled at random. With a reward-biased replay policy, state-action pairs were sampled in proportion to their Q-values, so that state-action pairs at which rewards had been experienced most frequently would be replayed most. With an RPE-prioritised replay policy, the state-action pair with the highest recent average RPE was sampled. With an RPE-proportional replay policy, state-action pairs were sampled in proportion to their recent average RPE. These latter policies offered two variations on preferentially updating state-action value(s) which had generated the greatest errors, concentrating efforts on correcting the most erroneous expectations of reward.

Compared to the no-replay Q-learning baseline, replay biased by RPE produced a more reliable model of learning, while replay that was random or biased by reward produced a less reliable model (fig. 4a; orange and purple compared to blue and green). Both the random and reward-biased replay policies resulted in higher reliability errors (p=8.8e^{-11} random, p=1.6e^{-08} reward-biased, Wilcoxon signed rank test, Bonferroni-corrected), even with a small amount of replay. Conversely, both the RPE-biased replay polices resulted in lower reliability errors (p = 6.6e^{-12} RPE-prioritised, p = 6.3e^{-10} RPE-proportional). This was true even when one additional sample was replayed between sessions (fig. 4b) and remained true when more samples were replayed between sessions (fig. 4c-4e). Replay of information encoded during trials associated with the most unexpected outcomes therefore significantly improved learning in the model, whereas replay of rewarded trials proved detrimental.

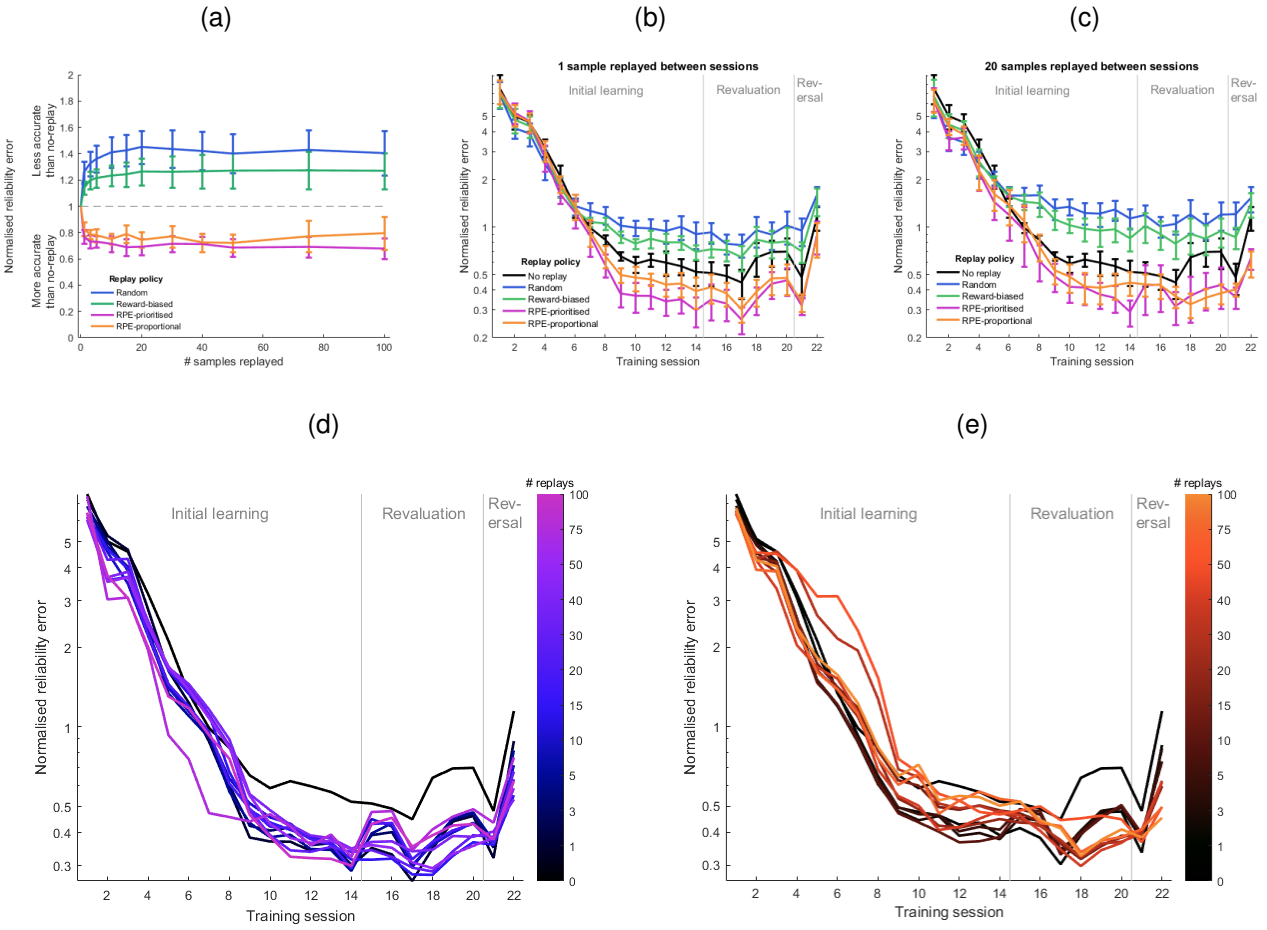


Figure 4: **A.** Normalised reliability error with varying numbers of samples replayed between sessions, averaged over all trials, according to the four replay policies shown. Reliability errors normalised to the average reliability with no replay, for each animal. Dashed line represents baseline with no replay. **B-C.** Average reliability error for each session, normalised to the average reliability error for no-replay for each animal. With 1 sample replayed between each session (B.) and 20 samples replayed between each session (C.). Error bars represent s.e.m. **D-E.** Average normalised reliability error for each session, with varying numbers of samples replayed. D. RPE-prioritised replay policy. E. RPE-proportional replay policy.

The superiority of the two RPE-biased replay policies was not uniform over the whole training period, however, and two patterns emerged. First, all replay policies showed improvements over no-replay in early sessions, but this effect disappeared in the random and reward-biased policies after roughly the seventh session. This initial superiority of all replay policies over no-replay cannot be due to replay itself because it begins in session 1, before any replay has taken place in the model; rather, it must be due to the non-replay parameters. Specifically, the optimised exploration parameter ϵ was higher in all replay policies than no-replay, so it may be the case that animals tended more towards exploration and relied on Q-values less in early training sessions. The higher ϵ value in the replay policies therefore better modelled behaviour in early sessions, whereas the differences in Q-values resulting from different replay policies impacted behaviour only later.

