The roles of online and offline replay in planning

Eran Eldar[1,2,3,†], Gaëlle Lièvre[2,3], Peter Dayan[4,*], Raymond J. Dolan[2,3,*]

1 Departments of Psychology and Cognitive Sciences, Hebrew University of Jerusalem, Israel
2 Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, UK
3 Wellcome Centre for Human Neuroimaging, University College London, UK
4 Max Planck Institute for Biological Cybernetics, Tübingen, Germany

† Correspondence to: eran.eldar@mail.huji.ac.il
* Equal contribution

Abstract

Animals and humans replay neural patterns encoding trajectories through their environment, both whilst they solve decision-making tasks and during rest. Both on-task and off-task replay are believed to contribute to flexible decision making, though how their relative contributions differ remains unclear. We investigated this question by using magnetoencephalography to study human subjects while they performed a decision-making task that was designed to reveal the decision algorithms employed. We characterized subjects in terms of how flexibly each adjusted their choices to changes in temporal, spatial and reward structure. The more flexible a subject, the more they replayed trajectories during task performance, and this replay was coupled with re-planning of the encoded trajectories. The less flexible a subject, the more they replayed previously and subsequently preferred trajectories during rest periods between task epochs. The data suggest that online and offline replay both participate in planning but support distinct decision strategies.

Introduction

Online and offline replay are both suggested to contribute to decision making[1–15], but their precise contributions remain unclear. Replay of experienced and expected state transitions during a task, either immediately before choice or following outcome feedback, is particularly well suited to mediate on-the-fly planning, where choices are evaluated based on the states to which they lead (this is known as model-based planning). Off-task replay might serve a complementary role of consolidating a model of a state space, specifying how each state can be reached from other states and the values of those states. According to this perspective, both types of replay help subjects make choices that are flexibly adapted to current circumstances.

However, a different possibility is that off-task replay also directly participates in planning, by calculating and storing a (so-called model-free) decision policy that specifies in advance what to do in each state[16–19]. Such a pre-formulated policy is inherently less flexible than a policy that is constructed on the fly, but at the same time it decreases a need for subsequent online planning when time itself might be limited. Thus, rather than online and offline replay both supporting the same form of planning, this latter perspective suggests a trade-off between them. In other words online replay promotes an on-the-fly model-based flexibility, whereas offline replay establishes a stable model-free policy.

Despite the wide-ranging behavioural implications of the distinction between model-based and model-free planning[20–23], and much theorising on the role of replay in one or the other form of planning, to date there is little data to suggest whether online and offline replay have complementary or contrasting impacts in this regard. Therefore, we tested the relationship between both online and offline replay and key aspects of decision flexibility that dissociate model-free (MF) and model-based (MB) planning[24]. For this purpose, we first recorded MEG signals from human subjects during rest and while they navigated a specially designed state space. We then characterized each individual subject's

flexibility and decision-making algorithm based on task behaviour, and we analysed their MEG signals seeking evidence of on-task[25] and off-task[10,12] sequences of state representations.

Results

Individual differences in decision flexibility

We used distinct visual images to represent 8 unique states, where occupancy of each state provided a different amount of reward (**Fig. 1a**). Subjects started each trial at a random state and had to choose a movement direction in order to collect reward from subsequent states (**Fig. 1b**). Subjects learnt beforehand how much reward was associated with each state, but they did not know initially where states were in relation to one another. The latter aspect of task structure had to be acquired through trial and error learning in order to be able to implement subsequent moves that delivered the maximal amount of reward.

We assessed subjects' flexibility in three ways. First, after the initial two blocks of trials, we changed the reward associated with each state (**Fig. 1a**; grey numbers) such that persisting with optimal previous moves would result in below-chance performance. Second, after two additional blocks of trials, we informed subjects that two specified pairs of states had switched positions (**Fig. 1a;** 'switch'), again rendering the optimal previous policy now suboptimal. A flexible model-based planner would be capable of re-planning its moves perfectly following each of these instructed changes, since such a planner has acquired

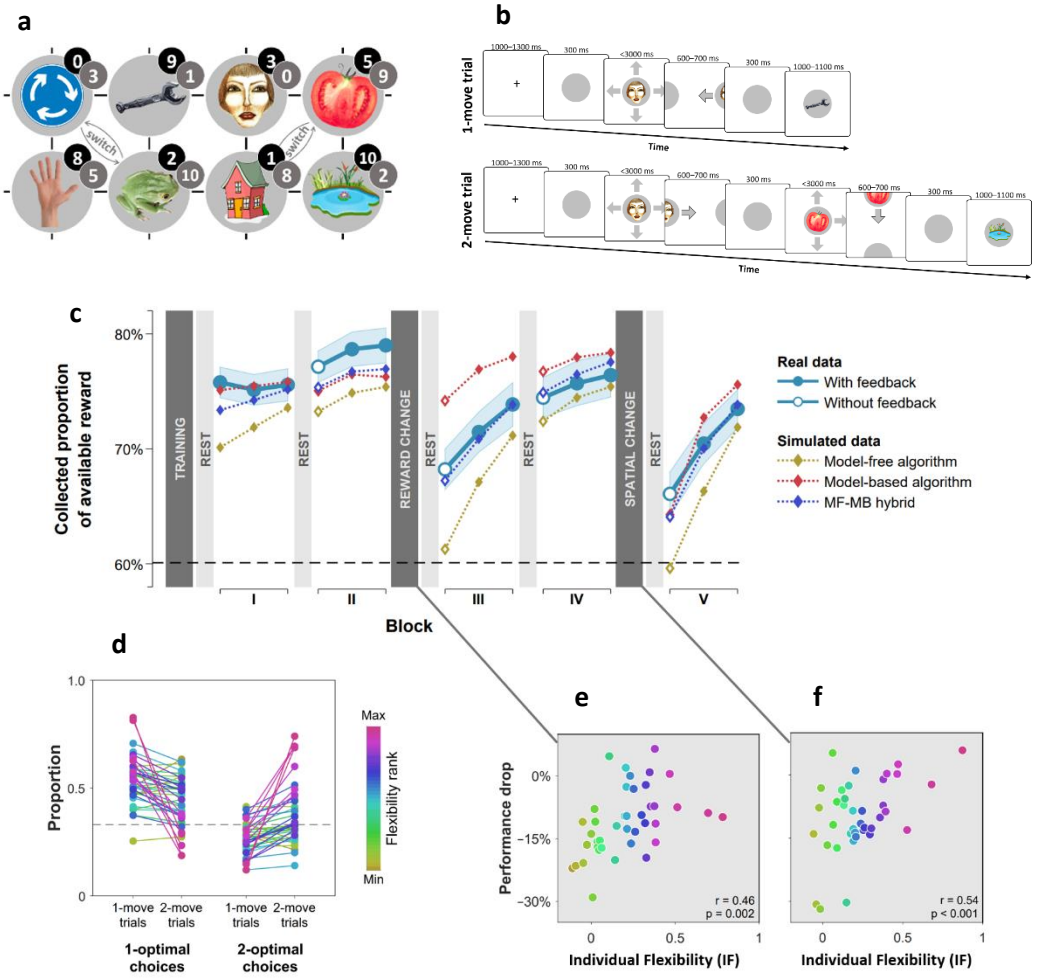


Fig. 1. Subjects differed in decision flexibility. (**a**) Experimental task space. Before performing the main task, subjects learned state-reward associations (numbers in black circles) and they were then gradually introduced to the state space in a training session. After performing the main task for two blocks of trials, subjects learned new state-reward associations (numbers in dark gray circles) and then returned to the main task. Before a final block of trials, subjects were informed of a structural task change such that 'House' switched position with 'Tomato', and 'Traffic sign' switched position with 'Frog'. The bird's eye view shown in the figure was never seen by subjects. They only saw where they started from on each trial and, after completing a move, the state to which their move led. The map was connected as a torus (e.g., starting from 'Tomato', moving right led to 'Traffic sign', and moving up or down from the tomato led to 'Pond'). (**b**) Each trial started from a pseudorandom location from whence subjects were allowed either one ('1-move trial') or two ('2-move trial') consecutive moves (signalled at the start of each set of six trials), before continuing to the next trial. Outcomes were presented as images alone, and the associated reward points were not shown. A key design feature of the map was that in 5 out of 6 trials the optimal (first) move was different depending on whether the trial allowed one or two moves. For instance, given the initial image-reward associations (black) and image positions, the best *single* move from 'Face' is LEFT (9 points), but when two moves are allowed it is best to move RIGHT and then DOWN (5+9 giving 15 total points). Note that the optimal moves differed also given the second set of image-reward associations. On 'no-feedback' trials (which started all but the first block), outcome images were also not shown (i.e., in the depicted trials, the 'Wrench', 'Tomato' and 'Pond' would appear as empty circles). (**c**) The proportion of obtainable reward points collected by the experimental subjects, and by three simulated learning algorithms. Each data point corresponds to 18 trials (six 1-move and twelve 2-move trials), with 54 trials per block. The images to which subjects moved were not shown to subjects for the first 12 trials of Blocks II to V (the corresponding 'Without feedback' data points also include data from 6 initial trials with feedback wherein starting locations had not yet repeated, and thus, subjects' choices still reflected little new information). All algorithms were allowed to forget information so as to account for post-change performance drops as best fitted subjects' choices (see **Materials and methods** for details). Black dashed line: chance performance. Shaded area: SEM. (**d**) Proportion of first choices that would have allowed collecting maximal reward where one ('1-optimal') or two ('2-optimal') consecutive moves were allowed. Choices are shown separately for what were in actuality 1-move and 2-move trials. Subjects are colour coded from lowest (gold) to highest (red) degree of flexibility in adjusting to one vs. two moves (see text). Dashed line: chance performance (33%, since up and down choices always lead to the same outcome). (**e**,**f**) Decrease in collected reward following a reward-contingency (**e**) and spatial (**f**) change, as a function of the index of flexibility (IF) computed from panel **d**. Measures are corrected for the impact of pre-change performance level using linear regression. *p* value derived using a premutation test.

knowledge as to how each state can be reached. Conversely, a pure model-free planner would require complete relearning by trial and error each time there is a change so as to establish a new policy, since such an agent only possesses a now counterproductive policy that specifies where to move from each state.

