Differential replay of reward and punishment paths predicts approach and avoidance

Jessica McFadyen[*1,2], Yunzhe Liu[3,4], & Raymond J Dolan[1,2]

* corresponding author: drjessicajean@gmail.com

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

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

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

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

Abstract

Neural replay is implicated in planning, where states relevant to a task goal are rapidly reactivated in sequence. It remains unclear whether, during planning, replay relates to an actual prospective choice. Here, using magnetoencephalography (MEG), we studied replay in human participants while they planned to either approach or avoid an uncertain environment containing paths leading to reward or punishment. We find evidence for forward sequential replay during planning, with rapid state-to-state transitions from 20 to 90 ms. Replay of rewarding paths was boosted, relative to aversive paths, prior to a decision to avoid and attenuated prior to a decision to approach. A trial-by-trial bias towards replaying prospective punishing paths predicted irrational decisions to approach riskier environments, an effect more pronounced in participants with higher trait anxiety. The findings indicate a coupling of replay with planned behaviour, where replay prioritises an online representation of a worst-case scenario for approaching or avoiding.

Introduction

When formulating a plan, we often face uncertainty as to whether a choice will lead to a good or bad outcome. For example, when we deliberate whether to go to a party or stay home, we might simulate potential sequences of events that are positive (e.g., arriving and seeing friends, meeting new people, coming home feeling happy) or negative (e.g., arriving and not knowing anybody, saying something embarrassing in front of new people, leaving early, and regretting the whole experience). Situations such as these can engender approach-avoidance conflict, wherein decision-making is rendered difficult by a need to weigh the benefits of a risky choice against a more certain, but less rewarding, choice to avoid.

Neural replay, originally characterised in the context of a rapid sequential reactivation of hippocampal place cells that map specific locations of recently experienced paths[1–7], is linked to a number of functions in both humans and animals, including memory consolidation of spatial[4,8–11] and temporal order relationships[7,12,13], inference[14–16], and credit assignment[17]. There is also evidence indicating that neural replay may relate to a simulation of potential outcomes during active planning[18–22].

A role for prospective replay in planning is supported by observations that when rodents pause during spatial navigation, the order of replayed place cell firing matches paths leading to the learned location of a reward[23,24], and is enhanced for paths leading to greater rewards[25]. Furthermore, the more a rewarding path is prospectively replayed, the more likely it is that the animal will pursue that path[21,23,24,26]. A disruption of replay events at decision points, such as by application of electric pulses to the hippocampus, leads to the expression of more vicarious trial and error behaviour[25,27] and a greater likelihood that an animal will take an incorrect path[10,28]. Remarkably, replay events also provide a mapping of potential trajectories to rewards that have never been experienced, evident in both online[29] and offline[30,31] sequential reactivation.

In contrast to reward, the question of how prospective aversive events modulate replay is under-investigated. Animal studies show that removal of a reward leads to a marked reduction in replay[32]. Paths leading to danger, however, are also more strongly replayed, and this is anticorrelated with an animal's chosen trajectory such that they tend to avoid the dangerous path[33]. Such findings have led to a proposal that hippocampal replay prioritises paths that are most immediately relevant for on-going behaviour[34]. Recent evidence, however, suggests the goal of a current plan might not, in fact, directly relate to which path is most strongly replayed. Instead, the selection of paths for replay appears to relate to mnemonic functions that support future planning, evident in replay being enhanced for paths leading to previously-rewarded locations that have not been visited recently[35,36], as well as for paths associated with sub-optimal decision-making[22]. Within this formulation, replay is proposed to support planning by consolidating memories of sequences that are susceptible to being forgotten, rather than reflecting a simulation of states leading to outcomes that directly relate to a current motivational goal[37].

A feature of many previous studies of replay has been the use of environments that contain either reward or punishment. Little is known about how replay is impacted by a prospective environment where paths can lead to either reward or punishment, especially where these environments give rise to an approach-avoidance conflict[38,39]. Notably, an inability to make optimal decisions under approach-avoidance conflict is a characteristic of clinical anxiety disorders, where the potential for experiencing a negative event leads to avoidance regardless of the likelihood of potential reward[40–42]. On the other hand, a tendency to approach, even when this might have negative consequences, is considered a risk factor for developing substance abuse disorders[43]. During approach-avoidance conflict, the magnitude and likelihood of threat is proposed to be monitored by anterior and ventral hippocampus interactions that arbitrate decisions to approach or avoid, in both humans[39,44] and rodents[45,46].

Replay is a candidate mechanism for this process, where a relative increase in prospective replay strength of one trajectory over another might relate to a bias towards deciding to approach versus avoid.

Here, we employed magnetoencephalography (MEG) to investigate whether there is an asymmetry between the replay of rewarding and aversive path sequences during planning. We designed a gambling-style task in which participants made decisions to either approach or avoid an uncertain environment containing paths leading to either gain or loss. By decoding rapid sequential replay related to sequences of transitioned states, we reveal a striking replay asymmetry that reflects prospective evaluations during planning and predicts trial-by-trial decision-making.

Results

Expected value guides decision-making

Participants learnt the structure of an environment containing two sequences (hereafter referred to as "paths") containing three images (hereafter referred to as "states"), each with an associated integer value (**Extended Data Fig. 1C**). In a gambling-type scenario, where the overall task goal was to earn as many points as possible, participants could choose to either "approach" the environment, thereby probabilistically transitioning to one of the two paths, or "avoid" the environment entirely (receiving a guaranteed sum of 1 point).

To make a rational choice, participants needed to mentally simulate a prospective accumulation of points along each path. The total value of each path was dependent on a visual cue presented at the beginning of each trial, which also guided participants towards a sequential evaluation of each path in a forwards direction (**Fig. 1;** see description of "odd rule" in **Methods**). In a majority of trials, one of the two paths resulted in an overall gain and the other in an overall loss. The likelihood of transitioning to either of the two paths (conditional on participants choosing to approach) spanned five probabilities (10-90%, 30-70%, 50-50%, 70-30%, and 90-10%), and these were displayed on screen during an allowed 30-second planning period.

If participants chose to approach, a screen then displayed which of the two available paths had been selected, as determined by the path transition probabilities displayed during the planning phase (**Fig. 1A**). Participants then deterministically transitioned to each state along the selected path, with the state value and cumulative sum of points along the trajectory displayed on-screen. Note that the first four trials of each block were forced-choice to approach, serving as a reminder of the images representing each state (images were replaced by text labels in all other free-choice trials to control for visual exposure) and their associated integer values (the value of one state from each of the two paths was updated at the beginning of each block). If participants chose to avoid, a screen was then displayed indicating that participants had earned one point.

In the task, rational decision-making required calculating the expected value of approach (i.e., the sum of points for paths 1 and 2, weighted by their probabilities) and then choosing to approach only if the overall expected value is greater than a certain value granted after choosing to avoid (i.e., ≥ 1). We calculated the accuracy of participants' choices by comparing them to perfectly rational choice behaviour. Simulations of different behavioural strategies showed that learning path values from experience, as opposed to the more cognitively burdensome optimal strategy of sequentially summing state values, could achieve approximately 69% to 85% accuracy (see **Supplementary methods** for model simulations). Moreover, only considering one, but not both, paths when computing the expected value achieved an hypothetical mean accuracy of approximately 63% (range = 53% to 75%). Two of 26 participants performed at 47.55% and 51.37% accuracy, respectively, and thus were excluded from all subsequent analyses, except for evaluation of replay for an overall state map.

We expected an ordered reactivation of state transitions to reflect the repeated visual experience of paths in sequential order (i.e., during learning, as well as the walkthrough phases of decision trials), as shown by previous studies[17,47], as opposed to reflecting a conscious mental calculation performed during planning. The sequential nature of determining path value was therefore a design feature that served to encourage perception of temporal order in the relationships between states, as well as provide a sufficient level of task difficulty. Moreover, the chosen design aligned with previous work using paradigms that incorporate a cumulative sum calculation[12], as well as investigations of spatial replay (which is inherently sequential[1–7]).

Participants performed significantly above chance, with 76.07% accuracy on average (SD = 7.35%, range = 60.27% to 89.73%; $t(23) = 17.373$, $p < 0.001$; **Fig. 2A**), correctly approaching when the expected value was 2.386 on average (SD = 0.57) and avoiding when the expected value was -1.552 on average (SD = 0.556; $t(23) = 21.152$, $p < 0.001$; **Fig. 2B**). Overall, participants tended to approach more (57.15% of trials) than avoid (42.85%; $t(23) = 4.176$, $p < 0.001$), consistent with reward-seeking or information-seeking behaviour. Experimental protocols were designed so that the expected value of approaching was > 1 on 50% of trials (**Extended Data Fig.1E**). As such, accuracy was significantly lower on trials where participants chose to approach (74.59%, SD = 7.53%) than avoid (79.27%, SD = 8.07%; $t(23) = -3.190$, $p = 0.004$; **Fig. 2A**). Participants were also significantly faster in their decision to approach (M = 8.446 seconds, SD = 2.03) than to avoid (M = 8.975 seconds, SD = 2.424; $t(23) = -2.319$, $p = 0.030$; **Fig. 2C**).