The second notable pattern is the fluctuations in the reliability errors over training sessions. In the no-replay baseline, reliability error increased in sessions 18-20 and in session 22 (t=3.54, p=1.8$e.^{-3}$, t-test compared to reliability error in sessions 15-17 and session 21). This mirrors an increase in optimal behaviour in these

sessions during the revaluation stage and reversal stage respectively, suggesting that the model failed to capture subtleties in the learning pattern at these points when animals were adapting their behaviour to changes in reward probabilities. As animals re-evaluated the state-action pairs in sessions 18-20 and adjusted their behaviour accordingly, replay by any policy was sufficient to overcome the increase in reliability error seen in the baseline, so there was no increase at these sessions (fig. 4c; p=0.37 for random replay, p=0.94 for reward-biased replay, p=0.081 for RPE-prioritised replay, p=0.06 for RPE-proportional replay with 20 samples replayed, sessions 18-20 compared to sessions 15-17). This may reflect the faster learning enabled by replaying recently experienced trials. However, as animals reversed their behaviour in session 22, requiring a substantial update to Q-values and a dramatic change in behaviour, increased random replay or reward-biased replay did not improve reliability error. With increased RPE-prioritised or RPE-proportional replay, on the other hand, increasing replay had a particularly strong effect on improving reliability error in session 22 (fig. 4d-4e). This raises the possibility that RPE-biased replay is especially important for behavioural flexibility of the kind seen in the reversal learning stage.

RPE-biased replay did not improve predictions when trained on shuffled data

Given the indication that replay might play different roles in different learning stages, it is important to control for the possibility that parameter values were optimised for the general statistics of rewards and actions in the task, rather than truly modelling the learning curve. Otherwise, the apparent superiority of RPE-biased replay may result from anomalous irregularities in the learning patterns and not true cognitive processes. Therefore, the same algorithms were trained on shuffled behavioural data in which the order of trials was randomly permuted 1,000-fold. This preserved the average frequency of state-action pairs and their associated rewards, as well as the lengths of training sessions, but altered the learning curve including revaluation and reversal learning.

Overall, the reliability errors for Q-learning with no replay were lower for shuffled data than real data, because shuffled behaviour was necessarily more consistent over time and therefore more predictable. Similarly to real data, reliability errors decreased sharply in early training sessions before reaching an asymptotic level (fig. 5), because Q-values in early training sessions were distorted by unrepresentative rewards as a result of a small sample size of trials experienced. Unlike real data, the approach to asymptotic reliability error was smooth and monotonic.

Crucially, compared to the no-replay baseline, no replay policy improved reliability error. All replay policies resulted in higher normalised reliability errors than no-replay (p=$6.9e^{-6}$ random, p=$6.9e^{-6}$ reward-biased, p=$1.6e^{-5}$ RPE-prioritised, p=$3.4e^{-5}$). This confirms that the improvement in reliability error in the real data is a result of better predictions of the learning process, and not better convergence to general statistics in the task.

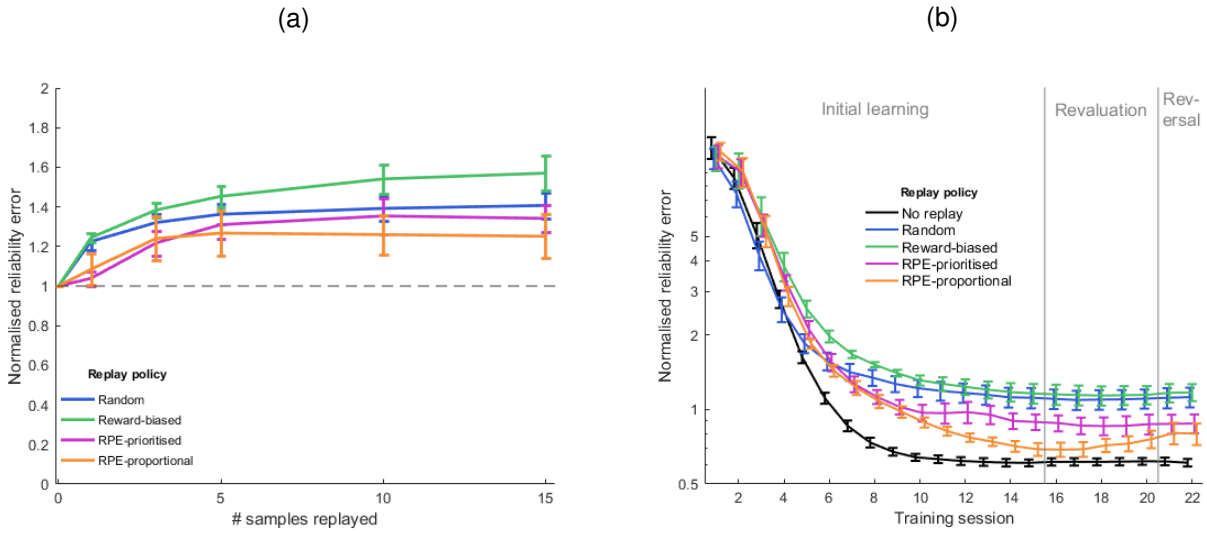


Figure 5: **A.** Normalised reliability error with varying numbers of samples replayed between sessions, trained on shuffled data in which trial data (state, action and reward) are randomly permuted. Dashed line represents baseline with no replay. **B.** Average reliability error for each session of shuffled data, normalised to the average reliability error for no-replay for each animal, with 15 samples replayed between each session. Error bars represent s.e.m.

Replay-biased RPE was the best predictor for all state-action pairs

We next accounted for the skew in training data towards the state-action pairs that were chosen most frequently. The transition from the high-probability arm to the mid-probability arm and vice versa (as they were in the initial and revaluation learning stages) were the most commonly experienced state-action pairs, representing 42% of trials overall, and the reliability error was weighted by the frequency of each state such that errors in the more common states contributed more to the overall reliability error than errors in the less common states. We therefore confirmed that Q-learning with RPE-biased replay learned to correctly predict all actions and not just the more-frequently chosen actions to which the cost function was skewed.