Examining how subjects' overall performance altered immediately following these changes revealed a decrement in average performance (**Fig. 1c**). However, there were substantial individual differences in this regard, with some subjects seamlessly adapting to reward and position changes, and others showing drops in performance to chance-levels. Subjects whose performance showed a strong decline following a reward change tended to cope poorly also with the position change ($\rho = 0.50$, partial correlation controlling for performance levels before the changes; $p = 0.001$, Permutation test).

As a third, more continuous, test of a different aspect of decision flexibility, we interleaved sets of six trials in which only a single move was allowed ('1-move trials') with trials which allowed two consecutive moves ('2-move trials'; **Fig. 1b**). In 2-move trials, subjects were rewarded for both states they visited, and thus, an optimal course of action often required subjects to move first to an initial low-reward state in order to gain access to a high reward state with their second move. Thus, we defined an individual index (IF) of decision flexibility as the difference between the proportion of moves that were optimal given the actual number of allotted moves and the proportion of moves that would have been optimal given a different number of allotted moves (i.e., had 1-move trials instead involved two moves and 2-move trials involved one move). A value of zero implies no net adjustment, while positive values imply advantageous flexibility.

The results indicate subjects adjusted their choices advantageously to the number of allotted moves (+0.21, SEM 0.05, $p < 0.001$, Bootstrap test), though there was evidence again of substantial individual differences (**Fig. 1d**). Importantly, IF correlated with how well a subject coped with the reward-contingency (**Fig. 1e**) and position (**Fig. 1f**) changes as well with how accurately they could sketch maps of the state space at the end of the experiment ($r = 0.51$, $p < 0.001$, Permutation test; **Supplementary Fig. 1**). Moreover, examining a subset of 2-move trials in which subjects made their second moves without seeing the consequence of their first moves, indicated subjects with high IF planned two steps into the future (**Supplementary Note 1**), as would be expected from MB planning.

Individual flexibility reflected MF-MB balance

These convergent results suggest that IF reflected deployment of a MB planning strategy. To test this formally, we compared how well different model-free and model-based decision algorithms, as well as a combination of both, explained subjects' choices. Importantly, we enhanced these algorithms to maximize their ability to mimic one another (see **Materials and methods** for details). Thus, for instance, the MF algorithm included separate 1-move and 2-move policies.

We found that a hybrid of MF and MB algorithms outperformed substantially either of them alone (Bayesian Information Criterion[26]: MF = 40821, MB = 43249, MF-MB hybrid = 39908), suggesting that subjects employed a mix of MF and MB planning strategies. Simulating task performance using the hybrid algorithm showed it captured adequately differences that were evident between subjects (correlation between real and simulated IF: $r = 0.92$, $p < 0.001$, Permutation test; **Supplementary Fig. 2a**). When we examined each subject's best-fitting parameter values, to determine which of these covaried with IF, we found 84% of inter-individual variance was explained by three parameters that control a

balance between flexible, model-based, and inflexible, model-free, planning (**Supplementary Fig. 2b**). Importantly, less flexible subjects had comparable learning rates and a higher model-free inverse temperature parameter (in 2-move trials), indicating that lower flexibility did not reflect a non-specific impairment, but rather, it was associated with enhanced deployment of a model-free algorithm. Thus, our index of flexibility specifically reflected the influence of model-based, as compared to model-free, planning.

On-task replay is induced by prediction errors and associated with high flexibility

In rodents, reinstatement of past states, potentially in the service of planning, is evident both prior to choices[27] and following observation of outcomes[2]. Thus, we determined firstly at what point states were neurally reinstated during our task. For this purpose, we trained MEG decoders to identify the images subjects were processing (**Fig. 2a**). Such decoders robustly reveal stimulus representations that are reinstated from memory and contribute to decision processes[25,28]. Crucially, image decoders were trained on MEG data collected prior to subjects having any knowledge about the task, ensuring that the decoding was free of confounds related to other task variables (see **Materials and methods**). Applying these decoders to MEG signals from the main task, we found no evidence of prospective representation of outcome states (images) to which subjects will transition at choice (**Supplementary Fig. 4a**). Instead, we found strong evidence that following outcomes (corresponding to new states to which subjects transitioned), subjects represented the states from which they had just moved ($\bar{t} = 3.4$, $p = 0.001$, Permutation test; **Supplementary Fig. 4b**). Consequently, we examined in detail the MEG data recorded following each outcome for evidence of replay of state sequences that subjects had just traversed.

To test for evidence of replay, we applied a measure of "sequenceness" to the decoded MEG time series, a metric we have previously shown is sensitive in detecting replay of experienced and decision-related sequences of states[10,12,25]. Importantly, sequenceness is not sensitive to simultaneous covariation, and thus, it is only found if stimulus representations follow one another in time[25] (as in previous work, we allowed for inter-stimulus lags of up to 200 ms). Thus, following each outcome, we computed sequenceness between the decoded representations of the preceding and the outcome state (**Fig. 2b**). Additionally, MEG signals recorded following the second outcome in 2-move trials were also tested for sequenceness reflecting the trial's first transition (i.e., between the starting state and first outcome; **Fig. 2c**).

Using an hierarchical Bayesian Gaussian Process approach (see **Methods** for details) we tested for timepoints at which sequenceness was evident and correlated with individual flexibility. This method directly corrects for comparison across multiple timepoints by accounting for the dependency between them[29]. Since replay is thought to be induced by surprising observations[16,17,30,31], we also included surprise about the outcome (i.e. the state prediction error inferred by the hybrid algorithm) as a predictor of sequenceness. We found significant sequenceness encoding the last experienced state transition (from 50 to 330 ms and from 820 to 950 ms following outcome onset; **Fig. 2b;** note that the median split is only for display purposes; analyses depended on the continuous flexibility index) and, at the conclusion of 2-move trials, also the penultimate transition (from 130 ms before to 350 ms following outcome onset; **Fig. 2c**). These sequences were accelerated in time, with an estimated lag of 130 ms between the images, and were encoded in a 'forward' direction corresponding to the order actually visited. Moreover, later in the post-outcome epoch, the

penultimate transition was also replayed backwards (from 440 to 940 ms following outcome onset).

Importantly, we found this evidence of replay, across all timepoints, was correlated with IF (mean $\beta = 0.17$, 95% Credible Interval = 0.13 to 0.20), with surprise about the outcome (mean $\beta = 0.06$, CI = 0.03 to 0.10) and with the interaction of these two factors (mean $\beta = 0.19$, CI = 0.15 to 0.22). Thus, sequenceness was predominantly evident following surprising outcomes in subjects with high index of flexibility, consistent with online replay contributing to model-based planning.