In the experimental design, there was consistency as to which of the two paths culminated in a reward or loss. Hence, for the first half of the experiment, path 1 resulted in reward and path 2 resulted in loss, and vice versa for the second half of the experiment (two protocols were used and counterbalanced across participants; see **Methods**). To encourage active engagement in sequential planning, rather than merely learning this tendency, we included catch trials (5%) where both paths either led to a reward or to a loss. Behavioural modelling of different strategies was consistent with participants performing online calculations (winning mental arithmetic model: N = 15/24) as opposed to a strategy of caching learned values (N = 9/24; see **Supplementary methods**). Thus, for a

majority of participants (N = 20/24), choice behaviour was best explained by a model in which both paths were considered when computing expected value, while that of the remaining participants was best explained by models that either reflected a consideration of only path 1 (N = 2/24) or only the path perceived to be consistently punishing (N = 2/24). Thus, a majority of participants engaged in sequential planning by mentally accumulating points along each path, with the majority considering both paths (rather than just one path, in an effort to conserve cognitive resources) when deliberating. Note that, of the two participants excluded from path-specific analyses due to overall poor task accuracy, one was best explained by a null model (i.e., a general bias towards approaching, irrespective of expected outcome) and the other by a caching strategy that considered aversive paths alone.

We next constructed a multilevel logistic regression model to more precisely examine how path values and transition probabilities influenced trial-by-trial decision-making. In this model, trial-by-trial choice was predicted by a three-way interaction between the value of the path with the highest prospective value (i.e., the rewarding path), the value of the path with the lowest prospective value (i.e., the loss path), and the probability of transitioning to the rewarding path (the probability of transitioning to one path was always relative to the other). We also included response time (RT) as a fixed effect, as well as certainty of path transition probabilities on each trial (uncertain: 50-50%, moderately certain: 30-70% or 70-30%, very certain: 10-90% or 90-10%).

Approach choices were significantly predicted both by the probability of transitioning to a more rewarding path ($\beta = 6.460$, $p < 0.001$) and by larger prospective rewards ($\beta = 0.113$, $p < 0.001$; **Fig. 2D**). Thus, participants approached environments containing larger rewards more when the probability of transitioning to reward was higher ($\beta = 4.991$, $p < 0.001$). Although the magnitude of potential loss also predicted decision-making (participants were more likely to choose to avoid when potential losses were larger: $\beta = 0.053$, $p = 0.011$), there was no interaction with transition probability ($\beta = -0.028$, $p = 0.744$; **Fig. 2E**). These findings support the idea that decision-making was guided by the total value of reward and loss paths, as well as the probability of transitioning to a rewarding path. We also observed a significant effect of certainty ($\beta = 0.317$, $p < 0.001$), such that participants were more likely to approach when transition probabilities were more certain overall (i.e., 90-10% or 10-90%, as opposed to 50-50%).

Finally, given that participants were more likely to approach when rewarding paths were more probable, we also tested whether participants experienced rewarding paths more frequently than aversive paths. On average, participants transitioned to a rewarding path 107 times (SD = 11) and to an aversive path 23 times (SD = 8), a difference that was significant ($t(23) = 25.577$, $p < 0.001$). Importantly, due to our counterbalanced design there was no significant difference in the likelihood of experiencing path 1 (M = 46, SD = 8) or path 2 (M = 46, SD = 7; $t(23) = -0.046$, $p = 0.964$).

Forward replay during planning

Our primary research questions with regard to replay were: 1) whether there is a sequential reactivation of state-to-state path transitions during planning, 2) whether this is influenced by each path's perceived value, and 3) whether this, in turn, relates to a subsequent choice to approach or avoid. In an initial functional localiser task, we trained classifiers on visually-evoked response fields (measured using MEG) for six unique state images (see **Fig. 3A-C** and **Extended Data Fig. 2**). Importantly, these state neural signatures were captured prior to participants learning the order of states in each sequence. Next, we applied each state classifier to MEG data acquired during the planning period of each decision trial, producing time series of decoded state reactivation (**Fig. 3D**). Using general linear modelling, we assessed evidence for temporally-ordered reactivation of each state pair (A-B and B-C in path 1, and D-E, and E-F in path 2) across different time intervals (10 to 600 ms, in steps of 10 ms), in both a forwards and backwards direction. We refer to this as "sequenceness", our index of replay.

As a first step, we asked whether there was evidence for replay of the entire state space (i.e., average sequenceness across all four transitions), discarding the first four trials in each block as these were forced-choice. We observed maximal forward state-to-state reactivation at 60 ms intervals (or "lags"), and maximal backward state-to-state reactivation at 110 ms (**Fig. 3E**). We then computed a forward-minus-backward sequenceness measure to remove common noise and increase sensitivity. A significance threshold generated by random permutations (see **Methods**) provided evidence for significant forward replay at 20 to 90 ms state-to-state intervals, indicating the state space was replayed during planning with a rapid temporal compression akin to that reported in previous studies[12,14,15,17,47]. Notably, we did not observe significant forwards replay at longer state-to-state intervals of up to 3 seconds, where this might be more indicative of conscious memory retrieval processes during path evaluation or choice deliberation (**Extended Data Fig. 3**).

Replay is modulated by prospective reward and loss

Having found evidence for forwards replay during planning, we next asked whether we could differentiate replay for paths that culminated in either a reward or a loss. For each trial, we averaged sequenceness across the two transitions present within each path (**Fig. 4A**). We then entered these trial-by-trial estimates of path replay at the significant state-to-state intervals identified within our previous analysis (20 to 90 ms) into a series of linear mixed-effects models that accounted for effects of subject, replay interval, and trial duration (i.e., response time; see **Supplementary methods** for detailed model specification).

We first asked whether the expression of replay was influenced by an eventual choice to approach or avoid. Overall, rewarding paths were replayed more strongly than aversive paths during planning ($\beta = 0.014$, $p < 0.001$; **Fig. 4B**). Notably, there was a significant interaction with choice ($\beta = -0.018$, $p < 0.001$) showing this was particularly the case when participants made an eventual decision to avoid ($EMM_{\Delta} = -0.008$, $p < 0.001$). Replay

strength did not differ between reward and loss paths when participants planned to approach (EMM_{Δ} = 0.002, p = 0.362). Thus, replay preceding a choice to avoid was stronger for potential paths leading to reward than for potential paths leading to punishment.

We next asked whether replay was modulated by factors other than choice; namely, recent path experience or the probability of transitioning to either path irrespective of path value. We operationalised recent experience as the number of trials within a block since participants last visited a particular path. We constructed a model in which replay was predicted by path experience (log-transformed to address a positive skew), path type (reward or loss), and path transition probability. Intriguingly, we found an interaction between path type and experience (β = 0.010, p < 0.001), showing that rewarding paths were more strongly replayed when they had been less recently experienced, whereas loss paths were more weakly replayed (**Extended Data Fig. 4B**). Path transition probability modulated this effect (β = 0.014, p = 0.010), such that less recently experienced rewarding paths were even more strongly replayed when the upcoming transition probability was higher.

In our next model, we assessed whether path replay, irrespective of reward or loss, was modulated by its transition probability. We modelled replay of each path per trial as being predicted by its transition probability, as well as the subsequent choice made on each trial. We found no evidence for an effect of path transition probability on replay (β < 0.001, p = 0.830), regardless of which choice was being planned (β = -0.004, p = 0.190; **Extended Data Fig. 4C**). This indicates that participants' beliefs about which path was more likely to be experienced did not impact the strength of replay.

Lastly, we asked whether evidence for a conscious retrieval of states during choice deliberation influenced the strength of path replay. Although our behavioural strategy modelling suggested participants did not have a bias towards evaluating one path more than another (**Extended Data Fig. 5**), we speculated that participants might differentially recollect states belonging to rewarding or aversive paths *after* appraising each path's value, as a way of simulating future outcomes during choice deliberation. We computed a measure of overall state reactivation throughout planning as an indicator of memory reactivation that might, in principle, be akin to conscious memory retrieval. We found that, overall, states belonging to paths leading to reward were reactivated more strongly overall (β < 0.001, p = 0.027; **Extended Data Fig. 6A**), but, crucially, a significant effect of choice and path type on replay remained even after accounting for such overall state reactivation (β = -0.008, p < 0.001; **Extended Data Fig. 6B**).

Replay predicts approach and avoidance

Stronger replay for rewarding paths when subjects planned to avoid indicates a relationship between the content of replay and subsequent decision-making. To investigate this further, we computed a measure of "differential" replay that captures a difference in the expression of sequenceness between each prospective path on a trial-by-trial basis. Specifically, we subtracted loss path replay from reward path replay, such that more positive

differential replay indicates a bias towards replaying paths leading to reward, and vice versa for more negative differential replay.

Using this differential replay measure, we modelled how replay content changed conditional on the choice being planned (i.e., to approach or avoid), as well as the environment prospects (i.e., the cumulative gain or loss for each path, and the probability of transitioning to each path). To simplify the model, we used the expected value of approaching on each trial as a summary metric of an environment's prospects (equivalent to the total sum of points for each path weighted by their respective transition probabilities). To then predict trial-by-trial decision-making, we constructed a model that allowed expected value to interact with differential replay at all significant replay intervals (20 to 90 ms), as well as the certainty of path transition probabilities and response times.

At a behavioural level, we observed a sigmoidal relationship between expected value and choice ($\beta = 0.432$, $p < 0.001$; **Fig. 4C**), such that participants were more likely to approach when the associated expected value was ≥ -1.2. This is below a rational indifference point of 1, indicating participants were more likely to approach environments with poorer prospects overall. Additionally, participants were more likely to approach when path transition probabilities were more certain ($\beta = 0.270$, $p < 0.001$).