Figure 6 shows the improvement in reliability errors for each replay policy over no-replay baseline, for each state-action pair separately. Despite the skew in training data, the RPE-biased replay policies outperformed random and reward-biased replay policies for every state-action pair, although the improvement was not identical in each case. Nevertheless, the broad conclusion can be reached that RPE-biased replay policies better predicted learning than either no-replay, random replay or reward-biased replay for all state-action pairs.

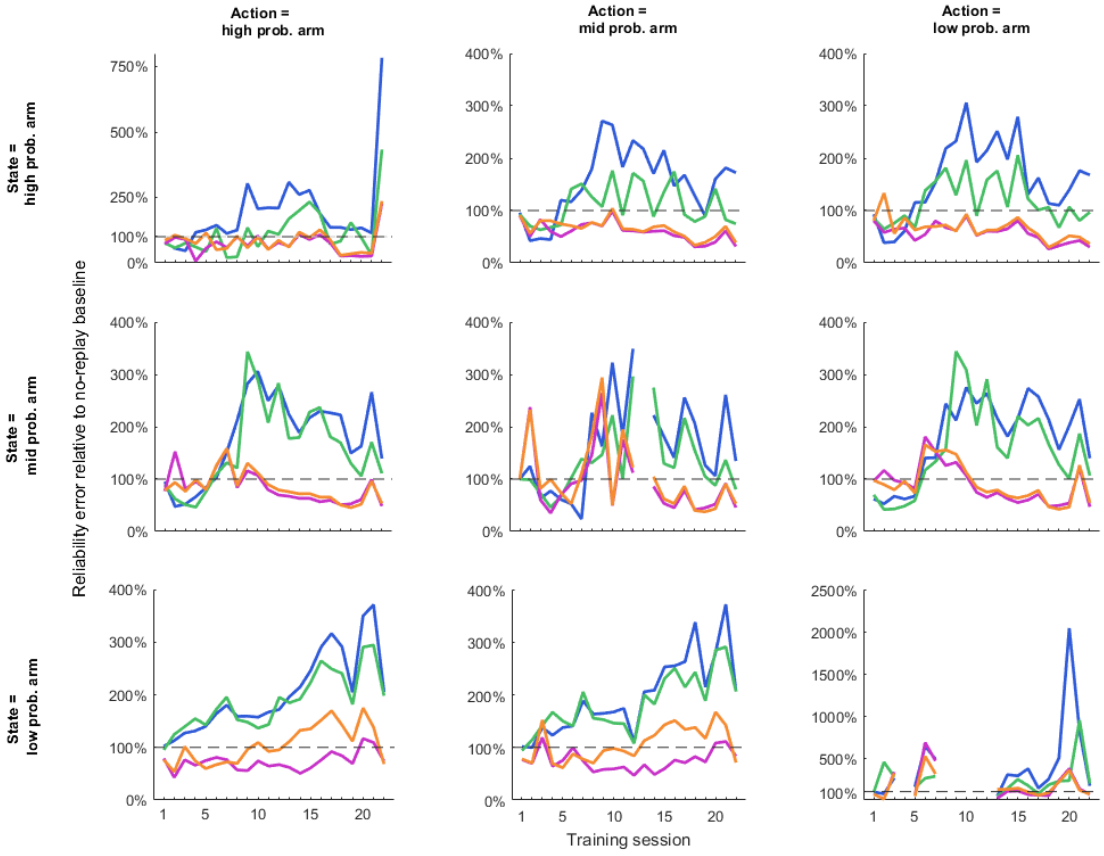


Figure 6: Change in reliability error for all trials on which a given state-action pair was expressed, with 15 samples replayed, relative to no-replay baseline. Intersection of "State = high prob. arm" and "Action = mid prob. arm" indicates a transition from high-probability arm to mid-probability arm.

3 Discussion

We trained rats on a reinforcement learning task designed to dissociate reward outcome (presence or absence of reward) from reward prediction error (RPE; an unexpected reward or absence of reward) on each trial. We trained variations of a Q-learning reinforcement learning model to predict behaviour on the task, and found that Q-learning with replay prioritised by RPE was the best predictor of learning.

Our first main result was that Q-learning can suitably model rats' learning of the stochastic reinforcement learning task, producing low reliability-errors when trained on rats' behaviour and predicting the likelihood of actions on each trial. This is consistent with other studies showing that Q-learning can predict behaviour in a range of tasks in rodents, monkeys and humans (Ito and Doya 2009). Given this result, we then proposed that adding replay to the Q-learning model between sessions might better reflect learning and therefore better predict behaviour. However, under a policy of replaying state-action pairs randomly, this produced higher reliability errors overall, indicating a worse model of the cognitive processes underlying reinforcement learning. Similarly, biasing replay by sampling from state-action pairs which had produced the largest recent reward also increased reliability errors relative to no-replay.

In contrast, biasing replay by sampling from state-action pairs which had produced the largest recent RPE decreased reliability errors. From this we conclude that the cognitive processes involved in the learning of this task are influenced by offline activity that takes place between sessions. Performance on memory tasks has widely been found to improve following a period of sleep (Stickgold 2005; Marshall and Born 2007; Diekelmann and Born 2010), associated with replay of activity which encodes recent experiences during hippocampal sharp-wave ripples (Ólafsdóttir et al. 2018). We therefore propose that such offline replay underlies the RPE-biased offline updating of state-action values which influenced reinforcement learning in this task.

The suggestion that hippocampal replay might be biased by RPEs differs from the commonly held view that replay is biased by reward itself (Ambrose et al. 2016; Atherton et al. 2015; Gruber et al. 2016; Singer and Frank 2009). However, the studies on which this conclusion is based generally do not use tasks which explicitly dissociate reward from RPE, so these results in the literature are not inconsistent with our suggestion that RPE biases replay.

Our conclusion that RPE-biased replay (but not random or reward-biased replay) improved model predictions is strengthened by the fact that this result did not hold when training data was shuffled. When the trial order was shuffled, such that there was no correlation between learning and behaviour, all replay policies produced higher reliability errors in predicting the animals' behaviour. This means that the influence of RPE is a feature of the learning process and not an epiphenomenon resulting from the general statistics of behaviour. Moreover, the result did hold for all state-action pairs, despite the overrepresentation in training data of those most frequently experienced. This gives credence to the notion that the Q-learning model with replay biased by RPE is a good overall model of state-action values held by the brain.