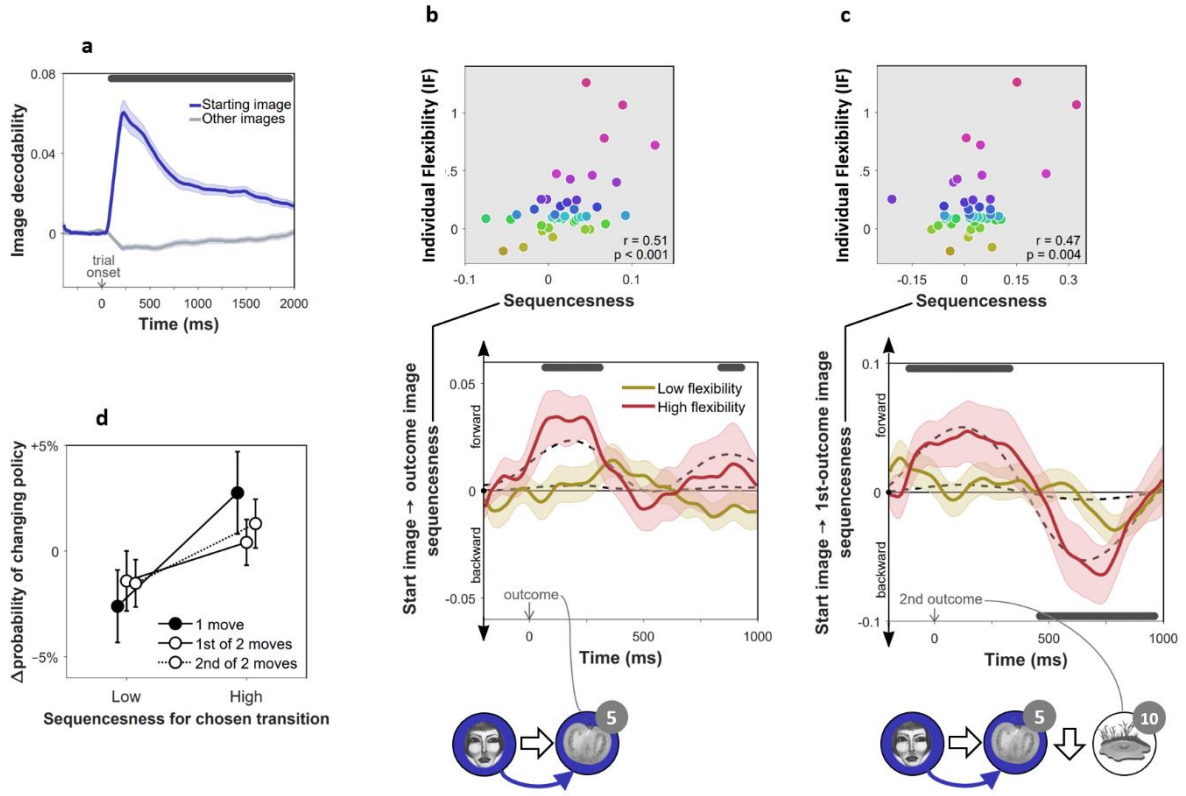


Fig. 2. On-task replay of state-to-state trajectories as a function of individual flexibility. n = 40 subjects. (**a**) Validation of the image MEG decoder used for the sequenceness analyses. The plot shows the decodability of starting images from MEG data recorded during the main task at trial onset. Decodability was computed as the probability assigned to the starting image by an 8-way classifier based on each timepoint's spatial MEG pattern, minus chance probability (0.125). (**b**) Sequenceness corresponding to a transition from the image the subject had just left ('Start image'; in the cartoon at the bottom, the face) to the image to which they arrived ('outcome image'; the tomato) following highly surprising outcomes (i.e., above-mean state prediction error). In the cartoon, the white arrow indicates the actual action taken on the trial; the blue arrow indicates the sequence that is being decoded. For display purposes only, mean time series are shown separately for subjects with high (above median) and low (below median) IF. Positive sequenceness values indicate forward replay and negative values indicate backward replay. As in previous work[25], sequenceness was averaged over all inter-image time lags from 10 ms to 200 ms, and each timepoint reflects a moving time window of 600 ms centred at the given time (e.g., the 1 s timepoint reflects MEG data from 0.7 s to 1.3 s following outcome). Dashed lines show mean data generated by a Bayesian Gaussian Process analysis, and the dark gray bars indicate timepoints where the 95% Credible Interval excludes zero and Cohen's $d > 0.1$. The top plot shows IF as a function of sequenceness for the timepoint where the average over all subjects was maximal. p value derived using a premutation test. (**c**) Sequenceness following the conclusion of 2-move trials corresponding to a transition from the starting image to the first outcome image. (**d**) Difference in the probability of subsequently choosing a different transition as a function of sequenceness recorded at the transition's conclusion. For display purposes only, sequenceness is divided into high (i.e., above mean) and low (i.e., below mean). A correlation analysis between sequenceness and probability of policy change showed a similar relationship (Spearman correlation: $M = -0.04$, $SEM = 0.02$, $p = 0.04$, Bootstrap test). Sequenceness was averaged over the first cluster of significant timepoints from panels **b** and **c**, in subjects with non-negligible inferred sequenceness (more than the standard deviation divided by 10; $n = 25$), for the first time the subject chose each trajectory. Probability of changing policy was computed as the frequency of choosing a different move when occupying precisely the same state again. 0 corresponds to the average probability of change (51%).

On-task replay is associated with changes of policy

Recent theorising regarding the role of replay in planning argues that replay should be preferentially induced when there is a benefit to changing one's policy[17]. This perspective predicts that, at least in our experiment, subjects should be more disposed to replay trajectories that they might not want to choose again, rather than trajectories whose choice reflects a firm policy. To determine whether decodable on-task replay was associated with policy changes, we tested the relationship between sequenceness corresponding to each move that subjects chose, and the probability of making a different choice when occupying the same state later on. We found that moves after which high forward sequenceness was evident corresponded to moves that were less likely to be re-chosen subsequently (**Fig. 2d**), and these policy changes increased the proportion of obtained reward ($M = +11.1\%$, $SEM = 1.5\%$, $p = 0.001$). Thus, evidence of online replay was coupled with advantageous re-planning in relation to the same trajectories.

Off-task replay is induced by prediction errors and associated with low flexibility

We next studied off-task replay, examining MEG data recorded during the 2-minute rest period that preceded each experimental block. Since each block included five frequently repeating starting states, we computed sequenceness for the five most frequent image-to-image transitions subjects chose before and after each rest period (mean choice frequency = 8.4 repetitions per block). As a control analysis, we also examined sequenceness for the five least frequently chosen transitions from the same starting states (mean choice frequency = 1.0 repetitions per block). We found significant evidence for sequenceness throughout the rest periods for frequent transitions (M = 0.002, SEM = 0.001, $p = 0.01$, Bootstrap test). By contrast, no sequenceness was found for the infrequent transitions (M < 0.001, SEM = 0.001, $p = 0.47$, Bootstrap test). Frequent transitions were replayed in a forward direction, with an estimated time lag of 180 ms between images, and prioritized trajectories that induced more reward prediction errors in the previous block (correlation of sequenceness with sum of absolute model-free reward prediction errors inferred by the hybrid algorithm: $M = 0.04$, $SEM = 0.018$, $p = 0.03$, Bootstrap test). Most importantly, off-task sequenceness negatively correlated with IF (**Fig. 3**). This association of sequenceness during rest with low flexibility is consistent with a proposed role of offline replay in establishing model-free policies[16–19].

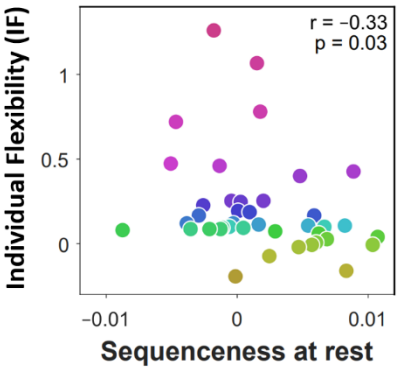


Fig. 3. Off-task replay of past and future trajectories . n = 40 subjects. Individual flexibility as a function of sequenceness in rest MEG data for the five most frequently experienced image-to-image transitions. For each rest period, sequenceness was averaged over transitions from both the preceding and following blocks of trials. p value derived using a premutation test.

Off-task replay can predict subsequently chosen sequences

If offline replay is involved in planning, then its content should predict subjects' subsequent choices. To test this, we dissociated the replay of experienced trajectories from that of planned trajectories, focusing on the third rest period after which the optimal image-to-image transitions changed entirely (due to a change in state-reward associations). As subjects had been taught about the reward change before this rest period, this afforded an opportunity to re-plan their choices accordingly during this rest epoch.

We first examined the behavioural effect of the state-reward change in more detail. The most frequently chosen transitions in the block that followed the third rest period differed from the transitions most frequently chosen in the preceding block (overlap: $M = 14\%$, $SEM = 3\%$), and this policy change was substantially greater than for the other rest periods (overlap: $M = 53\%$, $SEM = 2\%$). As expected, the newly chosen transitions from the following block were advantageous given the new state-reward associations (reward collected: $M = 71\%$, $SEM = 2\%$; chance = 60%) and disadvantageous given the state-reward associations that had so far applied ($M = 52\%$, $SEM = 2\%$).

Given the behavioural change, we focused our examination of the MEG data on evidence for sequenceness during this crucial third rest period. We found that subjects indeed replayed the transitions they subsequently chose ($M = 0.004$, $SEM = 0.002$, $p = 0.02$, Bootstrap test). This replay of subsequently chosen moves indicates subjects utilized a model of the task to re-plan their moves offline[16,17,19]. Our reasoning here is that re-planning in light of the new reward associations, before subjects experienced them in practice, requires a model that specifies how to navigate from one state to another. Indeed, multiple regression analysis showed that low IF was only associated with sequenceness encoding previously chosen transitions ($\beta = -0.35$, $t_{37} = 2.25$, $p = 0.03$), whereas the replay of subsequently chosen transitions did not correlate with IF ($\beta = -0.004$, $t_{37} = 0.03$, $p = 0.97$). On the other hand, the lack of a flexibility enhancement associated with prospective offline replay might indicate that, as might be expected, offline planning is ill-suited for enhancing trial-to-trial flexibility.

Discussion

We find substantial differences in the behaviour of individual subjects in a simple state-based sequential decision-making task that correspond also to a distinction in the nature, and apparent effects, of MEG-recorded on- and off-task replay of state trajectories. These results bolster important behavioural dissociations, as well as provide substantial new insights into the control algorithms that subjects employ. The findings fit comfortably with an evolving literature that addresses human replay and preplay[10–12,25,28].

There is an intuitive appeal to the distinction between model-based and model-free reasoning, confirmed by its close association with many well-established psychological distinctions[32,33]. However, tasks that have become popular for investigating this distinction[24,34–36] have been criticized for offering a better grasp on model-based compared to model-free reasoning processes[36,37]; for rewarding model-based reasoning indifferently[38]; and for admitting complex model-free strategies that can masquerade as being model-based[39].

In our new task, we show a convergence between superficially divergent methods for distinguishing model-based and model-free methods – flexibility to immediate task demands (one-step versus two-step control), preserved performance in the face of changes in the

location of rewards or structure, and an ability to reproduce explicitly, after the fact, the transition structure. Furthermore, the task effectively incentivizes flexible model-based reasoning, as this type of reasoning alone allows collection of substantial additional reward (93%) compared to our most successful MF algorithm (80%). These convergent observations suggest that the model-based and model-free distinction we infer from our task rests on solid behavioural grounds.

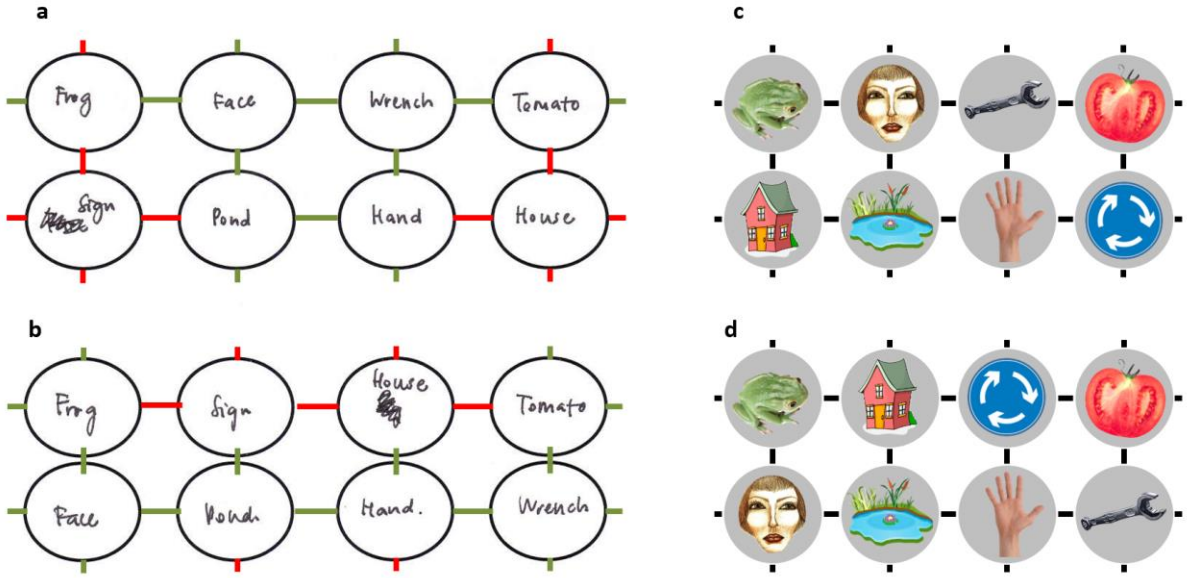
In human subjects, there is a growing number of observations of replay and/or preplay of potential trajectories of states that are associated with the structure of tasks that subjects are performing[10,11,28.] However, it has been relatively hard to relate these replay events to ongoing performance. By contrast, there is evidence that rodent preplay has at least some immediate behavioural function[8,27], and there are elegant theories for how replay should be optimally sequenced and structured in the service of planning[17]. In particular, it has been suggested that replay should prioritize trajectories that can soon be re-encountered, and for which one's policy can be improved. Our results are broadly consistent with this theoretical perspective, showing that new surprising observations precede evidence of corresponding replay, and which in turn predicts appropriate changes in policy. However, rather than preplay immediately prior to choice, we found evidence of on-task replay following feedback alone, suggesting a third potential factor impacting on the timing and content of replay – the need to minimize memory load by embedding new information in ones' policies as soon as it is received.