At a neural level, trial-by-trial differential neural replay predicted choice ($\beta = -0.713$, $p < 0.001$), such that participants were more likely to approach when differential replay during planning was less positive, reflecting a bias towards replaying paths leading to potential loss and/or a bias away from replaying paths leading to potential reward. Importantly, this effect of differential replay on decision-making interacted with expected value ($\beta = 0.133$, $p = 0.008$), such that a bias away from replaying paths leading to reward was even more pronounced when participants planned to approach environments with a more negative expected value.

Our use of a difference measure precludes knowing whether the above effect was driven by diminished replay of reward paths or enhanced replay of loss paths. To unpack this, we duplicated our model but replaced differential replay with two separate predictors, one for reward path replay and one for loss path replay, with each separately interacting with expected value. This revealed that path replay for reward and loss had opposing interactions with expected value, such that planning to approach a more hazardous environment (i.e., negative expected value) was predicted by enhanced replay of paths leading to loss ($\beta = 0.120$, $p = 0.090$) and an attenuated replay of paths leading to reward ($\beta = -0.146$, $p = 0.031$; **Fig. 4C**). Moreover, as highlighted by our earlier analyses of replay and path value, when participants planned to approach, replay of reward paths was significantly reduced ($\beta = -1.232$, $p < 0.001$). Replay of loss paths did not predict decision-making ($\beta = 0.189$, $p = 0.313$). Thus, the content of replay predicted subsequent decisions such that when, participants exhibited more rational decision-making (i.e., choosing to avoid when the expected value of an approach choice was lower), paths leading to reward were selectively replayed. By contrast, reduced replay of reward paths and relatively stronger replay of loss paths was associated with participants being more likely to approach riskier environments.

Trait anxiety and risk aversion

Next, we tested an hypothesis that a relationship between differential replay during planning and deciding to approach a risky environment would be amplified in participants with higher trait anxiety and/or a higher propensity towards risk-taking. An independent components analysis on subjects' self-report questionnaires yielded one component representing anxiety and another representing risk-aversion (see **Methods**). Based upon this, we then constructed a model in which these personality traits were allowed to interact with both differential replay during planning and expected value to predict future decision-making. We again included the degree of certainty about the path transition probabilities in the model, as well as trial duration (i.e., response time).

Within this model, anxiety and risk-aversion alone did not predict decision-making ($\beta = -0.028$, $p = 0.644$ and $\beta = 0.073$, $p = 0.284$, respectively), although there was a significant increase in approach rate at higher expected values (indicating more conservative decision-making) in participants with higher risk aversion ($\beta = 0.008$, $p = 0.048$). Instead, both anxiety and risk aversion significantly modulated the relationship between differential replay and decision-making. More anxious ($\beta = -0.314$, $p = 0.003$; **Fig. 4D**) and more risk-averse participants ($\beta = -0.377$, $p < 0.001$; **Fig. 4E**) showed a greater likelihood of approaching when replay was biased away from rewarding paths ($\beta = -0.314$, $p = 0.003$). For more risk-averse participants, this was the case regardless of expected value ($\beta = -0.018$, $p = 0.540$), whereas for more anxious participants this was predominantly the case when expected value was lower ($\beta = 0.096$, $p = 0.014$).

We repeated this model using separate interacting predictors for reward and loss path replay to detail what was driving the above effects. The model revealed that replay for paths leading to loss ($\beta = 0.450$, $p = 0.003$), but not reward ($\beta = -0.199$, $p = 0.189$), was boosted in more anxious participants when approaching more aversive environments. Similarly, replay for paths leading to loss ($\beta = 0.483$, $p < 0.001$), but not reward ($\beta = -0.225$, $p = 0.084$), was boosted for more risk-averse participants when planning to approach any environment. Additionally, more risk-averse participants had diminished replay of rewarding paths when planning to approach more aversive environments, while more risk-seeking participants had diminished replay of rewarding paths when planning to approach more lucrative environments ($\beta = -0.132$, $p = 0.004$). In contrast, more anxious participants had stronger replay of loss paths when planning to approach more aversive environments ($\beta = -0.201$, $p < 0.001$). This suggests that the more negative differential replay in participants with higher trait anxiety during planning was driven by an increase in loss path replay rather than a decrease in reward path replay, while the opposite was true for approach planning in more risk-averse participants.

Fronto-temporal theta activity underlies replay during planning

In a final analysis, we estimated the spatial sources of activity underlying onset of replay events. We defined a replay "event" as an above-chance reactivation of one state followed by reactivation of the following state within

a 20 to 90 ms lag, with additional stringent criteria (see **Methods**). We reconstructed source activity in either the theta (4 to 8 Hz) or high gamma (120 to 150 Hz) frequency band based upon a priori evidence for expression of hippocampal theta-related replay during planning[48,49], as well as high-frequency sharp-wave ripple events in hippocampus related to planning[19].

Across the whole brain ($p < 0.05$, FWE-corrected), there was a significant increase in theta power in the right thalamus (peak MNI: 5, -26, 8), as well as a cluster spanning the left middle temporal gyrus that overlapped left posterior hippocampus (peak MNI: -40, -31, -2). We also observed significant theta activity in dorsolateral prefrontal cortex (DLPFC; peak MNI: -35, 29, 28), right anterior cingulate cortex (ACC; peak MNI: 10, 49, 13), striatum (peak MNI: -15, 4, 13), and inferior occipital cortex (0, -101, -12). In contrast, we did not observe significant source activity in the gamma frequency range during replay events. We also investigated whether theta or high-gamma activity during replay events covaried with each subject's trait anxiety or overall performance accuracy, but we did not observe any significant effects.

The increased theta activity in medial temporal lobe accords with studies in rodent hippocampus, where a rapid "look-ahead" of spatial trajectories during route planning is reflected by rapid hippocampal replay events bounded by theta cycles[49]. In humans, theta activity in hippocampus and medial temporal lobe has been observed during prospective replay events when participants plan to avoid aversive outcomes[48], similar to the present study. Other studies in humans have localised replay onset during post-task rest periods (associated with memory consolidation of a cognitive map) to left hippocampus in the gamma frequency band[14,15,47]. This pattern is in line with the notion that planning-related replay in medial temporal cortex during is subserved by theta activity, whereas replay related to memory consolidation at rest is more closely linked to high-frequency sharp wave ripple events[18,49–51].

Our results support previous evidence for a role for ACC[52], DLPFC[53,54], striatum[55], and inferior occipital cortex[15] in prospective replay during planning that involves elements such as rule-switching or reward re-evaluation. Intriguingly, our results also hint at the thalamus being a significant source of replay-related theta activity. The thalamus purportedly coordinates reward-guided decision-making processes across hippocampus, medial temporal lobe, and prefrontal cortex[56], and thus might reasonably be involved in long-range communication of ordered state reactivation across these areas during planning.

Discussion

In rodent studies, replay content during planning has been found to reflect paths that should be pursued[23,24] as well as those that should be avoided[33]. Here, in the context of an approach-avoidance conflict in humans, we find that the content of forward replay during planning flexibly predicted subsequent decisions. Participants were more likely to avoid when replay was relatively stronger for paths leading to reward, and more likely to approach

when replay was relatively stronger for paths leading to loss, an effect most pronounced for risky environments (i.e., there was a negative expected value of approaching). Our findings indicate a role for replay during planning under uncertainty, where the relative strength of replay for paths leading to reward and loss is weighted towards counterfactual outcomes relating to a current plan to approach or avoid.

Based on rodent studies, we had expected prospective replay content to reflect the goals of approach (to obtain reward) and avoidance (to avoid punishment), such that replay would increase for rewarding paths being pursued[21,23,24,26,50] and for punishing paths being avoided[33]. Instead, we observed the opposite pattern, albeit in an environment that contained both reward and loss paths. Preceding a decision to avoid, replay was increased for paths that would lead to a foregone reward. By contrast, replay of paths leading to prospective reward was decreased preceding a decision to approach. Indeed, when there was greater risk associated with approach (i.e., a negative expected value), replay increased for paths leading to potential loss. This suggests a relationship between the content of prospective replay and rational decision-making under risk, where boosted replay of rewarding paths predicted rational avoidance and replay of more punishing paths predicted irrational approach.

Our findings echo a recent theoretical account which proposed replay provides a pessimistic reminder of counterfactual outcomes to a model-free learning system[37]. This proposal finds support in observations of increased replay for paths previously – but not currently – rewarded[35,36]. Furthermore, behavioural modelling has linked replay to model-based planning, such that replay of sub-optimal outcomes of a given choice (i.e., "pessimistic" replay) promotes more rational model-free decision-making by ensuring that negative outcomes of unchosen actions are not forgotten[22,37,48]. Intriguingly, we also observed enhanced replay for paths leading to counterfactual outcomes (albeit the hypothetical outcomes of a planned decision, rather than the observed outcomes of a previous decision), though the design of our experiment does not allow us to draw conclusions regarding a contribution of r to either model-based versus model-free learning mechanisms. Our task entailed a high degree of cross-trial volatility in state values and transition probabilities that rendered model-free learning of state-action contingencies futile, as participants needed to adopt a model-based strategy that explicitly considered path values and transition probabilities. Moreover, we found mixed evidence for whether paths were more strongly replayed when they were more susceptible to being forgotten. While we found paths were replayed more strongly when they had not been recently experienced, this was the case solely for rewarding, but not aversive, paths. Additionally, under an assumption that replay prevents forgetting of sequences and their associated values, then planning to avoid should theoretically increase replay of both paths (as neither will be experienced) while planning to approach should increase replay of the less probable path. However, this was only the case for rewarding paths, suggesting that counterfactual replay during model-based planning is not adequately explained by a role in memory maintenance[14,15,18,29–31,57].