Despite the prevalence of the idea that reward biases replay, our alternative theory that RPE biases replay fits better with existing research on the role of dopamine. Dopaminergic projections from the ventral tegmental area (VTA) to CA1 in the hippocampus have been found to modulate both replay during sleep following exposure to

a novel environment, and subsequent memory performance in the same environment (McNamara et al. 2014). It is suggested that dopaminergic neuromodulation might tag synapses by upregulating plasticity-related proteins, causing long-lasting potentiation which allows the stabilisation of the memory trace during subsequent sleep and rest (Frey and Morris 1998; Redondo and Morris 2011). Phasic dopaminergic inputs to the hippocampus are triggered not only in response to novelty, but also in the context of reward (Schultz et al. 1997), offering a likely mechanism by which reward-related information might influence replay. Indeed, post-task replay has been found in reward-related VTA cells (Gomperts et al. 2015; Valdés et al. 2015). However, such phasic dopamine activations are typically elicited in response to anticipation of reward and RPEs rather than reward itself (D'Ardenne et al. 2008; Dayan and Niv 2008; Montague et al. 1996; Schultz 1998; Schultz et al. 1997). These phasic dopamine signals could therefore bias hippocampal replay towards activity associated with RPEs; it is less clear how activity associated with reward per se might bias replay.

Several studies have expressly linked replay to reward, ostensibly in contrast with our results, but often RPE is a confounding factor in these which cannot be discounted. In humans, high monetary reward (but not low monetary reward) is linked to sleep-dependent improvements in associative memory (Igloi et al. 2015; Studte et al. 2017); in this task RPE was not estimated but would presumably be higher overall in the high-reward than low-reward condition, conflating reward-dependent effects with RPE-dependent effects. In rodents, newly-rewarded behaviour has been associated with replay more than behaviour which had been rewarded in previous sessions (Singer and Frank 2009); here, the authors attributed this replay bias to novelty, but it is also consistent with increased RPE when new behaviours are rewarded for the first time. Moreover, following extended reinforcement of both behaviours, the replay bias for the newly-rewarded behaviour was eliminated. In a third study, results were more mixed: following an increase in reward magnitude at one end of a linear track, there was more replay associated with the larger-magnitude end than the unchanged-magnitude end, correlated with both reward and RPE (Ambrose et al. 2016). However, following an elimination of reward at one end, there was a reduction in replay following a reduction in reward despite the increase in RPE. This is more consistent with reward-biased than RPE-biased replay, although the authors noted a rebound effect when the eliminated reward was reinstated: greater replay was found at the reinstated-reward end than the unchanged-reward end, despite identical reward magnitudes. This leaves open the possibility of bias by positive over negative RPEs. A fourth study found more replay of large-reward-related activity than small-reward-related activity on a maze task (Michon et al. 2019), but because reward was received on every trial analysed, any effects of reward magnitude are conflated with positive reward-prediction error.

Conversely, the specific case for RPE-biased replay is supported by findings that neural sensitivity to RPEs in humans predicts the amount of awake replay during a reinforcement learning task, and replay amount correlated with subsequent performance in a task requiring behavioural flexibility (Momennejad et al. 2018).

In addition to human and rodent studies, findings from the literature on machine learning show some consistency with our results. A number of machine learning studies have found that storing new information in memory buffers and sampling from it at regular intervals, similar to hippocampal replay, can speed up learning (Lin 1992; Mnih et al. 2013, Mnih et al. 2015), and more so when replay is biased by prediction errors (Cichosz 1999; Schaul et al. 2016). RPE-biased replay may therefore represent an adaptive focus whereby resources are focused on areas of a cognitive model which needs updating.

We do not claim that this tells the whole story: RPE is almost certainly not the only factor that biases replay

and the phenomenon is likely to be much more multifaceted than this model suggests. First, phasic dopamine signalling to hippocampus may encode other kinds of prediction errors or aspects of reward to which the VTA is sensitive (Keiflin et al. 2019; Sharpe et al. 2019; Takahashi et al. 2017), and bias replay by the same mechanism. Reward itself may bias replay, especially if positive RPEs influence replay more than negative RPEs; there is also evidence that novelty (Hirase et al. 2001; Kudrimoti et al. 1999), the expectation of reward (Gruber et al. 2016), frequency of experience (Gupta et al. 2010) and strength of encoding (Schapiro et al. 2018) bias replay too. Furthermore, in addition to aiding reinforcement learning, replay has been associated with other memory-related functions including planning (Ólafsdóttir et al. 2017; Pfeiffer and Foster 2013), processing of emotional memories (Genzel et al. 2015), creative problem-solving (Lewis et al. 2018), and generalising from episodic memories to abstractions (Lewis and Durrant 2011), all of which are likely to necessitate some biasing of replay distinct from RPEs. In sum, we submit that hippocampal replay is more complex than the model outlined here.

Our model assumes that a cache of all experience is stored from which to be sampled, which is expensive and unrealistic at large scales. This may not be necessary if memory for individual trials is gradually forgotten and subsumed into cortical long-term memory, for example over the course of hours over which cell assembly activation decays (Giri et al. 2019).

Finally, this model leaves open some questions. It will be necessary to directly test this theory by recording neural data from which replay can be directly observed, comparing replay of reward-associated activity with that of RPE-related activity in the VTA or striatum. There is also an open question about possible diverging roles of replay during behaviour compared to prolonged rest and sleep. Here we have considered replay between sessions, which is likely to take place at least partly during sleep; but replay during wake has also been shown to be necessary for learning (Jadhav et al. 2012).

In summary, we found that a Q-learning-based reinforcement learning model which assumes offline updates between sessions is a better predictor of learning behaviour than one which does not assume offline updates. Specifically, this is true when updates are prioritised according to experiences that have recently elicited high RPEs, and not when they are prioritised according to reward or random recent experiences. This finding offers a reinterpretation of how offline activity during rest and sleep might aid reinforcement learning, in terms of RPE rather than reward.

4 Materials and Methods

Behavioural task

All procedures were performed in accordance with the United Kingdom Animals (Scientific Procedures) Act 1986 and European Union Directive 2010/63/EU and were reviewed by the University of Bristol Animal Welfare and Ethical Review Board.