Critically, the timing and content of replay differed across individuals in a manner that links with their dominant mode of planning. More model-based subjects tended to replay trajectories during learning, predominantly reflecting choices they were likely to reconsider. There have been reports of preferential replay of deprecated trajectories in rodents[8,41]. However, those studies are consistent with a more general function for replay (e.g., maintaining the integrity of a map given a biased experience), whereas in our case, replay was closely related to future behaviour.

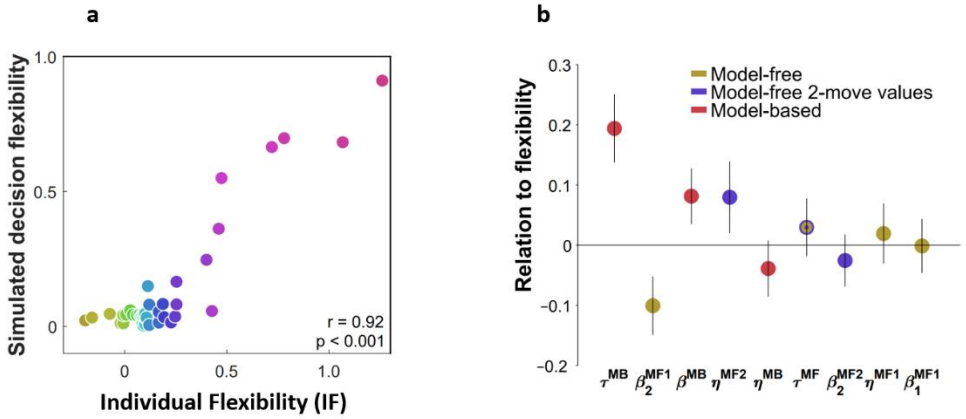
By contrast the decodeable replay of more model-free subjects centred on rest periods, during which DYNA-like mechanisms are hypothesized to compile information about the environment to create an effective model-free policy[17]. This replay of state-to-state transitions suggests that despite a general inability at the end of the task to draw a map accurately, model-free subjects do have implicit access to some form of model, though likely an incomplete one. In any case, generating a policy offline might not be a good strategy for a task that requires trial-to-trial flexibility, consistent with the lack of association here between offline replay and ultimate winnings.

Our work has a number of limitations. First, our experiment was not ideally suited to inducing compound representations that link states with those that succeed them, since succession here frequently changed both within and between blocks. However, algorithms that utilize such representations mimic both model-free and model-based behaviour, and future work could utilize our methods to investigate whether and how these algorithms are aided by online and offline forms of replay[41]. Second, the sequenceness measure that we use to determine replay suffers from a restriction of comparing forwards to backwards sequences. There is every reason to expect both forwards and backwards sequences co-exist, so focusing on a relative predominance of one or the other is likely to provide an incomplete picture. The problem measuring forwards and backwards replay against an absolute standard is the issue of a large autocorrelation in the neural decoding, and better ways of correcting for this are desirable in future studies. Nevertheless, despite these shortcomings the work we report here

is a further step towards revealing the rich and divergent structure of human choice in sequential decision making tasks.



Supplementary Fig. 1. Example sketches of the state space by a representative subject. Subjects sketched the state space at the end of the experiment, recalling how it had been structured before and after (**b**) the position change. On average, subjects sketched much of the state spaces accurately (correct state transitions: first map $M = 0.65$, $SEM = 0.06$; second map $M = 0.56$, $SEM = 0.06$; chance = 0.14, $p < 0.001$, Bootstrap test). (**a**,**b**) Sketches by a representative subject with 0.58 accuracy for the state space before (**a**) and after (**b**) the spatial change. Erroneous transitions are marked in red. (**c**,**d**) The actual state spaces the subject navigated before (**c**) and after (**d**) the position change.



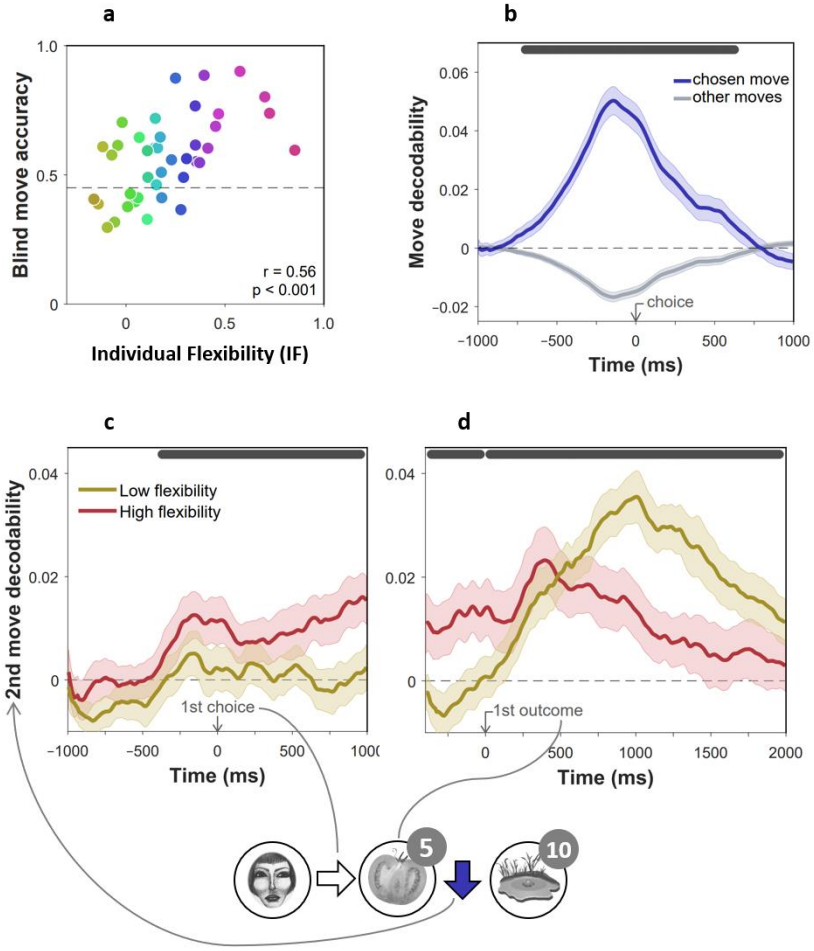
Supplementary Fig. 2. Individual flexibility reflected the balance between MB and MF planning. n = 40 subjects. (**a**) Actual and simulated individual flexibility (IF). Task performance was simulated using subjects' best-fitting parameter settings. IF was computed for each simulated subject and averaged over 100 simulations. (**b**) Relationship between IF and individually-fitted parameters. IF was regressed on subjects' best-fitting parameter settings, including all learning (η), memory (τ), and inverse temperature (β) parameters. Parameters are color-coded by the component of the algorithm they enhance. Error bars: 95% CI.

Supplementary Note 1: Planning two steps into the future

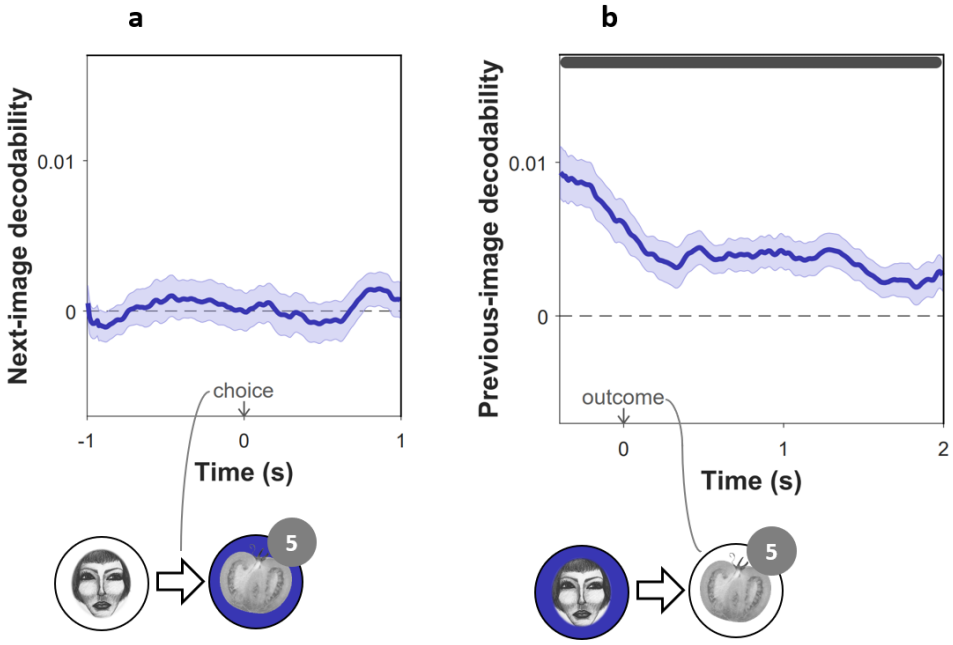
Having a cognitive model that specifies how states are spatially organized makes it possible to plan several steps into the future. To test whether subjects were able to do that, we challenged subjects with 12 'without-feedback' trials at the beginning of each of the last 4 blocks, during which outcome images were not shown. This meant that in 2-move trials subjects had to choose their second move 'blindly', without having seen the image to which their previous move had led (e.g., the tomato in **Fig. 1b**). We found that subjects performed above chance on these blind second moves (proportion of optimal choices: 0.56, SEM 0.03; chance = 0.45; $p < 0.001$, Bootstrap test), and this was the case even immediately following position and reward changes, when subjects could not have relied on previously tested 2-move sequences (0.52, SEM 0.03; $p = 0.01$, Bootstrap test). Most importantly, such blind-move success was correlated with IF (**Supplementary Fig. 3a**).