An alternative explanation for the pattern of replay we observed is that it reflects an anxiety-related simulation of counterfactual outcomes during planning. This would explain counterfactual replay being associated with

irrational decisions to approach under riskier conditions (i.e., when the expected value of approaching was negative), an effect most pronounced in participants with higher self-reported trait anxiety. Similarly, for participants who self-reported higher trait risk-aversion replay was biased towards paths leading to loss when planning to approach, regardless of the expected value of approaching. Replay has been speculated to play a role in clinical anxiety and depression[41] and our study provides tentative evidence for a relationship between differential replay and both trait anxiety and risk-aversion.

Dispositional anxiety is associated with a heightened, and sometimes uncontrollable, simulation of potential past (rumination) or future (worry) aversive events[40,58]. A functional role for replay in selectively sampling a prospective environment during planning provides a plausible explanation for a bias towards more aversive outcomes in people who have a greater tendency to worry[40]. Indeed, people with higher social anxiety engage in "counterfactual" updating, entailing greater deliberation of outcomes that have not, or will not, be experienced[59]. Thus, a simulation of "what if '' scenarios maps closely with our finding that replay content reflects a worst-case scenario of a plan to approach (i.e., the possibility of being punished) or avoid (i.e., foregoing potential reward). Note that more anxious and more risk-averse participants did not make more erroneous approach decisions overall, and an effect of anxiety and risk-aversion was only discernible at the neural level. Thus, our findings do not provide support for a suggestion that counterfactual replay drives a change in policy per se. Moreover, as our sample consisted solely of healthy controls, future studies involving participants with anxiety disorders, who show irrational risky choice behaviour[60], could determine the extent to which replay relates to anxiety-modulated model-based decision-making.

An important caveat to our study is that reward may have been perceived by participants as more salient than loss, in line with participants' choices being more sensitive to probability and magnitude of reward than that of loss. Playing to accumulate monetary rewards, as opposed to avoiding monetary losses, has been shown to enhance the utility of reward[61]. This might explain why replay reflected a worst-case scenario of choosing to avoid (i.e., foregoing potential reward) across all trials, irrespective of expected value. By contrast, replay reflected the worst-case scenario of choosing to approach (i.e., transitioning to a loss path) only when the expected value of approaching was more negative. An emphasis on reward might also explain why a relationship between path replay and memory maintenance was more evident for rewarding paths (as discussed above) but not punishing paths, in line with other recent findings[36]. Employing a variant of the current design using more arousing positive and negative stimuli (e.g., electric shocks or affective visual stimuli) could help adjudicate between these possibilities.

Overall, we present novel evidence for a relationship between the expression of replay and decision-making under uncertainty. A path prioritisation in prospective replay reflected a worst-case scenario of a decision to approach (increased replay for loss paths) or avoid (increased replay for reward paths). Our findings align with recent observations that replay reflects counterfactual outcomes associated with prospective decision-making[35-37] and

extends this to a domain in which choices to pursue reward also carry a risk of punishment. Scenarios such as this are particularly pertinent to survival where an outcome might be critical for the viability of an agent[62], as well as to understanding anxiety-related disorders that are characterised by an over-simulation of improbable, but often catastrophic, events[40].

Data availability

Data are freely available on the Open Science Framework: https://osf.io/6ndu9/.

Code availability

All code for the experimental paradigm and analysis pipeline is freely available on GitHub: https://github.com/jjmcfadyen/approach-avoid-replay.

Acknowledgements

The authors thank Toby Wise and Paul Sharp for their helpful discussions about the study design. This work is supported by the Wellcome Trust (098362/A/12/Z and 091593/Z/10/Z supporting RJD and JM, respectively). YL is supported by National Science and Technology Innovation 2030 Major Program (2022ZD0205500) and National Natural Science Foundation of China (32271093). The Max Planck University College London Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by University College London and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

JM designed the experiment with input from YL. JM collected the data, and JM and YL wrote the analysis code. JM and YL interpreted the data with input from RJD. JM wrote the manuscript with input and edits from YL and RJD.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1. Decision trials. (A) Participants began each trial in a planning phase. Here, they used the presented information (the odd rule states and the path transition probabilities) to mentally calculate the total outcome for each path and evaluate the utility of an approach vs avoid decision. This calculation involves summing the value (v) of each state (s) across each path, taking into account the 'odd rule', and multiplying the final sum (R_{path}) by the path transition probabilities (P_{path}), as described in B. The order of images and their respective values were learned during an initial training phase (Extended Data Fig. 1). MEG data from this planning period provided the focus for our replay analysis. If participants chose to approach, a screen then appeared displaying which of two potential paths they had probabilistically transitioned to ("Transition" screen), and participants then observed an animation of this sequence ("Walkthrough" screens). During this walkthrough, the number of points gained or lost at each state (light blue numbers), as well as the cumulative sum of points up to and including each state (dark blue numbers), was shown below the state image. Note that images were only shown in forced-choice trials, while text labels were shown in all other trials. The final sum of points for the sequence was then shown ("Outcome" screen). If participants chose to avoid, a fixed increase of one point was shown ("Safe outcome" screen). **(B)** The "odd rule" was introduced to reinforce the temporal order relationships between states by having participants appraise each sequence in a forwards direction. The rule was always applied to one state from each path, and this was indicated to participants on-screen during planning. The odd rule entailed that, if the cumulative sum of points collected up to (and including) a particular state was an odd number, then the sign of the sum would then be reversed (i.e., multiplied by -1) with this sum being carried over to any subsequent states in that path. Thus, the odd rule could significantly alter the total number of points collected along each path, depending on which state the odd rule was applied to, and enforced a need for online calculation. For example, using the values of path 2 illustrated in A, applying the odd rule to state 2 results in -3 points (s_1: 0 + 4 = 4 → s_2: 4 + 1 = 5, which is odd and so the sign is reversed to give -5 → s_3: -5 + 2 = -3), whereas applying the odd rule to state 1 results in 7 points (s_1: 0 + 4, as the sign is not reversed → s_2: 4 + 1 = 5 → s_3: 5 + 2 = 7). A rational planner first calculates the cumulative sum of points along each path (taking the odd rule into account), multiplies these by the respective path transition probabilities (which varied trial to trial), and then decides based on a comparison between the expected value of approaching (EV_{app}) and the expected value of avoiding (EV_{av}).

Figure 2. Behavioural results. (A) Accuracy is defined as the proportion of trials wherein participant responses matched an optimal response, based upon expected value. Overall, participants (individual markers; N = 26) made significantly more accurate avoid decisions than approach decisions (two-tailed t(25) = 4.023, p = 4.591E-4). Horizontal line indicates median and box bounds indicate 25th and 75th quantile. **(B)** The expected value of approaching was significantly higher when participants (N = 26) chose to approach than when participants chose to avoid (two-tailed t(25) = 12.250, p = 4.614E-12). **(C)** Participants (N = 26) were significantly faster to approach than to avoid (two-tailed t(25) = -2.360, p = 0.026). Boxplots indicate median and 25th and 75th percentiles of

average participant response times. **(D)** Approach rate estimated by a behavioural multilevel model showing participants were more likely to approach if the probability of transitioning to a rewarding path was higher, especially when prospective reward values were greater (error bars indicate 95% confidence interval). **(E)** Similarly, participants were more likely to approach if potential loss was lower, irrespective of path transition probability (error bars indicate 95% confidence interval). * p < .05, ** p < .01, *** p < 0.001.

Figure 3. State classification and replay analysis. (A) Before learning the order of images along each path, participants viewed each image in an initial functional localiser task. The visually-evoked event-related fields (measured using MEG) are displayed for each of the 12 images, or "states" (6 were randomly assigned to each participant), averaged across participants (shaded error indicates standard error of the mean). **(B)** Using functional localiser MEG data, we created classifiers for each state, per participant (example participant shown). A classifier was a set of beta weights per sensor. **(C)** Using K-folds cross-validation, we assessed average accuracy of state classifiers per participant. Classifiers trained at a 120 ms time point produced the highest average accuracy overall (error bars indicate standard error of the mean). **(D)** Classifiers trained on either 110, 120, or 130 ms (accounting for inter-subject variability in classifier performance) were applied to MEG data collected throughout the planning period of decision trials, producing matrices of predicted state reactivation per trial (example shown). **(E)** Using a two-level GLM approach, we estimated the intervals (or "lags") between maximal reactivation of each state during planning, in a forwards (left) and backwards (middle) direction. Plots display the sequenceness estimates averaged across all four transitions (shaded error indicates standard error of the mean). The significance threshold is indicated by an horizontal dashed line. Significant forwards-minus-backwards replay occurred at state-to-state intervals of 20 to 90 ms, peaking at 60 ms.

Figure 4. Replay of prospective reward and loss paths. (A) Replay strength for paths leading to either reward (green) or loss (red) during planning, split according to whether participants subsequently chose to approach (left) or avoid (right). Data is averaged across trials and participants. Significant replay intervals are highlighted by the yellow box. The difference between reward and loss replay is also shown (black). **(B)** Estimated marginal means produced by a mixed-effects model (N = 24 participants) predicting replay strength by the total value of a path (reward in green, loss in red) and the choice subsequently made by participants (approach or avoid). Error bars indicate standard error, and significance is given by a two-tailed statistic using a Satterhwaite approximation (p = 3.452E-6). * p < .05, ** p < .01, *** p < 0.001 **(C)** Approach rate (y axis) predicted by a mixed-effects model containing expected value (x axis) and differential replay. When differential replay was more negative (red, indicating relatively stronger replay of loss than reward paths), participants were more likely to approach environments with poorer prospects (i.e., negative expected value). A similar model using separate predictors for reward and loss path replay showed participants were more likely to approach on trials with a negative expected value when replay of rewarding paths was attenuated (green dashed) and when replay of loss paths was enhanced (red solid). The indifference point (i.e., the point at which approach rate should be 50%) is displayed for rational agent behaviour (vertical dashed line). **(D)** Same as **C**, except that participants' trait anxiety and risk-aversion

scores were included in the model. The interaction between differential replay and expected value on choice was driven predominantly by more anxious participants (low/high split is for visualisation purposes only). **(E)** Same as **D**, except data has been split into low and high risk-aversion. More risk-averse participants were more likely to approach when differential replay was more negative, regardless of expected value.