Six adult male Lister hooded rats (weighing 260-330g, Charles River Laboratories, UK) were individually housed with environmental enrichment, and food-restricted to no less than 85% of their pre-restriction body weight. They were trained during the light part of a 12:12 light/dark cycle to forage on a 3-armed radial maze for liquid sucrose rewards in a dimly-lit room. The maze consisted of a raised central platform 25cm in diameter, with three arms (60cm x 7cm) protruding from it (fig. 1a). Arms were separated from the central platform by inverted-guillotine doors, which raised to block access to the arms, and fell below the maze floor to allow access. Turning zones (10 x 10cm) with lick ports were positioned at the end of each arm, at which 20% sucrose solution rewards were delivered. Door movements and reward delivery were operated automatically according to the animal's position, tracked using a webcam mounted above the maze, using custom MATLAB (The MathWorks) code. Following at least three days of habituation to the recording room and maze-operation sounds, each animal performed 22 training sessions, between 5 and 7 days per week, lasting 1 hour each.

Trials began when a rat entered, or was placed by the experimenter on, the central platform with all doors closed. Doors opened following a 5-second delay period. When the animal reached the lick port, reward was probabilistically delivered or withheld, and doors to the other two arms were closed; the third door was closed when the animal re-entered the central platform to begin a new trial.

Each arm was assigned as either "high probability", "mid probability" or "low probability", which determined the protocol for reward delivery. These assignments remained fixed throughout training for each animal, but were counter-balanced between animals. For the first 15 training sessions, the high-probability arm delivered a reward on 6 out of 8 (75%) legitimate entries to the arm, the mid-probability arm on 4 out of 8 (50%), and the low-probability arm on 2 out of 8 (25%). A legitimate entry was one in which a different arm had been entered on the previous trial; entering the same arm twice in a row was incorrect and did not result in a reward delivery. For sessions 16-20, the reward probabilities for the high- and low-probability arms were amplified: reward was delivered on 7 out of 8 (87.5%) and 1 out of 8 (12.5%) legitimate entries respectively. For sessions 21-22 the reward probabilities for the high- and low-probability arms were switched, such that the (formerly) high- and low- probability arms delivered reward on 1 out of 8 (12.5%) and 7 out of 8 (87.5%) of legitimate entries respectively.

Q-learning

We trained several variations of a Q-learning algorithm on the behavioural data to predict choices of which arm would be entered on each trial. Q-learning is a reinforcement learning algorithm developed for Markov decision processes in which an agent selects actions in its environment and observes the outcome, recording at each time step t its starting state s_t, selected action a_t, resulting reward r_t, and resulting state s_{t+1}. The agent builds up a matrix Q of Q-value estimates for every state-action pair:

$$\begin{bmatrix} Q_{s_1,a_1} & Q_{s_1,a_2} & \cdots & Q_{s_1,a_A} \\ Q_{s_2,a_1} & Q_{s_2,a_2} & \cdots & Q_{s_2,a_A} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s_S,a_1} & Q_{s_S,a_2} & \cdots & Q_{s_S,a_A} \end{bmatrix} \tag{3}$$

corresponding to the future discounted expected reward, i.e. the temporal difference between the current state and the reward state. These Q-value estimates are used to guide actions to maximise reward. At each time step t, the Q-value for the state-action pair observed is updated by:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot (r_t + \gamma \cdot \max Q(s_{t+1}, a)) \tag{4}$$

where $\alpha \in (0, 1)$ is a learning rate parameter which determines the degree to which new information overrides old information, and $\gamma \in (0, 1)$ is a discount parameter which determines the importance of long-term gains.

In this task, entries into a chosen arm (and arrival at the goal location at the end of the arm) were modelled as actions, while the arm entered on the previous trial, on which reward probabilities were contingent, were modelled as states. Each trial therefore gave rise to one state-action transition out of nine possible state-action pairs.

Q-learning with replay

We used four variants of Q-learning in which additional "offline" updates are performed between "online" trials, based on sequences already experienced, to boost learning. This has the effect of learning from several trials per actual trial of experience, and is similar to the Dyna-Q algorithm which has been shown to speed up learning compared to Q-learning alone (Sutton 2014) in a manner which may underlie the function of hippocampal replay (Johnson and Redish 2005). Generally, sequences are selected randomly from a memory buffer of recently-acquired experiences, without bias towards any trial or type of trial. Given the observed bias reported in the literature towards salient experiences, such as those rewarded or aversive, we modified Dyna-Q to perform updates only between sessions and to reflect hypothesised biases in four different ways.

Parameter-fitting

Parameter-fitting for Q-learning

First, a Q-learning algorithm (without replay) was trained, to obtain a baseline score against which various replay policies could be compared. Q-values were stored for each state-action pair on the task, and updated according to each animal's experience. A state s_t was defined as the arm visited on the previous trial $t-1$, and an action a_t was defined as the arm chosen on the current trial t. Following each trial of an animal's training, the Q-value $Q(s_t, a_t)$ was updated according to the reward received, $r \in \{0, 1\}$ by equation 4, and Q-values were transformed into a forecast probability of choosing each arm on the subsequent trial.

The learning rate α, discount factor γ, and exploration factor ϵ were free parameters that were tuned to each rat, using the following optimisation procedure. Here we used a reliability score (Murphy and Murphy 1973), generated based on the forecast probabilities of all trials, to quantify the consistency of the forecast probabilities with the animals' behaviour. The mean observed frequency was calculated for each state-action pair, i.e. the proportion of trials on which a given action was chosen in a given state, and the reliability score R_t for a given trial t was calculated according to:

$$R_t = n_{s_t} \cdot \sum_{a=1}^{n_a} (p_a - o_{s_t,a})^2 \tag{5}$$

where s_t is the animal's state on trial t, n_{s_t} is the number of trials on which the animal was in state s_t, n_a is the number of possible actions (3) p_a is the forecast probability for entering arm a, and $o_{s,a}$ is the mean observed frequency of state-action pair s, a.

Parameter optimisation was performed using the reliability error as the cost function. Because the parameter state-space was vulnerable to local minima, and also because it was highly stochastic under replay policies (see description below), a two-step approach was taken to optimise parameters. In the first step, simulated annealing was run 32 times for a maximum of 1000 iterations (or until the reliability error could not be improved by more than $1e^{-6}$), using the MATLAB function `simulannealbnd`. Reliability error was averaged over 1,000 runs when computing the cost function, to minimise stochasticity. This function performs a probabilistic variation of gradient descent by taking increasingly smaller steps in random directions, to approximate a global minimum without becoming stuck in local minima. The resulting 32 rough estimates of the optimal parameter values were used as the initial values for the second step: a simple quasi-Newton method using gradient descent, implemented by the MATLAB optimisation function `fmincon`, for a further maximum 1000 iterations. Of the 32 final sets of parameter values, the one which produced the smallest reliability error was used for analysis. All analyses were performed on the average reliability error over 1,000 runs using the given parameter values.