This result indicates that more flexible subjects were better able to plan two steps into the future when required. Examining response times suggested flexibility was associated with advance planning also when it was not required. Thus, we found that IF correlated with quicker execution of second moves in general (Spearman correlation with median reaction time: $r = -0.61$, $p < 0.001$, Permutation test). To determine whether advance planning was indeed generally associated with flexibility, we examined at what point during a trial their choices became decodeable from MEG signals. For this purpose, we trained a decoder to decode chosen moves from MEG signals recorded outside of the main task (see **Materials and methods** for details). Validating the decoder on MEG data from the main task showed that chosen moves became gradually more evident over the course of the trial, their decodability peaking 140 ms before a choice was made (**Supplementary Fig. 3b**).

Thus, we used the move decoder to test whether second-move choices began to materialize in the MEG signal even before subjects observed the outcomes of their first moves. We found that chosen second moves were indeed decodeable already during first-move choices (decodability: $M = 0.006$, 95% Credible Interval $= 0.004$ to 0.008, Bayesian Gaussian Process analysis; **Supplementary Fig. 3c**) and prior to the appearance of the first outcome (decodability: $M = 0.004$, 95% Credible Interval $= 0.002$ to 0.006; **Supplementary Fig. 3d**). Importantly, this early decodability was correlated with IF (β: $M = 0.29$, 95% Credible Interval $= 0.24$ to 0.34). By contrast, later decodability, following the onset of the second image, did not correlate with IF (β: $M = 0.02$, 95% Credible Interval $= -0.02$ to 0.05). Thus, neural and behavioural evidence concur with the notion that flexibility was associated with planning second moves prospectively.



Supplementary Fig. 3. Evidence of advance prospective planning in flexible subjects. n = 40 subjects. (**a**) Proportion of optimal choices in second moves for trials without feedback, as a function of individual index of flexibility (IF). In such trials, second moves were enacted without seeing the state they were made from. Measures are corrected using linear regression for accuracy of non-blind moves from the same phases of the experiment. (**b**) Validation of move decoder. The plot shows the decodability of chosen and unchosen moves from MEG data recorded during the main task. Decodability was computed as the probability assigned to the chosen move (right, left, up or down) by a 4-way classifier based on each timepoint's spatial MEG pattern, minus the average probability assigned to the same moves at baseline (400 ms preceding trial onset). A separate decoder was trained for each subject on MEG data recorded outside of the main task, during the image-reward association training phases. (**c**,**d**) Decodability of second moves (the blue arrow in the bottom example cartoon) in 2-move trials during first move choice (**c**) and presentation of the first outcome (**d**), as a function of IF. For display purposes only, mean time series are shown separately for subjects with high (above median) and low (below median) IF. In all panels, dark gray bars indicate timepoints where the 95% Credible Interval excludes zero and Cohen's $d > 0.1$ (Bayesian Gaussian Process analysis). Dashed lines: chance decodability level.



Supplementary Fig. 4. Previous, not subsequent, images were encoded in MEG. (**a**) Decodability during choice, of the image to which the chosen move led subsequently, in high- and low-flexibility subjects. (**b**) Decodability following outcome, of images subjects had visited earlier in the trial. In both panels, the analysis excluded decoded probabilities assigned to the image presently on the screen. Dark gray bars indicate timepoints where the 95% Credible Interval excludes zero and Cohen's $d > 0.1$ (Bayesian Gaussian Process analysis). Dashed lines: chance decodability level. Example trials are shown below the plots with decoded elements marked in blue.

Materials and Methods

Subjects. 40 human subjects, aged 18–33 years, 25 female, were recruited from a subject pool at University College London. Exclusion criteria included age (younger than 18 or older than 35), neurological or psychiatric illness, and current psychoactive drug use. To allow sufficient statistical power for comparisons between subjects, we set the sample size to roughly double that used in recent magnetoencephalography (MEG) studies on dynamics of neural representations[28,42], and in line with our previous study of individual differences using similar measurements (including 'sequenceness')[25]. Subjects received monetary compensation for their time (£20) in addition to a bonus (between £10 and £20) reflecting how many reward points subjects earned in the experiment task. The experimental protocol was approved by the University of College London local research ethics committee, and informed consent was obtained from all subjects.

Experimental design. To study flexibility in decision making, we designed a 2x4 state space where each location was identified by a unique image. Each image was associated with a known number of reward points, ranging between 0 and 10. Subjects' goal was to collect as much reward as possible by moving to images associated with a high numbers of points. Subjects were never shown the whole structure of the state space, and thus, had to learn by trial and error which moves lead to higher reward.

Subjects were first told explicitly how many reward points were associated with each of the eight images. Subjects were then trained on these image-reward associations until they reliably chose the more rewarding image of any presented pair (see **Image-reward training**).

Next, the rules of the state-space task were explained (see **State-space task**), and multiple-choice questions were used to ensure that subjects understood these instructions. To facilitate learning, subjects were then gradually introduced to the state space, and were allowed one move at a time from a limited set of starting locations (see **State-space training**). Following this initial exposure, the rules governing two-move trials were explained and subjects completed a series of exercises testing their understanding of a distinction between one-move and two-move trials (see **State-space exercise**). Once these exercises were successfully completed, subjects played two full blocks of trials in the state pace, that included both one-move and two-move trials.

We next tested how subjects adapted to a change in the rewards associated with images. For this purpose, we instructed and trained subjects on new image-reward associations (see **State-space design**). Subjects then played two additional state-space blocks with these modified rewards.

Finally, we tested how subjects adapted to changes in the spatial structure of the state space. For this purpose, we told subjects that two pairs of images would switch locations, informing them precisely which images these were (see **State-space design**). Multiple-choice questions were used to ensure that subjects understood these instructions. Subjects then played a final state-space block with this modified spatial map.

At the end of the experiment, we also tested subjects' explicit knowledge, asking them to sketch maps of the state spaces and indicate how many points each image was associated with before, and after, the reward contingency changed.

Stimuli. To ensure robust decoding from MEG, we used 8 images that differed in colour, shape, texture and semantic category[43–45]. These included: a frog, a face, a traffic sign, a tomato, a hand, a house, a pond, and a wrench.

State-space task. Subjects started each trial in a pseudorandom state, identified only by its associated image. Subjects then chose whether to move right, left, up, or down, and the chosen move was implemented on the screen, revealing the new state (i.e., as its associated image) to which the move led. In 'one-move' trials, this marked the end of the trial, and was followed by a short inter-trial interval. The next trial then started from another pseudorandom location. In 'two-move' trials, subjects made an additional move from the location where their first move had led. This second move disallowed backtracking the first move (e.g., moving right and then left). Subjects were informed they would be awarded points associated with any image to which they move. Thus, subjects won points associated with a single image on one-move trials, and the combined value of the two images on two-move trials. The numbers of points awarded were never displayed during the main task. Every 6 trials, short text messages informed subjects what proportion of obtainable reward they had collected in the last 6 trials (message duration 2500 ms).

Each state-space block consisted 54 trials, 18 one-move and 36 two-move trials respectively. The first 6 trials were one-move, the next 12 were two-move trials, then the next 6 were again one-move trials, the next 12 two-move, and so on. Every 6 trials, short text messages informed subjects whether the next 6 trials were going to be one-move or two-move trials (message duration 2000 ms). Every six consecutive trials featured 6 different starting locations. The one exception to this were the first of the 24 two-move trials of the experiment, where in order to facilitate learning, each starting location repeated for two consecutive trials (a similar measure was also implemented for one-move trials during training; see **State-space training**). Subjects' performance improved substantially in the second of such pairs of trials (Δproportion of optimal first choices = +0.15, 95% CI = +0.11 to +0.18, $p < 0.001$, Bootstrap test).

At the beginning of every block (except the first one), we tested how well subjects could do the task without additional information, based solely on the identity of the starting locations. For this purpose, images to which subjects' moves led were not shown for the first 12 trials. In two-move trials, this meant subjects implemented a second move from an unrevealed image (i.e., state).

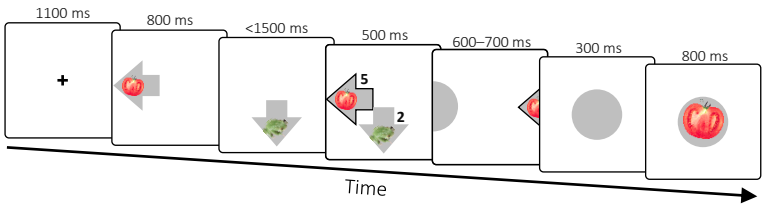
State-space design. The mapping of individual images to locations and rewards was randomly determined for each subject, but rewards were spatially organized in a similar manner for all subjects. To test whether subjects could flexibly adjust their choices, the state space was constructed such that there were five locations from which the optimal initial move was different depending on whether one or two moves were allowed. We tested subjects predominantly on these starting locations, using all five of them in every six consecutive trials. Following two blocks, the rewards associated with each image were changed, such that the optimal first moves in both 1-move and 2-move trials, given the new reward associations, were different from the optimal moves under the initial reward associations. The initial and modified reward associations were weakly anti-correlated across images ($r = -0.37$). Finally, before the last block, we switched the locations of two pairs of images, such that the optimal first move changed for 15 out of 16 trial types (1- and 2-move trials x 8 starting locations).

State-space training. Subjects played six short training blocks, each block consisted 12 one-move trials starting in one of two possible locations. If a subject failed to collect 70% of the points available in one of these short blocks, the block was repeated. The majority of subjects (35 out of 40) had to repeat the first block, whereas only 12% of the remaining blocks were repeated (mean 0.6 blocks per subject, range 0 to 2). Very rarely, a block had to be repeated twice (a total of 5 out of 240 blocks for the whole group). Lastly, subjects played a final

training block consisting 48 one-move trials starting at any of the 8 possible locations. To facilitate learning, during the first half of the block, each starting location was repeated for two consecutive trials. In the second half of the block, starting locations were fully interleaved.