Figure 5. Beamforming analysis on replay onsets. Sources underlying the onset of replay events for any state-to-state transition included the middle temporal gyrus, hippocampus, anterior cingulate cortex (ACC), and thalamus. Significant activity not pictured: striatum, dorsolateral prefrontal cortex (DLPFC), and inferior occipital cortex. Viewing coordinates: left and middle = MNI [-30, -30, 3], right = MNI [5, 47, 7]. Clusters are thresholded at $p < 0.05$, whole brain FWE-corrected.

References

1. Skaggs, W. E. & McNaughton, B. L. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870–1873 (1996).
2. Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145–156 (2001).
3. Diba, K. & Buzsáki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242 (2007).
4. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).
5. Lee, A. K. & Wilson, M. A. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* **36**, 1183–1194 (2002).
6. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
7. Schuck, N. W. & Niv, Y. Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, (2019).
8. Wu, X. & Foster, D. J. Hippocampal replay captures the unique topological structure of a novel environment. *J. Neurosci.* **34**, 6459–6469 (2014).
9. Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63**, 497–507 (2009).
10. Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454–1458 (2012).
11. Karlsson, M. P. & Frank, L. M. Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* **12**, 913–918 (2009).
12. Kurth-Nelson, Z., Economides, M., Dolan, R. J. & Dayan, P. Fast Sequences of Non-spatial State

Representations in Humans. *Neuron* **91**, 194–204 (2016).
13. Friston, K. & Buzsáki, G. The functional anatomy of time: What and when in the brain. *Trends Cogn. Sci.* **20**, 500–511 (2016).
14. Nour, M. M., Liu, Y., Arumuham, A., Kurth-Nelson, Z. & Dolan, R. J. Impaired neural replay of inferred relationships in schizophrenia. *Cell* **184**, 4315–4328.e17 (2021).
15. Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. J. Human Replay Spontaneously Reorganizes Experience. *Cell* **178**, 640–652.e14 (2019).
16. Wallach, H. *et al.* Coordinated hippocampal-entorhinal replay as structural inference. in *Advances In Neural Information Processing Systems 32 (NIPS 2019)* (eds. Wallach, H. et al.) vol. 32 13 (Neural Information Processing Systems (NIPS), 2019).
17. Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D. & Dolan, R. J. Experience replay is associated with efficient nonlocal learning. *Science* **372**, (2021).
18. Ólafsdóttir, H. F., Bush, D. & Barry, C. The Role of Hippocampal Replay in Memory and Planning. *Curr. Biol.* **28**, R37–R50 (2018).
19. Buzsáki, G. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* (2015).
20. Ólafsdóttir, H. F., Carpenter, F. & Barry, C. Task Demands Predict a Dynamic Switch in the Content of Awake Hippocampal Replay. *Neuron* **96**, 925–935.e6 (2017).
21. Xu, H., Baracskay, P., O'Neill, J. & Csicsvari, J. Assembly Responses of Hippocampal CA1 Place Cells Predict Learned Behavior in Goal-Directed Spatial Tasks on the Radial Eight-Arm Maze. *Neuron* **101**, 119–132.e4 (2019).
22. Eldar, E., Lièvre, G., Dayan, P. & Dolan, R. J. The roles of online and offline replay in planning. *Elife* **9**, (2020).
23. Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. *Neuron* **77**, 1163–1173 (2013).
24. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
25. Ambrose, R. E., Pfeiffer, B. E. & Foster, D. J. Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. *Neuron* **91**, 1124–1136 (2016).
26. Zheng, C., Hwaun, E., Loza, C. A. & Colgin, L. L. Hippocampal place cell sequences differ during correct and error trials in a spatial memory task. *Nat. Commun.* **12**, 3373 (2021).
27. Papale, A. E., Zielinski, M. C., Frank, L. M., Jadhav, S. P. & Redish, A. D. Interplay between Hippocampal Sharp-Wave-Ripple Events and Vicarious Trial and Error Behaviors in Decision Making. *Neuron* **92**, 975–982 (2016).
28. Igata, H., Ikegaya, Y. & Sasaki, T. Prioritized experience replays on a hippocampal predictive map for learning. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
29. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Hippocampal replay is not a simple

549 function of experience. *Neuron* **65**, 695–705 (2010).
550 30. Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. Hippocampal place cells construct
551 reward related sequences through unexplored space. *Elife* **4**, e06063 (2015).
552 31. Dragoi, G. & Tonegawa, S. Preplay of future place cell sequences by hippocampal cellular assemblies.
553 *Nature* **469**, 397–401 (2011).
554 32. Singer, A. C. & Frank, L. M. Rewarded outcomes enhance reactivation of experience in the hippocampus.
555 *Neuron* **64**, 910–921 (2009).
556 33. Wu, C.-T., Haggerty, D., Kemere, C. & Ji, D. Hippocampal awake replay in fear memory retrieval. *Nat.*
557 *Neurosci.* **20**, 571–580 (2017).
558 34. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nature*
559 *Neuroscience* vol. 21 1609–1617 Preprint at https://doi.org/10.1038/s41593-018-0232-z (2018).
560 35. Carey, A. A., Tanaka, Y. & van der Meer, M. A. A. Reward revaluation biases hippocampal replay content
561 away from the preferred outcome. *Nat. Neurosci.* **22**, 1450–1459 (2019).
562 36. Gillespie, A. K. *et al.* Hippocampal replay reflects specific past experiences rather than a plan for
563 subsequent choice. *Neuron* (2021) doi:10.1016/j.neuron.2021.07.029.
564 37. Antonov, G., Gagne, C., Eldar, E. & Dayan, P. Optimism and Pessimism in Optimised Replay. *PLOS*
565 *Computational Biology* **18**, e1009634 (2022).
566 38. Quartz, S. R. Reason, emotion and decision-making: risk and reward computation with feeling. *Trends Cogn.*
567 *Sci.* **13**, 209–215 (2009).
568 39. Bach, D. R. *et al.* Human Hippocampus Arbitrates Approach-Avoidance Conflict. *Curr. Biol.* **24**, 1435 (2014).
569 40. Gagne, C., Dayan, P. & Bishop, S. J. When planning to survive goes wrong: predicting the future and
570 replaying the past in anxiety and PTSD. *Current Opinion in Behavioral Sciences* **24**, 89–95 (2018).
571 41. Heller, A. S. & Bagot, R. C. Is Hippocampal Replay a Mechanism for Anxiety and Depression? *JAMA*
572 *Psychiatry* **77**, 431–432 (2020).
573 42. Aupperle, R. L. & Paulus, M. P. Neural systems underlying approach and avoidance in anxiety disorders.
574 *Dialogues Clin. Neurosci.* **12**, 517–531 (2010).
575 43. Loijen, A., Vrijsen, J. N., Egger, J. I. M., Becker, E. S. & Rinck, M. Biased approach-avoidance tendencies in
576 psychopathology: A systematic review of their assessment and modification. *Clin. Psychol. Rev.* **77**, 101825
577 (2020).
578 44. Loh, E. *et al.* Parsing the Role of the Hippocampus in Approach–Avoidance Conflict. *Cereb. Cortex* **27**, 201–
579 215 (2016).
580 45. Schumacher, A. *et al.* Ventral Hippocampal CA1 and CA3 Differentially Mediate Learned Approach-
581 Avoidance Conflict Processing. *Curr. Biol.* **28**, 1318–1324.e4 (2018).
582 46. Schumacher, A., Vlassov, E. & Ito, R. The ventral hippocampus, but not the dorsal hippocampus is critical for
583 learned approach-avoidance decision making. *Hippocampus* **26**, 530–542 (2016).
584 47. Wimmer, G. E., Liu, Y., Vehar, N., Behrens, T. E. J. & Dolan, R. J. Episodic memory retrieval success is

associated with rapid replay of episode content. *Nat. Neurosci.* **23**, 1025–1033 (2020).
48. Wise, T., Liu, Y., Chowdhury, F. & Dolan, R. J. Model-based aversive learning in humans is supported by preferential task state reactivation. *Sci Adv* **7**, (2021).
49. Wikenheiser, A. M. & Redish, A. D. Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294 (2015).
50. Johnson, A. & Redish, A. D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12176–12189 (2007).
51. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Segmentation of spatial experience by hippocampal theta sequences. *Nat. Neurosci.* **15**, 1032–1039 (2012).
52. Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *Elife* **7**, (2018).
53. Kaefer, K., Nardin, M., Blahna, K. & Csicsvari, J. Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching. *Neuron* **106**, 154–165.e6 (2020).
54. Berners-Lee, A., Wu, X. & Foster, D. J. Prefrontal cortical neurons are selective for non-local hippocampal representations during replay and behavior. *J. Neurosci.* (2021) doi:10.1523/JNEUROSCI.1158-20.2021.
55. Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. A. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* **7**, e1000173 (2009).
56. Vertes, R. P. Interactions among the medial prefrontal cortex, hippocampus and midline thalamus in emotional and cognitive processing in the rat. *Neuroscience* **142**, 1–20 (2006).
57. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
58. Hirsch, C. R. & Mathews, A. A cognitive model of pathological worry. *Behav. Res. Ther.* **50**, 636–646 (2012).
59. Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M. & Daw, N. D. Increased and biased deliberation in social anxiety. *Nat Hum Behav* (2021) doi:10.1038/s41562-021-01180-y.
60. Hartley, C. A. & Phelps, E. A. Anxiety and decision-making. *Biol. Psychiatry* **72**, 113–118 (2012).
61. Kuhnen, C. M. Asymmetric learning from financial information. *J. Finance* **70**, 2029–2062 (2015).
62. Mobbs, D., Headley, D. B., Ding, W. & Dayan, P. Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends Cogn. Sci.* **24**, 228–241 (2020).