Parameter-fitting for Q-learning with replay

Against the baseline of no-replay, the same optimisation procedure was performed with increasing amounts of replay according to four replay policies. Following each session, a specified number of samples were chosen from all the trials experienced so far. How the samples were selected depended on the replay policy (detailed below); a probability $P(s, a)$ was assigned to each state-action pair to determine which pair to sample from. From the chosen state-action pair, a sample trial was chosen according to the probability $P(i)$ in which a recency parameter ensured that more recent trials were exponentially more likely to be chosen. Q-values were then updated according to the state, action and reward of the sampled trial, in the same manner as "online" Q-value updates described in equation 4.

Each replay policy required the same three parameters to be optimised as in Q-learning without replay, plus additional parameters for recency and/or RPE-weighting. Table 2 shows the number of free parameters for each replay policy.

Replay policy	Number of parameters
No replay	3
Random replay	4
Reward-biased replay	4
RPE-prioritised replay	5
RPE-proportional replay	5

Table 2

These were optimised according to the same procedure as for Q-learning with no replay, described above, for $n = \{1, 3, 5, 10, 15, 20, 30, 40, 50, 75, 100\}$ replay events between each session, resulting in 11 sets of parameter values for each replay policy and each animal. Comparing this to plausible quantities of replay events in animals is not trivial, but studies in which discrete replay events are enumerated report 100-200 bursts of hippocampal activity that can be statistically related to prior experience, over the first one or two hours after experience (Ólafsdóttir et al. 2016; Michon et al. 2019). Separately, reactivation of cell pairs has been found to decay to baseline well within that time period following exposure to familiar environments (Giri et al. 2019), so the first one to two hours is likely to be when most replay of recent experience in a familiar environment occurs.

Random replay

Random replay, biased by nothing but the recency of an action, was included as a control. For each replay event, a state-action pair was chosen at random out of all state-action pairs experienced so far:

$$P(s, a) = \frac{1}{n_{sa}} \tag{6}$$

where n_{sa} is the number of state-action pairs experienced (up to 9). The subset of trials experienced, $i \in (1, I)$, which represented this state-action pair were ordered chronologically, and the probability $P(i)$ of a trial i being replayed was determined according to a recency parameter φ:

$$P(i) = \frac{i^{\varphi}}{\sum_{i=1}^{I} p_i} \tag{7}$$

480 **Reward-biased replay**

481 Reward-biased replay represents the predominant interpretation of how reward influences replay (Atherton et al.
482 2015, Carr et al. 2011). For each replay event, a state-action pair s, a was chosen probabilistically in proportion
483 to its Q-value:

$$P(s, a) = \frac{Q(s, a)}{\sum_{s=1}^{n_s} \sum_{a=1}^{n_a} Q(s, a)} \tag{8}$$

484 The subset of trials experienced which represented the chosen state-action pair were ordered chronologically,
485 and determined according to equation 7.

486 **RPE-prioritised replay**

487 RPE-prioritised replay represents the policy of replaying trials associated with the most surprising outcomes,
488 i.e. where the difference between expectation (Q-values) and experience (reward) was greatest. For each trial
489 t, RPE was calculated as the difference between actual reward and expected reward:

$$\text{rpe}_t = r + \gamma \cdot Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{9}$$

490 For every trial $i \in (1, I)$ which was an example of a given state-action pair, its absolute value was weighted,
491 determined by a parameter φ raised to the power of its recency i:

$$\text{wrpe}_i = \mid rpe_i \mid \cdot \varphi^i \tag{10}$$

492 The weighted RPEs, wpre, were then averaged to produce an overall weighted-average RPE, $\text{RPE}_{s,a}$, for each
493 state-action pair s, a, which was more heavily influenced by recent trials:

$$\text{RPE}_{s,a} = \frac{\sum_{i=1}^{I} \text{wrpe}_i}{I} \tag{11}$$

494 The state-action pair with the highest RPE was selected, and the subset of trials experienced which represented
495 the chosen pair were ordered chronologically, and determined according to equation 7. Once replayed, the rpe_t
496 for the trial sampled was updated to reflect the RPE resulting from the replay event.

497 **RPE-proportional replay**

RPE-proportional replay is a variant of RPE-prioritised replay, in which state-action pairs are chosen in proportion to their weighted-average-RPE instead of choosing the pair with the highest weighted-average-RPE. The RPE was calculated according to eq. 11 and a state-action pair to be sampled from was chosen probabilistically according to:

$$p_{s,a} = \frac{\text{RPE}_{s,a}}{\sum \text{RPE}_{s,a}} \tag{12}$$

The subset of trials experienced which represented the chosen state-action pair were ordered chronologically, and determined according to equation 7. Once replayed, the rpe_t for the trial sampled was updated to reflect the RPE resulting from the replay event.

Shuffling procedure

As an additional control, the parameters were also optimised for shuffled data, in which trial order was randomly permuted 1,000-fold. This preserved the large-scale information in the training data, such as the mean observed frequency and average rewards of state-action pairs and the number of trials in each session between replays, but disrupted the specific structure of how this information was acquired over time.

Code Availability

All data and code used in this study are available at https://github.com/eroscow/QlearningReplay.

References

Ambrose, R. E., Pfeiffer, B. E., Correspondence, D. J. F. & Foster, D. J. (2016). Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. *Neuron*, *91*. doi:10.1016/j.neuron.2016.07.047

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., ... Zaremba, W. (2017). Hindsight Experience Replay. Retrieved from http://papers.nips.cc/paper/7090-hindsight-experience-replay

Antony, J. W., Gobel, E. W., O'Hare, J. K., Reber, P. J. & Paller, K. A. (2012). Cued memory reactivation during sleep influences skill learning. *Nat. Neurosci. 15*(8), 1114–6. doi:10.1038/nn.3152

Atherton, L. A., Dupret, D. & Mellor, J. R. (2015). Memory trace replay: the shaping of memory consolidation by neuromodulation. *Trends Neurosci. 38*(9), 560–70. doi:10.1016/j.tins.2015.07.004

Bendor, D. & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nat. Neurosci. 15*(10), 1439–44. doi:10.1038/nn.3203

Cairney, S. A., Durrant, S. J., Jackson, R. & Lewis, P. A. (2014). Sleep spindles provide indirect support to the consolidation of emotional encoding contexts. *Neuropsychologia*, *63*, 285–292. doi:10.1016/j.neuropsychologia.2014.09.016