State-space exercise. Following the state-space training, which only included one-move trials, we ensured subjects understood how choices should differ in one- and two-move trials by asking them to choose the optimal moves in a series of random, fully visible state spaces. Subjects were given a bird's eye view of each state space, with each location showing the number of reward points with which it was associated. The starting location was indicated in addition to whether one, or two, moves were available from which to collect reward. In all exercises, the optimal initial move was different depending on whether one or two moves were allowed. Every 10 consecutive exercises consisted of 5 one-move trials and 5 two-move trials. To illustrate the continuity of the state space, the exercise included one-move and two-move trials, wherein the optimal move required the subject to move off the map and arrive at the other end (e.g., moving left from a leftmost location to arrive at the rightmost location). In another two-move trial, the optimal moves involved moving twice up or twice down, thereby returning to the starting location. Subjects continued to do the exercises until fulfilling a performance criterion of 9 correct answers in 10 consecutive exercises. This criterion was relaxed to 8 correct answers if at least 60 exercises had been completed. Only one subject required 60 exercises to reach criterion (mean required exercises = 24.5 exercises, SD 9.3).

Image-reward training. To ensure subjects remembered how many points each image awarded, we required subjects to select the more rewarding image out of any pair of presented images. First, subjects were asked to memorize the number of points each image would awards. Then, each round of training consisted of 28 trials, testing subjects on all 28 possible pairs of images (**Supplementary Fig. 5**). Each trial started with the presentation of one image, depicted on an arrow pointing either right, left, up or down. 800 ms later, another image appeared on an arrow pointing in a different direction. Subjects had then to press the button corresponding to the direction of the more rewarding image. Here, as throughout the experiment, subjects were instructed to press the 'left' and 'up' buttons with their left hand, and the 'right' and 'down' buttons with their right hand. During training, images were mapped to directions such that each of the four directions was equally associated with low- and high-reward images. Once subjects made their choice, the number of points associated with each of the two images appeared on the screen, and if the choice was correct the chosen move was implemented on the screen. Subjects repeated this training until they satisfied a performance criterion, based on how many points they missed consequent upon choosing less rewarding images. The initial performance criterion allowed 4 missed points, or less, in a whole training round (out of a maximum of 130 points). This criterion was gradually relaxed, to 8 missed points in the second training round, to 12 missed points in the third training round, and to 16 missed points thereafter. Once subjects satisfied the performance criterion without time limit, they repeated the training with only 1500 ms allowed to make each choice, until satisfying the same re-set gradually relaxing criterion. Overall, subjects required an average of 3.4 training rounds (SD 1.0) to learn the initial image-reward associations (1.3 rounds without, and then 2.1 rounds with, a 1500 ms time limit), and 4.3 rounds (SD 1.3) to learn the second set image-reward associations (2.0 rounds without, and 2.3 rounds with, a time limit). Questioning at the end of the experiment validated that subjects had explicit recall for both sets of image-reward associations (mean error 0.36 pts, SEM = 0.07 pts; chance = 4.05 pts).



Supplementary Fig. 5. Image-reward training. Timeline of a trial.

Modelling. To test what decision algorithm subjects employed, and in particular, whether they chose moves that had previously been most rewarding from the same starting location (model-free planning), or whether they learned how the state space is structured and used this information to plan ahead (model-based planning), we compared between model-free and model-based algorithms in terms of how well they fitted subjects' actual choices. These models were informed by previous work[46,47], adjusted to the present task, and validated using model and parameter recovery tests on simulated data.

Model-free learning algorithm (free parameters: η^{MF1}, η^{MF2}, τ^{MF}, τ'^{MF}, θ, $\beta_{1,2}^{MF1}$,β_{2}^{MF2}, $\gamma_{\mathrm{up,down,left,right}}$). This algorithm learns the expected value of performing a given move upon encountering a given image. To do this, the algorithm updates its expectation Q^{MF} from move m given image s whenever this move is taken and its outcome is observed:

$$Q_{t+1}^{\mathrm{MF1}}(s_{t,1}, m_t) = Q_t^{\mathrm{MF1}}(s_{t,1}, m_t) + \eta^{\mathrm{MF1}}\delta_t^{\mathrm{MF1}}, \quad (1)$$

where $s_{t,1}$ is trial t's starting image, δ_t^{MF} is the reward prediction error, and η^{MF1} is a fixed learning rate between 0 and 1. Reward prediction errors are computed as the difference between actual and expected outcomes:

$$\delta_t^{\mathrm{MF1}} = R_g(s_{t,2}) - Q_t^{\mathrm{MF1}}(s_{t,1}, m_t), \quad (2)$$

where the actual outcome consists of the points associated with the new image to which the move led, $R_g(s_{t,2})$. $g = 1$ refers to the initial image-rewards associations, and $g = 2$ refers to the second set of image-rewards associations about which subjects were instructed in the middle of the experiment.

On 2-move trials, the algorithm also learns the expected reward for each pair of moves given each starting image. Thus, another set of Q values is maintained (Q^{MF2}), one for each possible pair of moves for each starting image, and these are updated every time a pair of moves is completed based on the total reward obtained by the two moves. This learning proceeds as described by Eqs. 1 and 2, but with a different learning rate (η^{MF2}).

All expected values are initialized to θ, and decay back to this initial value before every update:

$$Q^{\mathrm{MF}} \leftarrow \tau^{\mathrm{MF}}Q^{\mathrm{MF}} + (1 - \tau^{\mathrm{MF}})\theta, \quad (3)$$

where τ^{MF} value retention. This allows learned expectations to be gradually forgotten.

Following instructed changes to the number of points associated with each image, or to the spatial arrangement of the images, previously learned Q values are of little use. Thus, we allow the Q values to then return back to θ, as in Eq. 3, but only for a single timestep and with a different, potentially lower, memory parameter τ'^{MF}.

Finally, the algorithm chooses moves based on a combination of its learned expected values. On 1-move trials, only single-move Q values are considered:

$$\mathrm{p}(m_t = a|s_t) \propto e^{\gamma_m + \beta_1^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}(s_{t,1},m)}, \quad (4)$$

where γ_m is a fixed bias in favor of move m ($\sum_m \gamma_m = 0$), and β_1^{MF1} is an inverse temperature parameter that weighs the impact of expected values on choice. On 2-move trials, both types of Q values are considered. Thus, the first move is chosen based on a weighted sum of the single-move Q values and the move-pair Q values:

$$\mathrm{p}\left(m_{t,1} = m\middle|s_{t,1}\right) \propto e^{\gamma_m + \beta_2^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}(s_{t,1},m) + \beta_2^{\mathrm{MF2}} Q_t^{\mathrm{MF2}}(s_{t,1},m)}, \quad (5)$$

wherein the latter are integrated over possible second moves each weighted by its probability:

$$Q_t^{\mathrm{MF2}}(s_{t,1},m) = \sum_{m^*} \mathrm{p}\left(m_{t,2} = m^*\middle|s_{t,1},m_{t,1}\right) Q_t^{\mathrm{MF2}}(s_{t,1},m,m^*) \quad (6)$$

Then, in choosing the second move the algorithm takes into account the state to which the first move led:

$$\mathrm{p}\left(m_{t,2} = m\middle|s_{t,1},m_{t,1},s_{t,2}\right) \propto e^{\gamma_m + \beta_2^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}(s_{t,2},m) + \beta_2^{\mathrm{MF2}} Q_t^{\mathrm{MF2}}(s_{t,1},m_{t,1},m)}. \quad (7)$$

However, when the newly reached image $s_{t,2}$ is not known (i.e., in trials without feedback, or when estimating $\mathrm{p}\left(m_{t,2} = m^*\middle|s_{t,1},m_{t,1}\right)$ in Eq. 6 before $s_{t,2}$ is reached), Q^{MF1} values are averaged over all settings of $s_{t,2}$.

Model-based learning algorithm (free parameters: η^{MB}, τ^{MB}, τ'^{MB}, ρ, ω, β^{MB}, κ, $\gamma_{\mathrm{up,down,left,right}}$). This algorithm learns the probability of transitioning from one image to another following each move. To do this, the algorithm updates its probability estimates, $\boldsymbol{T}$, whenever a move is made and a transition is observed:

$$T_{t+1}(s_{t,1},m_t,s_{t,2}) = T_t(s_{t,1},m_t,s_{t,2}) + \eta^{\mathrm{MB}}\delta_t^{\mathrm{MB}}, \quad (8)$$

where δ_t^{MB} is the image-transition prediction error, and η^{MF} is a fixed learning rate between 0 and 1. Image-transition prediction errors reflect the difference between actual and expected transitions:

$$\delta_t^{\mathrm{MB}} = 1 - T_t(s_{t,1},m_t,s_{t,2}). \quad (9)$$

To ensure that transition probabilities sum to 1, the transition matrix is renormalized following every update:

$$\forall s \qquad T_{t+1}(s_{t,1},m_t,s) \leftarrow \frac{T_{t+1}(s_{t,1},m_t,s)}{\sum_{s'} T_{t+1}(s_{t,1},m_t,s')}. \quad (10)$$

Learning may also take place with respect to the opposite transition. For instance, if moving right from image $s_{t,1}$ leads to image $s_{t,2}$, the agent can infer that moving left from image $s_{t,2}$ would lead to image $s_{t,1}$. Such inference is modulated in the algorithm by free parameter ρ:

$$T_{t+1}(s_{t,1},\tilde{m}_t,s_{t,2}) = T_t(s_{t,1},\tilde{m}_t,s_{t,2}) + \rho\eta^{\mathrm{MB}}\delta'^{\mathrm{MB}}_t, \quad (11)$$

where $\tilde{m}_t$ is the opposite of m_t, and δ' is the opposite transition prediction error:

$$\delta'^{\mathrm{MB}}_t = 1 - T_t(s_{t,2},\tilde{m}_t,s_{t,1}). \quad (12)$$

Self-transitions are impossible and thus their probability is initialized to 0. All other transitions are initialized with uniform probabilities, and these probabilities decay back to their initial values before every update:

$$T \leftarrow \tau^{\mathrm{MB}} T + (1 - \tau^{\mathrm{MB}}) \frac{1}{7}, \tag{13}$$

where τ^{MB} is the model-based memory parameter. A low τ^{MB} results in faster decay of expected transition probabilities towards uniform distributions, decreasing the impact of MB knowledge on choice.