Methods

Participants

The study was approved by the University College London Research Ethics Committee (9929/002). We recruited 32 healthy volunteers via online advertisements to participate in the first session, which served as an opportunity to practice and as a screening point to exclude participants who found the memorisation or arithmetic in the task

too difficult (see **Methods, Experimental task**). We excluded 1 participant who scored < 80% accuracy when tested on the image order, and 4 participants who scored < 60% accuracy in the decision trials. Thus, 27 participants completed session 2. One of these participants was excluded due to a technical error with MEG data collection. The final sample consisted of 26 right-handed participants (8 males, 18 females) aged between 18 and 35 years (M = 25, SD = 5).

All participants were fluent or native English speakers with normal vision and no current use of psychiatric medication. Each participant provided written consent for each session and were paid £50 (£10 for behavioural session and £40 for MEG session), plus up to £15 bonus (up to £5 for the behavioural session and up to £10 for the MEG session) upon completing the study. Bonuses were calculated by converting the accuracy of each block (i.e., the proportion of times participants made the correct choice) into a monetary value between £0 and £1.

Experimental task

Image learning

The experiment was created for web browser using jsPsych v6.1.0. The experiment was presented in the format of a computer game where participants played the role of an astronaut exploring rooms within a spaceship. There were six rooms in total, arranged as two sequences (or "paths"): path 1 contained rooms A, B, and C, and path 2 contained rooms D, E, and F. Each room (or "state") was represented by a unique image randomly selected from a set of 12 for each participant (**Extended Data Fig. 1A**). During the image learning phase, participants watched an animation of the transitions along each path, in which the images for each room were presented one at a time for 3 seconds each (**Extended Data Fig. 1C**). Participants were then tested on their memory for the order of images in each sequence. Participants were given up to two attempts to reach at least 80% accuracy.

Value learning

After successfully completing the image learning phase, participants then learned to associate an integer value (ranging from -5 to 5, excluding 0) with each room. This integer represented the number of points subjects stood to gain or lose in each room. To learn these values, participants were presented with each sequence four times, with the integer value presented underneath each image (4-second presentation; **Extended Data Fig. 1D**). Participants were then tested on their memory for each individual room's value, as well as their ability to calculate the cumulative sum of points in each room. This process was repeated until participants scored at least 80% accuracy (up to two attempts).

Decision trials

After completing image and value learning, participants then partook in decision trials. At the beginning of each decision trial, participants were placed conceptually "outside" of the environment containing the two learned sequences and could choose to either approach or avoid it (**Fig. 1A**). Avoidance resulted in a guaranteed point increase of +1 and no transition to either path. Approach decisions took participants down one of the two paths, as chosen by the computer. Crucially, however, there was always a degree of uncertainty as to which of the two paths the participant would transition to if an approach decision was made. The transition probability of each path varied from trial to trial and was explicitly conveyed to the participant at the beginning of each trial. There were five possible sets of probabilities: 10-90%, 30-70%, 50-50%, 70-30%, and 90-10% for transitions to paths 1 and 2, respectively. Once transitioned to a path, the transitions to each room within the sequence were deterministic.

Participants were required to use the value map they had learned in the previous stage, in conjunction with the path transition probabilities presented on each trial, to evaluate the utility of making an approach versus an avoid decision. Optimally, this evaluation would reflect an expected value calculation for both approach and avoid choices, such that:

$$EV_{app} = P_1 R_1 + P_2 R_2 \quad (1)$$

where EV_{app} is the expected value of approaching, P_1 and P_2 are the probabilities of transitioning to paths 1 and 2, respectively, and R_1 and R_2 are the total sums of points for paths 1 and 2, respectively, taking into account the odd rule states (see Methods, Planning manipulation). The expected value of avoiding, EV_{av} , was always 1. The decision to approach was considered correct if $EV_{app} \geq EV_{av}$ and the decision to avoid was considered correct if $EV_{app} \leq EV_{av}$.

After each block, the proportion of correct responses was converted into a monetary value and displayed as a bonus. Participants did not receive feedback on the accuracy of their choices throughout the block. They did, however, observe an animation of their subsequent transitions and change in points (**Fig. 1A**). For "avoid" decisions, a screen was displayed with text stating that they had received 1 point (3 seconds). For "approach" decisions, participants were first shown which path had been selected by the computer according to the transition probability ("Path 1" or "Path 2", for 3 seconds). Participants were then shown each state within that path one at a time (2 second presentation), underneath which the state value as well as the running total of points collected along the path was displayed. A blank screen was presented between states (randomly jittered duration between

0.5 and 0.8 seconds). A final screen conveyed the total number of points earned for that trial (2 seconds). Trials were separated by a blank screen (1 second).

After an initial practice block, participants completed 6 (behavioural session) or 10 (MEG session) blocks. Each block contained 18 decision trials. In the practice block, participants were given unlimited time to make their decision and did not earn bonus money. In test blocks, participants were given 30 seconds (indicated by an on-screen timer) to make their choice. Responses were disabled for the first 5 seconds to prevent accidental presses and encourage planning. If no response was made after 30 seconds, participants were penalised -1 point and prompted with a warning message ("Too slow!") and the trial ended.

Planning manipulation

A number of additional features were incorporated into the design of the decision trials to encourage planning, as well as to control for certain variables. One feature was what we term the "odd rule". The purpose of the odd rule was to allow the sum of points along each path to vary from trial to trial, thus encouraging participants to engage in sequential planning. On each trial, the odd rule was applied to two states: one from each path. These two odd rule states were displayed on-screen (as images on forced-choice trials or as text labels on free-choice trials) at the beginning of each trial, alongside the path probabilities (**Fig. 1A**). Participants were instructed that, if the sum of points accumulated up until (and including) an odd rule state was an odd number, then the sign of this cumulative sum would "flip" (i.e., a negative cumulative sum will become positive, and vice versa). This new sum would then be carried over to any subsequent states along the path.

By way of example, assume the values of states A, B, and C in path 1 are -5, -2, and 3, respectively. If state B is the odd rule state, then one must mentally sum the number of points up until (and including) state B (-5 + -2 = -7). One must then consider whether the current sum of points is an odd number. In this case, it is (-7), and thus the sign of the sum is flipped (becoming +7). This value is then carried over to the next state, C (7 + 3 = 10), producing a final outcome of 10. If, instead, state C is the odd rule state, then one sums the number of points up until state C (-5 + -2 + 3 = -4). In this case, the sum of points at the odd-rule state is an even number (-4), and thus no sign-flipping occurs, producing a final outcome of -4. Hence, the final value of each path is entirely dependent on the position of the odd rule state in each path (see **Fig. 1A** for another example). This manipulation increased the variability of final path values across trials. In the MEG session, participants were instructed to refrain from verbalising numbers aloud to minimise movement-related artefacts in the MEG activity.

To further increase the variability in final path values across the experiment, the value of one state from each path changed at the beginning of each block. All state values then remained constant for the duration of the block. So that participants knew which values had changed at the beginning of each block, the first four trials in each block (first six in the practice block) were forced-choice, such that participants could only choose to approach. Forced-

choice trials were controlled so that they lead to an equal number of transitions to path 1 and path 2. Any points gained or lost on these trials did not count towards bonus payment and were not included in planning-related MEG analyses, as participants were unable to plan until having observed the updated values in both paths.

Forced-choice trials were also the only trials in which the images were displayed, both during the planning period (where the states with the odd rules were displayed) and the sequence animation. In all other free-choice trials, images were replaced by their text labels (e.g., "cat" or "bicycle"), which had already been shown to participants during the functional localiser (see **Procedure** below). This was done to control for any potential biased visual exposure to the state images during free-choice trials based on choice behaviour (e.g., only deciding to approach when path 1 is more likely) while still periodically reminding participants of the images associated with each room.

Participants were assigned to one of two experimental protocols in a counterbalanced fashion (**Extended Data Fig. 1E**). Each protocol was designed to minimise the repetition of odd rule state pairs across trials. These two protocols also captured another feature of the design, in which one path more often resulted in a positive outcome and the other in a negative outcome. This was done to maximise the difference in replay between rewarding and aversive paths, by allowing for some degree of association by repetition. To prevent participants from relying on this consistency (and thus not engaging in sequential planning), 5% of trials were catch trials, where either both paths produced a gain or both produced a loss, thus increasing the utility of planning on every trial. Furthermore, the rewarding and aversive paths swapped positions halfway through the experiment (e.g., if path 1 was consistently rewarding at the beginning, it became consistently aversive, and vice versa for path 2). The starting positions of the rewarding and aversive paths were counterbalanced across the two protocols.