Calabresi, P., Picconi, B., Tozzi, A. & Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci. 30*(5), 211–219. doi:10.1016/J.TINS.2007.03.001

Carr, M. F., Jadhav, S. P. & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci. 14*(2), 147–53. doi:10.1038/nn.2732

Cheng, S. & Frank, L. M. (2008). New experiences enhance coordinated neural activity in the hippocampus. *Neuron*, *57*(2), 303–13. doi:10.1016/j.neuron.2007.11.035

Cichosz, P. (1999). An analysis of experience replay in temporal difference learning. *Cybern. Syst. 30*(5), 341–363. doi:10.1080/019697299125127

D'Ardenne, K., McClure, S. M., Nystrom, L. E. & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–7. doi:10.1126/science.1150605

Daw, N. D., Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci. 8*(12), 1704–1711. doi:10.1038/nn1560

Day, J. J., Roitman, M. F., Wightman, R. M. & Carelli, R. M. (2014). Associative learning mediates dynamic shifts in DORamine signaling in the nucleus accumbens Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. doi:10.1038/nn1923

Dayan, P. & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Curr. Opin. Neurobiol. 18*(2), 185–196. doi:10.1016/J.CONB.2008.08.003

Diekelmann, S. & Born, J. (2010). The memory function of sleep. *Nat. Rev. Neurosci. 11*(2), 114–126. doi:10.1038/nrn2762

Dupret, D., O'Neill, J. & Pleydell-Bouverie, B. (2010). The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nature*. Retrieved from http://www.nature.com/neuro/journal/v13/n8/abs/nn.2599.html

Ego-Stengel, V. & Wilson, M. A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, *20*(1), 1–10. doi:10.1002/hipo.20707

Foster, D. J. & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683. doi:10.1038/nature04587. arXiv: 440: 680âĂŞ683

Frey, U. & Morris, R. G. (1998). Synaptic tagging: implications for late maintenance of hippocampal long-term potentiation. *Trends Neurosci. 21*(5), 181–188. doi:10.1016/S0166-2236(97)01189-2

Genzel, L., Spoormaker, V., Konrad, B. & Dresler, M. (2015). The role of rapid eye movement sleep for amygdala-related memory processing. *Neurobiol. Learn. Mem. 122*, 110–121. doi:10.1016/J.NLM.2015.01.008

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci. 12*(10), 1222–1223. doi:10.1038/nn.2384

Girardeau, G., Inema, I. & Buzsáki, G. (2017). Reactivations of emotional memory in the hippocampus–amygdala system during sleep. *Nat. Neurosci. 20*(11), 1634–1642. doi:10.1038/nn.4637

Giri, B., Miyawaki, H., Mizuseki, K., Cheng, S. & Diba, K. (2019). Hippocampal Reactivation Extends for Several Hours Following Novel Experience. *J. Neurosci. 39*(5), 866–875. doi:10.1523/JNEUROSCI.1950-18.2018

Gomperts, S. N., Kloosterman, F., Wilson, M. A., Cardinal, R., Parkinson, J., Hall, J., ... Sejnowski, T. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. *Elife*, *4*, 321–352. doi:10.7554/eLife.05360

Gruber, M. J., Ritchey, M., Wang, S.-F., Doss, M. K. & Ranganath, C. (2016). Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events. *Neuron*, *89*(5), 1110–1120. doi:10.1016/j.neuron.2016.01.017

Gupta, A. S., van der Meer, M. A., Touretzky, D. S. & Redish, A. D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. *Neuron*, *65*(5), 695–705. doi:10.1016/J.NEURON.2010.01.034

Hirase, H., Leinekugel, X., Czurkó, A., Csicsvari, J. & Buzsáki, G. (2001). Firing rates of hippocampal neurons are preserved during subsequent sleep episodes and modified by novel awake experience. *Proc. Natl. Acad. Sci. U. S. A. 98*(16), 9386–90. doi:10.1073/pnas.161274398

Igloi, K., Gaggioni, G., Sterpenich, V. & Schwartz, S. (2015). A nap to recap or how reward regulates hippocampal-prefrontal memory networks during daytime sleep in humans. doi:10.7554/eLife.07903.001

Ito, M. & Doya, K. (2009). Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia. *J. Neurosci. 29*(31), 9861–9874. doi:10.1523/jneurosci.6157-08.2009

Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. (2012). Awake Hippocampal Sharp-Wave Ripples Support Spatial Memory. *Science (80-.). 336*(6087), 1454–1458. doi:10.1126/SCIENCE.1217230

Johnson, A. & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, *18*(9), 1163–1171. doi:10.1016/J.NEUNET.2005.08.009

Karimpanal, T. G. & Bouffanais, R. (2017). Experience Replay Using Transition Sequences. *Front. Neurorobot. 12*, 32. doi:10.3389/fnbot.2018.00032. arXiv: 1705.10834

Keiflin, R., Pribut, H. J., Shah, N. B. & Janak, P. H. (2019). Ventral Tegmental Dopamine Neurons Participate in Reward Identity Predictions. *Curr. Biol. 29*(1), 93–103.e3. doi:10.1016/J.CUB.2018.11.050

Kim, H., Lee, D. & Jung, M. W. (2013). Signals for Previous Goal Choice Persist in the Dorsomedial, but Not Dorsolateral Striatum of Rats. *J. Neurosci. 33*(1), 52–63. doi:10.1523/jneurosci.2422-12.2013

Kudrimoti, H. S., Barnes, C. A. & McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. *J. Neurosci. 19*(10), 4090–101. doi:10.1523/JNEUROSCI.19-10-04090.1999

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. A. (2009). Hippocampus Leads Ventral Striatum in Replay of Place-Reward Information. *PLoS Biol. 7*(8), e1000173. doi:10.1371/journal.pbio.1000173

Lewis, P. A. & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends Cogn. Sci. 15*(8), 343–351. doi:10.1016/j.tics.2011.06.004

Lewis, P. A., Knoblich, G. & Poe, G. (2018). How Memory Replay in Sleep Boosts Creative Problem-Solving. *Trends Cogn. Sci. 22*(6), 491–503. doi:10.1016/J.TICS.2018.03.009

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn. 8*(3-4), 293–321. doi:10.1007/BF00992699

Marshall, L. & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends Cogn. Sci. 11*(10), 442–450. doi:10.1016/J.TICS.2007.09.001