When instructed about changes to the image locations, the agent rearranges its transition probabilities based on the instructed changes with limited success, as indexed by free parameter ω:

$$T \leftarrow (1 - \omega) T + \omega T^{\mathrm{rearranged}}. \tag{14}$$

Since some subjects may simply reset their transition matrix following instructed changes, the algorithm also 'forgets' after such instruction, as in Eq. 13, but only for a single time point and with a different memory parameter, τ'^{MB}.

Finally, the probability the algorithm will choose a given move when encountering a given image depends on its model-based estimate of the move's expected outcome:

$$\mathrm{p}\left(m_t = m \middle| s_{t,1}\right) \propto e^{\gamma_m + \beta^{\mathrm{MB}} Q_t^{\mathrm{MB}}(s_{t,1}, m)}. \tag{15}$$

The algorithm estimates expected outcomes by multiplying the number of points associated with an image with the probability of transitioning to that image, integrating over all potential future images:

$$Q_t^{MB}(s_{t,1}, m) = \sum_s T_t(s_{t,1}, m, s) R_g(s). \tag{16}$$

When two moves are allowed, the calculation also accounts for the number of points obtainable with the second move, $m_{t,2}$:

$$Q_t^{MB}(s_{t,1}, m) = \sum_s T_t(s_{t,1}, m, s) \left(R_g(s) + \kappa \max_{m'} \sum_{s'} T_t(s, m', s') R_g(s') \right), \tag{17}$$

where κ is a fractional parameter that determines the degree to which reward obtained by the second move is taken into account.

Following the first move, Eq. 15 is used to choose a second move based on the observed new location ($s_{t,2}$). However, if the next location is not shown (i.e., in trials without feedback), the agent chooses its second move by integrating Eq. 15 over the expected $s_{t,2}$, as determined by $T_t(s_{t,1}, m_{t,1}, s_{t,2})$.

MF-MB hybrid algorithm. This algorithm employs both model-free (MF) and model-based (MB) planning, choosing moves based on a combination of the expected values estimated by the two learning processes:

$$\mathrm{p}(m_t = m | s_t) \propto e^{\gamma_m + \beta_1^{\mathrm{MF}} Q_t^{\mathrm{MF}}(s_{t,1}, m) + \beta^{\mathrm{MB}} Q^{\mathrm{MB}}(s_{t,1}, m)}, \tag{18}$$

In 2-move trials, the algorithm makes a choice based on a combination of the model-based Q values and both the single-move and two-move model-free Q values. For the first move, the combination is:

$$\mathrm{p}\left(m_{t,1}=m\middle|s_{t,1}\right)\propto e^{\gamma_m+\beta_2^{\mathrm{MF1}}Q_t^{\mathrm{MF1}}\left(s_{t,1},m\right)+\beta_2^{\mathrm{MF2}}Q_t^{\mathrm{MF2}}\left(s_{t,1},m\right)+\beta^{\mathrm{MB}}Q^{\mathrm{MB}}\left(s_{t,1},m\right)}, \quad (19)$$

with $Q^{\mathrm{MB}}\left(s_{t,1},m\right)$ computed according to Eq. 17. For the second move, the choice is made according to:

$$\mathrm{p}\left(m_{t,1}=m\middle|s_{t,2}\right)\propto e^{\gamma_m+\beta_2^{\mathrm{MF1}}Q_t^{\mathrm{MF1}}\left(s_{t,2},m\right)+\beta_2^{\mathrm{MF2}}Q_t^{\mathrm{MF2}}\left(s_{t,1},m_{t,1},m\right)+\beta^{\mathrm{MB}}Q^{\mathrm{MB}}\left(s_{t,2},m\right)}. \quad (20)$$

When the image is not shown following the first move (i.e., in a no-feedback trial), the agent averages the model-free values over all images.

Parameter fitting. To fit the free parameters of the different algorithms to subjects' choices, we used an iterative hierarchical expectation-maximization procedure[26]. We first sampled 10000 random settings of the parameters from predefined group-level prior distributions. Then, we computed the likelihood of observing subjects' choices given each setting, and used the computed likelihoods as importance weights to re-fit the parameters of the group-level prior distributions. These steps were repeated iteratively until model evidence ceased to increase (see **Model Comparison** below for how model evidence was estimated). This procedure was then repeated with 31623 samples per iteration, and finally with 100000 samples per iteration. To derive the best-fitting parameters for each individual subject, we computed a weighted mean of the final batch of parameter settings, in which each setting was weighted by the likelihood it assigned to the subject's choices. Fractional parameters (η^{MF}, τ^{MF}, τ'^{MF}, η^{MB}, τ^{MB}, τ'^{MB}, ρ, ω, α) were modelled with Beta distributions (initialized with shape parameters a = 1 and b = 1) and their values were log-transformed for the purpose of subsequent analysis. Initial Q values (θ) and bias parameters (γ_{up}, γ_{down}, γ_{left}, γ_{right}) were modelled with normal distributions (initialized with μ = 0 and σ = 1) to allow for both positive and negative effects, and all other parameters were modeled with Gamma distributions (initialized with shape = 1, scale = 1).

Algorithm comparison. We compared between pairs of algorithms, in terms of how well each accounted for subjects' choices, by means of the integrated Bayesian Information Criterion (iBIC)[48,49]. To do this, we estimated the evidence in favour of each model ($\mathcal{L}$) as the mean likelihood of the model given 100000 random parameter settings drawn from the fitted group-level priors. We then computed the iBIC by penalizing the model evidence to account for algorithm complexity as follows: $\mathrm{iBIC}=-2\ln\mathcal{L}+k\ln n$, where k is the number of fitted parameters and n is the number of subject choices used to compute the likelihood. Lower iBIC values indicate a more parsimonious fit.

Algorithm and parameter recovery tests. We tested whether our dataset was sufficiently informative to distinguish between the MF, MB and hybrid algorithms and recover the correct parameter values. For this purpose, we generated 10 simulated datasets using each algorithm and applied our fitting and comparison procedures to each dataset. To reduce processing time, only 10000 parameter settings were sampled. To maximize the chances of confusion between algorithms, we implemented all algorithms with the parameter values that best fitted subjects' choices. Algorithm comparison implicated the correct algorithm in each of the 30 simulate datasets, and the parameters values that best fitted the simulated data consistently correlated with the actual parameter values used to generate these data (Pearson's r: $M = 0.57$, $SEM = 0.05$), and this correlation was stronger for parameters

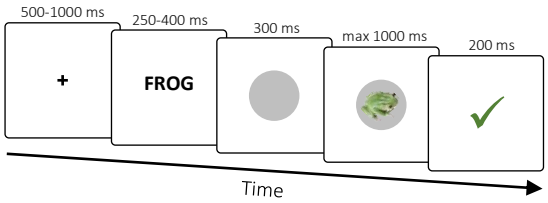
whose values were used for multiple trials when computing the fit to data (e.g., learning rates and inverse temperature parameters; $M = 0.67, SEM = 0.04$).

Additional algorithms. To test whether the algorithms described above were most suitable for describing subjects' behaviour, we compared them to several additional algorithms, all of which failed to fit subjects' choices as well as the above counterparts, and so we do not describe them in detail. These alternative algorithms included a MF algorithm that only learns single-move Q values, but employs temporal difference learning[50] to backpropagate second outcomes in 2-move trials back to the Q values of the starting location (BIC = 41559); a MB algorithm that employs Bayesian inference with a uniform Dirichlet prior[26] to learn the multinomial distributions that compose the state transition matrix (BIC = 43301); a MF-MB hybrid algorithm where state-transition expectations are only used to account for prospective second-move Q values when choosing the first move in 2-move trials (BIC = 40920); and an algorithm that combines two MF algorithms with different parameters (BIC = 40715).

MEG acquisition. MEG was recorded continuously at 600 samples/second using a whole-head 275-channel axial gradiometer system (CTF Omega, VSM MedTech, Canada), while subjects sat upright inside the scanner. A projector displayed the task on a screen ~80 cm in front of the subject. Subjects made responses by pressing a button box, using their left hand for 'left' and 'up' choices and their right hand for 'right' and 'down' choices. Pupil size and eye gaze were recorded at 250 Hz using a desktop-mounted EyeLink II eyetracker (SR Research).

MEG preprocessing. Preprocessing was performed using the Fieldtrip toolbox51 in MATLAB (MathWorks). Data from two sensors were not recorded due to a high level of noise detected in routine testing. Data were first manually inspected for jump artefacts. Then, independent component analysis was used to remove components that corresponded to eye blinks, eye movement and heart beats. Based on previous experience[25], we expected stimuli to be represented in low frequency fluctuations of the MEG signal. Therefore, to remove fast muscle artefacts and slow movement artefacts, we low-pass filtered the data with a 20 Hz cutoff frequency using a sixth-order Butterworth IIR filter, and we baseline-corrected each trial's data by subtracting the mean signal recorded during the 400 ms preceding trial onset. Trials in which the average standard deviation of the signal across channels was at least 3 times greater than median were excluded from analysis (0.4% of trials, SEM 0.2%). Finally, the data were resampled from 600 Hz to 100 Hz to conserve processing time and improve signal to noise ratio. Therefore, data samples used for analysis were length 273 vectors spaced every 10 ms.

Pre-task stimulus exposure. To allow decoding of images from MEG we instructed subjects to identify each of the images in turn (**Supplementary Fig. 6**). On each trial, the target image was indicated textually (e.g., 'FACE') and then an image appeared on the screen. Subjects' task was to report whether the image matched (LEFT button) or did not match (RIGHT button) the preceding text. 20% of presented images did not match the text. The task continued until subjects correctly identified each of the images at least 25 times. Subjects were highly accurate on both match (M = 97.2%, SEM = 0.4%) and no-match (M = 90.2%, SEM = 0.6%) trials. To ensure robust decoding from MEG, we chose eight images that differed in colour, shape, texture and semantic category[43,44] (**Fig. 1a**). Importantly, at this point subjects had no knowledge as to what the main task would involve, nor that the images would be associated with state-space locations and rewards. This ensured that no task information could be represented in the MEG data at this stage.



Supplementary Fig. 6. Pre-task stimulus exposure. Timeline of a trial.