Procedure

Initial session

Participants completed two sessions on consecutive days. The first session was a behavioural-only practice, where participants completed three questionnaires: the 12-item Intolerance of Uncertainty Scale[63], 16-item Penn State Worry Questionnaire[64], and 30-item Domain-Specific Risk-Taking Scale[65], each presented in a random order on a computer (approximately 15 minutes). Participants then completed a shorter 45-minute version of the experiment. The aim of this session was to ensure participants were capable of performing the task (at least 80% performance on the image and value memory tests, and at least 60% correct choices on decision trials) before continuing to the MEG session the following day.

Functional localiser

The second session comprised an MEG session. Participants first completed a functional localiser task (30-minutes) and then completed a full 1.5-hour task. In the functional localiser, participants were shown the six unique images (randomly selected per participant) used in the main task. Crucially, these images were different from those shown in the initial behavioural session. On each trial, an image was presented on screen for 1 second (**Extended Data Fig. 1B**). After the image disappeared, two words were presented on the left and the right of the screen. One of these was the correct label for the previous image (e.g. "cat") and the other label was randomly selected from a pool of invalid words. Participants pressed either the left or right button of a 4-button response pad to indicate the correct label. After making a response, the words were replaced by a fixation cross for a randomly jittered inter-trial interval between 0.5 and 1.5 seconds. Correct and incorrect responses produced a green or red cross, respectively. There were four blocks, within which each image was randomly presented 20 times, giving 80 trials in total per image. Across the 26 participants, the mean response accuracy was 97.48% (SD = 2.48%, range = 90.63 to 99.79%).

MEG analysis

MEG acquisition and preprocessing

Participants' neural activity was measured using a CTF Omega MEG scanner with a 275-channel axial gradiometer whole-head system (CTF Omega, VSM MedTech) at University College London. Participants were seated upright in the scanner and head position was continuously monitored by three head position indicator coils located at the nasion and left and right pre-auricular fiducial points. Data were acquired continuously at 1,200 Hz and participants' eye movements were recorded using an Eyelink eye-tracking system. Triggers were recorded using a photodiode positioned behind the stimulus presentation screen that detected the onset of a flashing white stimulus (hidden from view) that was synchronised with event onsets.

MEG data from the functional localiser and decision trials were preprocessed using SPM12 (Wellcome Centre for Human Neuroimaging), Fieldtrip (2019), and custom code written in MATLAB R2018b (MathWorks). All code is available on GitHub: https://github.com/jjmcfadyen/approach-avoid-replay. CTF data for each block were imported using OSL (the OHBA Software Library, from OHBA Analysis Group). Trigger onset times and durations were extracted from the photodiode signal and semi-automatically checked for errors. Next, the data were high-pass filtered at 0.5 Hz to reduce slow drift, and a notch filter for 50 Hz was applied to remove line frequency. The data were then downsampled to either 100 Hz (for replay analysis, to reduce temporal autocorrelation) or 600 Hz (for source reconstruction), thereby reducing computational load and increasing signal to noise ratio. OSL also

identified potential bad channels whose characteristics fell outside the normal distribution of values for all sensors.

Independent component analysis was then performed on the data (FastICA, http://research.ics.aalto.fi/ica/fastica), decomposing it into 150 independent spatiotemporal components. Artefactual components were automatically classified using the combined spatial topography, time course, time course kurtosis, and frequency spectrum of all components. For example, eye blink artifacts exhibited high kurtosis (>20), a repeated pattern in the time course, and consistent spatial topographies. The number of excluded components was limited to a maximum of 20. Artefacts were rejected by subtracting them out of the data. All subsequent analyses were performed directly on the filtered, cleaned MEG signal, in units of femtotesla.

The data were then divided into different epochs using the trigger onsets and durations. For the functional localiser, epochs were created for the image onset (-0.1 to 0.8 seconds post-stimulus onset). For the decision trials in the main task, epochs were created for the planning time (-0.1 seconds before trial onset to the response time). Artefactual sensors identified by OSL were interpolated for all epochs, and artefactual functional localiser trials were excluded from the classification procedure.

Image classification

We used Temporal Delayed Linear Modelling (TDLM) to characterise patterns of neural dynamics during the task[66], as performed in previous studies[12,14,15,17,47]. First, for each participant, we classified patterns of multivariate neural activity evoked by each image in the functional localiser (**Fig. 2A**). The purpose of these classifiers was to detect reinstatement of each image representation during planning, likely indicating memory reactivation. This approach capitalises on the similarity between spatial patterns of neural activity evoked by the visual onset of stimuli during conscious viewing and memory retrieval, which has previously been demonstrated in both MEG and fMRI[67,68]. Specifically, an interplay between hippocampus and distributed cortical networks during memory retrieval produces spatial patterns of activity that closely resemble patterns of activity that were produced when stimuli were first experienced[69]. Notably, our stimuli were visually and categorically unique, thus maximising our ability to detect features reinstated during planning (e.g., visual imagery, conceptual associations, etc.)[47].

We selected data from 0 to 300 ms from each functional localiser epoch, excluding incorrect and artefactual trials, as well as trials where response time was > 5 standard deviations from the mean per participant (average of 78 trials per stimulus, per participant; SD = 2, range = 73 to 80). We then constructed a series of Lasso-regularised logistic regression models. Each model received data from a single time sample (0 to 300 ms, at 10-ms resolution) across all trials. Hence, we constructed separate models (per time sample, and per image; 31 × 6) per participant, each using a trials × sensors (e.g., 480 × 275) data matrix and a binary vector indicating which trials belonged to

that image. For each model, we appended a duplicate-sized matrix of zeros to the data matrix to reduce the spatial correlation between each model.

Each lasso-regularised logistic regression model used a range of 100 regularisation parameters (λ) sampled from a half-Cauchy distribution (γ = 0.05, range = 0.0001 to 1). Thus, each model produced a λ × sensors (100 × up to 275) matrix of slope coefficients (**Fig. 2B**), as well as a vector of intercept coefficients for each λ. We refer to these coefficients as our binomial classifiers, each of which are trained to distinguish the sensor data associated with one image as compared to all other images.

To evaluate the accuracy of each classifier per participant, we conducted a *K*-folds cross-validation procedure. *K* was set to the minimum number of trials per stimulus for that participant. In each fold, a test set was created by randomly taking one sample from one exemplar trial per stimulus. The remaining data was used for training. Random selection of the test data was controlled to maximise equal sampling across trials. The classifiers per state generated from the training dataset were then applied to the six test trials (one for each stimulus). Thus, for a given fold, a score of 1 or 0 was given for whether each state classifier maximally predicted the correct trial. The accuracy of each state classifier was given by the average score across folds.

For each subject, we selected λ that produced the highest mean accuracy across state classifiers (λ: M = 0.0017, SD = 0.0015). We then averaged the classification accuracy across states per subject and examined which training times produced the highest accuracy across subjects (**Fig. 2C**). Overall average state classification accuracy exceeded chance (16.66%) for all subjects from 80 ms onwards, peaking at 120 ms (48.97%). Classifier training times from 110 to 150 ms made up the top 15% performance (all > 45.80% accuracy).

Sequential state reactivation

Using our state classifiers, we then estimated the degree to which images were sequentially reactivated in the brain while participants planned whether to approach or avoid the state space in each trial. We utilised an updated general linear modelling approach, which encapsulates a lagged cross-correlation between the evidence for state-to-state transitions. This method produces an overall "sequenceness" statistic at different time intervals, or "lags". We employed this approach on a trial-by-trial basis per participant, using neural data collected during the planning period.

In the first step, we estimated the degree to which each state was reactivated during the planning period of free-choice decision trials by multiplying the spatiotemporal MEG data by each state classifier's beta estimates. We used state classifiers trained at 120 ms post-stimulus onset, which had the highest cross-validated accuracy across subjects. We then entered the resultant time series of predicted state reactivation (states × time matrix;

Fig. 2D) per trial into a 2-level general linear model designed to test whether reactivation of each stimulus occurred in a specific order at different time intervals.

At the first level, we performed a family of multiple regressions for each state's reactivation time series ($i \in [1{:}\,6]$), in which a time-lagged copy of the reactivation time series for state j ($X(t\Delta)_j$) predicts the original, unshifted reactivation time series of state i (X_i). The time lags ranged from 0 to 600 ms, in 10 ms bins. Hence, this analysis evaluated the average likelihood that stimulus i is followed by stimulus j after a time lag of $t\Delta$. Separate linear models were estimated for each stimulus i and each time lag $t\Delta$:

$$X_i = \sum_{i=1}^{6} X(t\Delta)_j \times \beta(t\Delta)_{ij} + C \tag{2}$$

where C is a constant term and $\beta(t\Delta)_{ij}$ is a coefficient derived from ordinary least-squares that captures the unique influence of X_i on $X(t\Delta)_j$. These coefficients are then used to form 6 × 6 empirical transition matrices, $\beta(t\Delta)$, for each time lag.