McClure, S. M., Berns, G. S. & Montague, P. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron*, *38*(2), 339–346. doi:10.1016/S0896-6273(03)00154-5

McNamara, C. G., Tejero-Cantero, Á., Trouche, S., Campo-Urriza, N. & Dupret, D. (2014). Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nat. Neurosci. 17*(12), 1658–1660. doi:10.1038/nn.3843

Michon, F., Sun, J.-J., Kim, C. Y., Ciliberti, D. & Kloosterman, F. (2019). Post-learning Hippocampal Replay Selectively Reinforces Spatial Memory for Highly Rewarded Locations. *Curr. Biol. 29*(9), 1436–1444.e5. doi:10.1016/J.CUB.2019.03.048

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. arXiv: 1312.5602. Retrieved from http://arxiv.org/abs/1312.5602

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. doi:10.1038/nature14236

Momennejad, I., Otto, R., Daw, N. D. & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. doi:10.7554/eLife.32548.001

Montague, P. R., Dayan, P. & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci. 16*(5), 1936–47. doi:10.1523/JNEUROSCI.16-05-01936.1996

Morris, G., Schmidt, R. & Bergman, H. (2010). Striatal action-learning based on dopamine concentration. *Exp. Brain Res. 200*(3-4), 307–317. doi:10.1007/s00221-009-2060-6

Murphy, A. H. & Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *J. Appl. Meteorol. 12*(4), 595–600. doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. doi:10.1016/S0896-6273(03)00169-7

Ólafsdóttir, H. F., Bush, D. & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Curr. Biol. 28*(1), R37–R50. doi:10.1016/j.cub.2017.10.073

Ólafsdóttir, H. F., Carpenter, F. & Barry, C. (2016). Coordinated grid and place cell replay during rest. *Nat. Neurosci. 19*(6), 792–794. doi:10.1038/nn.4291

Ólafsdóttir, H. F., Carpenter, F. & Barry, C. (2017). Task Demands Predict a Dynamic Switch in the Content of Awake Hippocampal Replay. *Neuron*, *96*(4), 925–935.e6. doi:10.1016/J.NEURON.2017.09.035

Pagnoni, G., Zink, C. F., Montague, P. R. & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nat. Neurosci. 5*(2), 97–98. doi:10.1038/nn802

Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74–9. doi:10.1038/nature12112

Rasch, B., Büchel, C., Gais, S. & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science (80-.).* Retrieved from http://science.sciencemag.org/content/315/5817/1426.short

Redondo, R. L. & Morris, R. G. M. (2011). Making memories last: the synaptic tagging and capture hypothesis. *Nat. Rev. Neurosci. 12*(1), 17–30. doi:10.1038/nrn2963

Roesch, M. R., Calu, D. J. & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci. 10*(12), 1615–1624. doi:10.1038/nn2013

Rudoy, J., Voss, J., Westerberg, C. & Paller, K. (2009). Strengthening individual memories by reactivating them during sleep. *Science (80-.).* Retrieved from http://science.sciencemag.org/content/326/5956/1079.short

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C. & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat. Commun. 9*(1), 3920. doi:10.1038/s41467-018-06213-1

Schaul, T., Quan, J., Antonoglou, I., Silver, D. & Deepmind, G. (2016). *Prioritized Experience Replay*. arXiv: 1511.05952v4. Retrieved from https://arxiv.org/pdf/1511.05952.pdf

Schultz, W., Dayan, P. & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9. doi:10.1126/SCIENCE.275.5306.1593

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *J. Neurophysiol. 80*(1), 1–27. doi:10.1152/jn.1998.80.1.1

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues Clin. Neurosci. 18*(1), 23–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27069377%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4826767

Sharpe, M. J., Batchelor, H. M., Mueller, L. E., Chang, C. Y., Maes, E. J., Niv, Y. & Schoenbaum, G. (2019). Dopamine transients delivered in learning contexts do not act as model-free prediction errors. *bioRxiv*, 574541. doi:10.1101/574541

Singer, A. C. & Frank, L. M. (2009). Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus. *Neuron*, *64*(6), 910–921. doi:10.1016/j.neuron.2009.11.016

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*(7063), 1272–1278. doi:10.1038/nature04286

Studte, S., Bridger, E. & Mecklinger, A. (2017). Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain Lang. 167*, 28–35. doi:10.1016/J.BANDL.2016.03.003

Sutton, R. S. (2014). Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. *Mach. Learn. Proc. 1990*, 216–224. doi:10.1016/b978-1-55860-141-3.50030-4

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M. & Schoenbaum, G. (2017). Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards Article Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron*, *95*. doi:10.1016/j.neuron.2017.08.025

Valdés, J. L., McNaughton, B. L. & Fellous, J.-M. (2015). Offline reactivation of experience-dependent neuronal firing patterns in the rat ventral tegmental area. *J. Neurophysiol. 114*(2), 1183–95. doi:10.1152/jn.00758.2014

Watkins, C. J. (1989). Learning form delayed rewards. *Ph. D. thesis, King's Coll. Univ. Cambridge*. Retrieved from https://ci.nii.ac.jp/naid/10007782517/

Wimmer, G. E., Li, J. K., Gorgolewski, K. J. & Poldrack, R. A. (2018). Reward Learning over Weeks Versus Minutes Increases the Neural Representation of Value in the Human Brain. *J. Neurosci. 38*(35), 7649–7666. doi:10.1523/jneurosci.0075-18.2018

Wu, C.-T., Haggerty, D., Kemere, C. & Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nat. Neurosci. 20*(4), 571–580. doi:10.1038/nn.4507

Yu, J. Y., Kay, K., Liu, D. F., Grossrubatscher, I., Loback, A., Sosa, M., ... Frank, L. M. (2017). Distinct hippocampal-cortical memory representations for experiences associated with movement versus immobility. *Elife*, *6*. doi:10.7554/eLife.27621

Acknowledgements: We are grateful to Rui Ponte Costa and Mark Humphries for useful discussions and comments, and to Aleksander Domanski and Andrew New for assistance with experimental set-up. This research was funded by a Wellcome Trust PhD scholarship.

Author Contributions: E.L.R., M.W.J. and N.F.L. conceived and designed the study. E.L.R. carried out the experiments and analysed the data. E.L.R. performed the computational modelling under the guidance of N.F.L. E.L.R. and N.F.L. prepared the paper with critical revision from M.W.J.

Competing interests: The authors declare no competing interests.