MEG decoding. We used support vector machines (SVMs) to decode images and moves from MEG. All decoders were trained on MEG data recorded outside of the main state-space task and validated within the task. As in previous work[25], we trained a separate decoder for each time bin between 150 and 600 ms following the relevant event, either image onset or move choice, resulting in 46 decoders whose output was averaged. Averaging over decoders trained at different time points reduces peak decodability following stimulus onset, but can increase decodability of stimuli that are being processed when not on the screen[25]. To avoid over-fitting, training and testing were performed on separate sets of trials following a 5-fold cross validation scheme. These analyses were performed using LIBSVM's implementation of the C-SVC algorithm with radial basis functions[52]. Decoder training and testing were performed with each of 16 combinations of the algorithms' cost parameter (10^{-1}, 10^{0}, 10^{1}, 10^{2}) and basis-function concentration parameter ($10^{-2}/n$, $10^{-1}/n$, $10^{0}/n$, $10^{1}/n$), where n is the number of MEG features (273 channels). Where classes differed in number of instances, weighting was used to ensure classes were equally weighted.

To decode the probability of each of eight possible images being presented (8-way classification), we used MEG data recorded during pre-task stimulus exposure. Decoding was evaluated based on the mean probability the decoders assigned to the presented image. To decode the probability of each of the four possible moves (LEFT, RIGHT, UP, DOWN) being chosen (4-way classification), we used MEG data recorded during the image-reward training. For both types of decoder, the parameter combination of cost = 10^{2} and concentration = $10^{-2}/n$ yielded the best cross-validated decoding performance and was thus used for all ensuing analyses.

Sequenceness measure. To investigate how representations of different images related to one another in time, we used a measure recently developed for detecting sequences of representations in MEG[10]. 'Sequenceness' is computed as the difference between the cross-correlation of two images' decodability time-series with positive and negative time lags. By relying on asymmetries in the cross-correlation function, this measure detects sequential relationships even between closely correlated (or anti-correlated) time series, as we have previously demonstrated on simulated time series[25]. Positive values indicate that changes in the first time series are followed by similar changes in the second time series ('forward sequenceness'), negative values indicate the reverse sequence ('backward sequenceness'), and zero indicates no sequential relationship. As in previous work, cross correlations were computed between the z-scored time series over 400 ms sliding windows with time lags of up to 200 ms. This timescale is sufficient for capturing the relationship between successive alpha cycles, which is important given the possibility that such oscillations may reflect temporal quanta of information processing[53].

Bayesian hierarchical Gaussian Process time series analysis. To determine whether sequenceness time-series recorded following outcomes provided robust evidence of replay that correlated with individual index of flexibility, we modelled each mean sequenceness time-series as a summation of two zero-mean Gaussian Processes with squared exponential kernels: a group-level process and an individual-level process. The group-level process

identifies the timepoints in which sequenceness systematically deviates from zeros, and the individual-level processes (one for each time series) account for deviations of individual time series from the group-level process.

To enable completion of the MCMC sampling within a reasonable timeframe, we reduced the trial-to-trial sequenceness data to four mean time series per subject: sequenceness encoding the last or penultimate transition following highly or weakly surprising outcomes. High and low surprise were determined based on the state prediction error generated by the hybrid algorithm, whose parameters were fitted to the individual subject's choices (i.e., high – above-mean prediction error, low – below-mean prediction error). Since we assumed last and penultimate transitions could be replayed in different timepoints, these two types of time series each had their own group-level Gaussian Process. To account for the factors of IF and surprise, the group-level process was multiplied for each time series by a weighted linear combination of the two factors, their interaction, and an intercept (thus involving four parameters: $\beta, \beta^{\text{subject}}, \beta^{\text{surprise}}, \beta^{\text{interaction}}$). The two types of Gaussian Process were parameterized by different length-scales (ρ^{group}, $\rho^{\text{inidivdual}}$) and marginal standard deviations (α^{group}, $\alpha^{\text{individual}}$), and an a standard deviation parameter (σ) accounted for additional normally distributed noise across all observations.

Bayesian estimation was performed in R[54] using the STAN55 package for Markov Chain Monte Carlo (MCMC) sampling. Prior distributions were set so as to be weakly informative and have broad range on the scale of the variables[29]. Thus, β coefficients were drawn from normal distributions with a mean of zero and a standard deviation of 10. All predictor variables were standardized. Standard deviations parameters (α, σ) were drawn from a truncated normal distribution limited to positive values, with a mean of zero and a standard deviation that matches the standard deviation of the predicted variable. Length-scales (ρ) were drawn from log-normal distributions whose mean is the geometric mean of two extremes: the distance in time between two successive timepoints, and the distance in time between the first and last timepoints. Half of the difference between these two values was used as the standard deviation of the priors. $\beta^{\text{interaction}}$ was limited to positive values for the sake of identifiability, since the group-level Gaussian Processes were multiplied by the β coefficients.

We ran six MCMC chains each for 1400 iterations, with the initial 400 samples used for warmup. STAN's default settings were used for all other settings. Examining the results showed there were no divergent transitions, and all parameters were estimated with effective sample sizes larger than 1000 and shrink factors smaller than 1.1. Posterior predictive checks showed good correspondence between the real and generated data (**Fig. 2b,c**).

Decodability time series analyses. Decodability was tested for difference from zero and covariance with individual flexibility using the Bayesian Gaussian Process approach outlined above with the exclusion of the surprise predictor, which is inapplicable to timepoints that precede outcome onset.

Other statistical Methods. Significance tests were conducted using nonparameteric methods that do not assume specific distributions. Differences from zero were tested using 10000 samples of bias-corrected and accelerated Bootstrap with default MATLAB settings. Correlations and differences between groups were tested by comparison to null distributions generated by 10000 permutations of the pairing between the two variables of interest. All tests are two-tailed.

Data and Code availability

The data and custom code used in this study have been deposited on the Open Science Framework under DOI 10.17605/OSF.IO/GUHJE.

Acknowledgements

We thank Zeb Kurth-Nelson and Yunzhe Liu for helpful comments on a previous version of the manuscript. E.E. holds an Alon Fellowship from the Israeli Council for Higher Education. P.D. is funded by the the Max Planck Society. R.J.D. holds a Wellcome Trust Investigator award (098362/Z/12/Z). The Max Planck UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by the Max Planck Society and University College London. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (091593/Z/10/Z).

Competing interests

The authors declare that they have no conflict of interest.

References

1. Olafsdottir, H. F., Carpenter, F., & Barry, C. (2017). Task demands predict a dynamic switch in the content of awake hippocampal replay. *Neuron*, *96*(4), 925-935.
2. Pezzulo, G., van der Meer, M. A., Lansink, C. S., & Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in cognitive sciences*, *18*(12), 647-657.
3. Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241.
4. Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680.
5. Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145-156.
6. Skaggs, W. E., & McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, *271*(5257), 1870-1873.
7. Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience*, *10*(1), 100.
8. Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695-705.
9. Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife*, *4*, e06063.
10. Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. *Neuron*, *91*(1), 194-204.
11. Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181.
12. Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*.
13. Foster, D. J. (2017). Replay comes of age. *Annual review of neuroscience*, *40*, 581-602.
14. Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490-509.
15. Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, 20(11), 1643.
16. Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife*, 7, e32548.
17. Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609.
18. Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, *2*(4), 160-163.
19. Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182.

20. Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*(13), 6035-6044.

21. Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

22. Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature neuroscience*, *8*(11), 1481.

23. Gillan, C. M., Fineberg, N. A., & Robbins, T. W. (2017). A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychological medicine*, *47*(9), 1528-1548.

24. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204-1215 (2011).

25. Eldar, E., Bae, G. J., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2018). Magnetoencephalography decoding reveals structural differences within integrative decision processes. *Nature Human Behaviour*, *2*(9), 670.

26. Bishop. C.M. (2006) Pattern Recognition and Machine Learning (Springer).

27. Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74.

28. Kurth-Nelson, Z., Barnes, G., Sejdinovic, D., Dolan, R. & Dayan, P. Temporal structure in associative retrieval. *Elife* 4, e04919 (2015).

29. Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

30. Moore, A. W., Atkeson, C. G. (1993) Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13, 103–130.

31. Peng, J., & Williams, R. J. (1993) Efficient learning and planning within the Dyna framework. *IEEE International Conference on Neural Networks* 168–174. DOI: https://doi.org/10.1109/ICNN.1993.298551.

32. Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

33. Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences* 23, 645-665.

34. Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585-595.

35. Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science* 27, 848-858.

36. Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience* 15, 523-536.

37. da Silva, C. F., & Hare, T. (2019). Model-free or muddled models in the two-stage task? *bioRxiv* 682922.

38. Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology* 12, e1005090.

39. Akam, T., Rodrigues-Vaz, I., Zhang, X., Pereira, M., Oliveira, R., Dayan, P., & Costa, R. M. (2017). Single-Trial Inhibition of Anterior Cingulate Disrupts Model-based Reinforcement Learning in a Two-step Decision Task. *bioRxiv* 126292.

40. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, *13*(9), e1005768.

41. Carey, A. A., Tanaka, Y., & Van Der Meer, M. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nature Neuroscience* 22, 1450-1459 (2019).

42. Hunt, L. T. et al. Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci*. 15, 470–476 (2012)

43. Carlson, T., Tovar, D. A., Alink, A. & Kriegeskorte, N. Representational dynamics of object vision: the first 1000 ms. *J. vis.* 13, 1 (2013)

44. Isik, L., Meyers, E. M., Leibo, J. Z. & Poggio, T. The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* 111, 91–102 (2014)

45. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462 (2014).

46. Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT press.

47. Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704.

48. Huys, Q.J.M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J.P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comp. Biol. *8*, e1002410.

49. Eldar, E., Hauser, T.U., Dayan, P., and Dolan, R.J. (2016) Striatal structure and function predict individual biases in learning to avoid pain. Proc. Natl. Acad. Sci. USA 113, 4812–4817.

50. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron* 38, 329-337.

51. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J. M. (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. intel. Neurosci. 2011*, 156869.

52. Chang, C.C., and Lin, C.J. (2011). LIBSVM: a library for support vector machines. ACM T. Intel. Syst. Tec. *2*, 27.

53. Busch, N. & VanRullen, R. Is visual perception like a continuous flow or a series of snapshots. In: Arstila, V. & Lloyd. D. (Eds.) *Subjective time: The philosophy, psychology, and neuroscience of temporality* (MIT Press, 2014)

54. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

55. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software 76(1).