At the second level, we quantified the evidence for specific, hypothesised state-to-state transitions. In this task, the key state-to-state transitions were A → B and B → C (path 1), as well as D → E and E → F (path 2). These transitions were declared by separate 6 x 6 binary matrices for hypothesised forward (T_F) and backward (T_B) transitions, where $T_F = T_B{}'$. The evidence for the hypothesised transitions was then quantified by:

$$B(\Delta t) = \sum_r Z(r) \times T_r \tag{3}$$

where r is the total number of all regressors included in the second level. These regressors included T_F, T_B, T_{auto}(an identity matrix of self-transitions to control for autocorrelation), and T_{const} (a constant matrix that models away the average of all transitions, ensuring that any weight on T_F and T_B was not due to general dynamics in background neural dynamics). Note that there were four versions of T_F and T_B, one for each hypothesised transition (A → B, B → C, D → E, and E → F). This allowed us to examine the evidence of replay of each transition specifically, which was critical to our path-specific analyses. Z is the weight for each regressor, representing the evidence for the hypothesised state-to-state transitions. Z_F and Z_B are evidence for forward and backward transitions, respectively. A forwards-minus-backwards sequenceness measure, Z_D, was also computed by performing $Z_F - Z_B$, thus removing common variance. Repeating equation 3 at each time lag produces a time

series of sequenceness at different intervals, where smaller intervals indicate more time-compressed replay (Fig. 2E).

To determine the statistical significance of Z (averaged across the four transitions and all trials per participant), we employed non-parametric permutation testing at the second level. We generated a null distribution by generating all possible invalid versions of T_F and T_B, such that they only included cross-path transitions (e.g., A to E, B to D, etc.). This produced 40 null versions of Z. We then calculated a significance threshold for our valid Z by taking the maximum absolute value of each null and computing the 95th percentile for Z_F and Z_B (one-sided test) or the 2.5th and 97.5th percentile for Z_D (two-sided test). Thus, values of Z were deemed statistically significant (FWE < 0.05) if they exceeded these significance thresholds.

To account for inter-subject variability in classification accuracy across training times and their relevance to replay, we also computed sequenceness using classifiers trained on 110 ms and 130 ms (10 ms either side of the winning training time). Thus, we computed sequenceness three times per subject, and chose the classifier training time (110 ms, 120 ms, or 130 ms) that produced the greatest absolute value of Z_D across lags, averaged across all transitions (110 ms = 11 subjects, 120 ms = 10 subjects, 130 ms = 5 subjects).

Source localisation

To investigate the neural sources underlying replay during planning, we used a procedure for identifying replay onsets similar to previous studies[14,15,47]. Replay onsets were defined as time samples where reactivation of one state was followed by reactivation of the following state to a higher degree than that expected by chance. For each trial, we multiplied the state reactivation matrix (X) by a time-shifted version of the state reactivation matrix by lag t ($X(\Delta t)$). We did this separately for each lag found to be significant in the group-level replay analysis (20 to 90 ms) and only investigated forward transitions, as only forwards replay was significant at the group level (**Fig. 3E**). Next, we multiplied X by a state transition matrix (P) that either represented the true sequential order of states or a randomised order (40 randomisations in total, matching the null iterations used in the replay analysis). Then, for each lag and for each iteration of P, we multiplied $X(\Delta t)$ by P to produce a matrix of sequential state reactivation (i.e., replay) per transition across time. We then summed across transitions to produce a vector (R) reflecting an overall estimate of replay.

To demarcate the onset of a replay event, we estimated a significance threshold in a similar manner to the replay analysis. For each null iteration of P, we concatenated R vectors for all lags and all trials into a single vector, which were combined to create matrix N (40 columns: one per randomised state order). We then calculated a significance threshold by computing the maximum value across the columns of matrix N, and then computing its 95th percentile. Thus, this permutation approach controlled for multiple comparisons across time samples and lags, and also maximised our ability to distinguish signal from noise. Individual replay events were marked as

879 instances where replay at any lag exceeded the overall significance threshold. Finally, we excluded any replay
880 events that were preceded by another replay event (of any lag) in the preceding 100 ms.

881 We epoched the MEG data according to the replay onsets (-100 to 150 ms surrounding replay onset) and baseline
882 corrected the data using a -100 and -50 ms window. We then transformed these data to a three-dimensional grid
883 in MNI space (grid step = 5 mm) using a linearly constrained minimum variance beamformer[70,71], as implemented
884 in OSL. Forward models were generated on the basis of a single shell using superposition of basis functions that
885 approximately corresponded to the plane tangential to the MEG sensor array. The sensor covariance matrix for
886 beamforming was estimated using data separately in theta (4 to 8 Hz) and high gamma (120 to 150 Hz) frequency
887 ranges.

888 At the first level, we computed one-sample tests on whole-brain source activity at each time point using
889 nonparametric permutation testing[72] as implemented in OSL. We selected the resultant t-maps for each
890 participant and smoothed the images in SPM12 using a 12 mm FWHM Gaussian kernel. We then entered these
891 into one-sample t-tests (averaged from 0 to 100 ms post-replay onset) in SPM12 for group-level inference, with
892 or without participant trait anxiety or overall performance accuracy added as a covariate. All statistics are $p <$
893 0.05, FWE-corrected at the whole brain cluster level. Anatomical labelling was determined via the Automated
894 Anatomical Labelling Atlas (AAL3) add-on to SPM12[73].

895 Multi-level modelling

896 All analyses were conducted on the MEG session, as the initial behavioural session served purely to acquaint
897 participants with the task structure. We adopted a multi-level modelling approach, which allowed us to examine
898 effects on a trial-by-trial basis. This approach also allowed us to compare conditions with unbalanced trial
899 numbers (e.g., "approach" decisions mostly consisted of trials where reward probability was high, and vice versa
900 for "avoid" decisions).

901 We used the lme4 package implemented in R v3.6. We constructed a series of models that either used: a) choice
902 as a binomial dependent variable, or b) sequenceness as a linear dependent variable. In all models, forced choice
903 trials and catch trials (i.e., trials where both paths resulted in an overall loss or both resulted in an overall gain)
904 were excluded. All predictors were mean-centred. To ensure convergence, the bobyqa optimiser was used and
905 set to 10^6 iterations. Significant interaction terms were followed up by simple slopes analyses using the
906 "interactions" package in R, FDR-corrected for multiple comparisons, and the "emmeans" package in R. We also
907 ensured that all models produced a variable inflation factor (VIF) below 5 and that autocorrelation within the
908 residuals of each model was minimal, as assessed by a Durbin-Watson test[74]; see **Supplementary methods**).

For models including individual differences, we used principal components analysis to reduce the dimensionality of the three self-report questionnaires (intolerance of uncertainty, worry, and risk-taking across 7 domains: ethical, social, health, financial, and recreational) completed at the beginning of the behavioural session. We identified two principal components that together explained 60.55% of the variance (41.52% and 29.49%, respectively; eigenvalues = 1.548, 1.357, 1.040, 0.944, 0.684, 0.457, 0.350). The first component mapped positively on to risk-taking questionnaire scores, while the second component mapped negatively on to intolerance of uncertainty and worry. We refer to these two components as risk-seeking and anxiety, respectively. For interpretability, we inverted these factors, such that more positive values represented higher risk-aversion and higher anxiety, respectively.

Statistics and reproducibility

No statistical methods were used to pre-determine sample sizes but our sample size was similar to those reported in previous publications observing significant replay of state transitions as measured with MEG [12,15], as well as a relationship between replay and individual differences in performance [47]. All statistical analyses were performed using computer code available online (see **Code availability**). Raw behavioural and MEG data are also available online in the interest of experimental reproducibility (see **Data availability**). As the study was a within-subjects design, there was no randomisation to experimental conditions and thus no blinding during data collection or analysis. The stimuli presented to each participant was, however, randomised using a random seed generator based on computer time at the beginning of the experiment. Assumptions of all tests were formally tested. In cases where assumptions of normality were violated, data were log-transformed. Two participants' data were excluded from path-specific MEG replay analysis due to poor behavioural performance in the task (< 60% accuracy), meaning that these participants were unlikely to have processed that the two paths in the experiment resulted in an overall reward or loss.

Methods-only references

63. Carleton, R. N., Norton, M. A. P. J. & Asmundson, G. J. G. Fearing the unknown: a short version of the Intolerance of Uncertainty Scale. *J. Anxiety Disord.* **21**, 105–117 (2007).
64. Meyer, T. J., Miller, M. L., Metzger, R. L. & Borkovec, T. D. Development and validation of the Penn State Worry Questionnaire. *Behav. Res. Ther.* **28**, 487–495 (1990).
65. Blais, A.-R. & Weber, E. U. A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations. (2006).
66. Liu, Y. *et al.* Temporally delayed linear modelling (TDLM) measures replay in both animals and humans. *eLife* vol. 10 Preprint at https://doi.org/10.7554/elife.66917 (2021).
67. Kurth-Nelson, Z., Barnes, G., Sejdinovic, D., Dolan, R. & Dayan, P. Temporal structure in associative retrieval. *Elife* **4**, (2015).
68. Polyn, S. M., Natu, V. S., Cohen, J. D. & Norman, K. A. Category-specific cortical activity precedes retrieval

during memory search. *Science* **310**, 1963–1966 (2005).
69. Horner, A. J., Bisby, J. A., Bush, D., Lin, W.-J. & Burgess, N. Evidence for holistic episodic recollection via hippocampal pattern completion. *Nat. Commun.* **6**, 7462 (2015).
70. Woolrich, M., Hunt, L., Groves, A. & Barnes, G. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *Neuroimage* **57**, 1466–1479 (2011).
71. Van Veen, B. D., van Drongelen, W., Yuchtman, M. & Suzuki, A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans. Biomed. Eng.* **44**, 867–880 (1997).
72. Hunt, L. T. *et al.* Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.* **15**, 470–6, S1–3 (2012).
73. Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J. & Joliot, M. Automated anatomical labelling atlas 3. *Neuroimage* **206**, 116189 (2020).
74. Fox, J. *Applied Regression Analysis and Generalized Linear Models*. (SAGE Publications, 2015).