ARTICLES
DOI: 10.1038/s41562-017-0180-8

nature human behaviour

The successor representation in human reinforcement learning

I. Momennejad[1]*, E. M. Russek[2], J. H. Cheong[3], M. M. Botvinick[4], N. D. Daw[1] and S. J. Gershman[5]

Theories of reward learning in neuroscience have focused on two families of algorithms thought to capture deliberative versus habitual choice. 'Model-based' algorithms compute the value of candidate actions from scratch, whereas 'model-free' algorithms make choice more efficient but less flexible by storing pre-computed action values. We examine an intermediate algorithmic family, the successor representation, which balances flexibility and efficiency by storing partially computed action values: predictions about future events. These pre-computation strategies differ in how they update their choices following changes in a task. The successor representation's reliance on stored predictions about future states predicts a unique signature of insensitivity to changes in the task's sequence of events, but flexible adjustment following changes to rewards. We provide evidence for such differential sensitivity in two behavioural studies with humans. These results suggest that the successor representation is a computational substrate for semi-flexible choice in humans, introducing a subtler, more cognitive notion of habit.

How do humans and other animals discover adaptive behaviours in dynamic task environments? Identifying such behaviours poses a particular challenge in sequential decision tasks like chess or maze navigation, in which the consequences of an action may unfold gradually, over many subsequent steps and choices. In the past two decades, much attention has focused on a distinction between two families of reinforcement learning algorithms for solving multi-step problems, known as model-free (MF) and model-based (MB) reinforcement learning[1,2]. Although both approaches formalize the problem of choice as comparing the long-term future reward expected following different candidate actions, they differ in the representations and computations they use to estimate these values[3] (Fig. 1). The MF versus MB dichotomy has been influential, in part, because it poses an appealingly clean tradeoff between decision speed and accuracy: MF algorithms store ('cache') pre-computed long-run action values directly, whereas in MB algorithms, state transition learning enables more flexibility at greater computational expense: action values are re-computed using an internal model of one-step state transitions. This tradeoff has been put forward as a computational basis for phenomena relating to automaticity, deliberation and control, and it has been argued that the inflexibility of MF learning in particular explains maladaptive, compulsive behaviours such as drug abuse.

Although experiments suggest that people (and other animals) can flexibly alter their decisions in situations that would defeat fully MF choice, there remains surprisingly little evidence for how, or indeed whether, the brain carries out the sort of full MB re-computation that has typically been invoked to explain these capabilities. Furthermore, there exist other computational and representation learning shortcuts along the spectrum between MF and MB learning, which might suffice to explain many of the available experimental results. For the brain, such shortcuts provide plausible strategies for maximizing fitness; for the theorist, they enrich and complicate the theoretical tradeoffs involved in controlling decisions and managing habits. Here, we report two experiments examining whether humans employ one important class of such shortcuts that lie between the MB and MF strategies. This intermediate algorithm for representation learning, based on the successor representation (SR)[4,5], caches long-range or multi-step state predictions. In the remainder of the introduction we will focus on summarizing MF and MB algorithms and explaining SR-based algorithms in relation to them.

MF strategies such as temporal difference learning cache fully computed long-run action values as decision variables. Caching makes action evaluation at the decision time computationally cheap, as stored action values can be simply retrieved. Action values (Q in Fig. 1) can be estimated and updated by using reward prediction error signals and aggregating the net value over a series of events and rewards unfolding over time. This means MF learners do not store any information about the relationships between different states, and hence fail to solve problems involving distal changes in the reward value[6]. This has been empirically demonstrated by 'reward revaluation' studies[7] and latent learning[8].

In contrast, MB algorithms do not rely on cached value functions. Instead, they store a full model of the world and compute trajectories at the decision time. Specifically, they learn and store a one-step internal representation or model of the short-term environmental dynamics: specifically, a state transition function T and a reward function R (T and R in Fig. 1). By iterative computation using this one-step model, action values can be computed at the decision time. This is analogous to mental simulation using a 'cognitive map', stringing together the series of outcomes expected to follow each action according to learned representations. This planning capacity endows MB algorithms with sensitivity to distal changes in the reward (as in reward revaluation) and also to changes in the transition structure (such as detour problems in spatial tasks). This flexibility comes at a higher computational cost compared with caching: computations traversing a model are intensive in time and working memory. Such computations may be intractable (requiring error-prone approximations) in large search spaces such as wide and deep trees[9,10].

[1]Department of Psychology, Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Center for Neural Science, New York University, New York, NY, USA. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [4]DeepMind and Gatsby Computational Neuroscience Unit, University College London, London, UK. [5]Department of Psychology, Center for Brain Science, Harvard University, Cambridge, MA, USA. I. Momennejad and E. M. Russek are contributed equally to the work. *e-mail: idam@princeton.edu

The SR was originally introduced as a method for rapid generalization in reinforcement learning[4]. The SR simplifies evaluation via multi-step representation learning: it caches long-term predictions about the states it expects to visit in the future. Namely, for each starting state, the SR caches how often the agent expects or needs to visit each of its successor states in the future (which can be learned via simple temporal difference learning[5]). When the agent faces a decision, the SR is combined with the reward function (R) to evaluate the optimal trajectory to reward: it combines how often successor states are expected to be visited on average in the future with their reward.

Mathematically, the SR is a matrix M, where the i-th row is a vector in which element $M_{i,j}$ stores the expected discounted future occupancy of state j following initial state i. To understand what this means, imagine starting a trajectory in state i and counting the number of times each state j is encountered subsequently, while exponentially discounting visits that occur farther in the future. This representation is useful because at the decision time, action values can be computed by linearly combining the SR for the current state with the one-step reward function. This obviates the MB strategy's laborious iterative simulation of future state trajectories using MB's one-step model, but stops short of storing the fully computed decision variable, as MF learning does. Thus, action evaluation with the SR has similar computational complexity to MF algorithms, while at the same time retaining some of the flexibility characteristic of the MB strategy. This form of predictive caching, if it exists in the brain, would provide a compromise between fully flexible deliberation and complete automaticity, allowing choices to be adjusted nimbly in some circumstances but still producing inappropriate, habit-like behaviour in others. Such a representation learning strategy is particularly well suited to environments in which the trajectories of states are fairly reliable, but rewards and goals change frequently. In such 'multi-goal' environments, as they are termed by the reinforcement learning literature, a compromise between MF and MB strategies becomes an appealing algorithm. The evidence that people and animals can solve (at least small and simple) reward revaluation tasks despite the computational complexity of MB algorithms lends further support to the validity of a more cost-efficient algorithm.

Several lines of evidence motivate consideration of the SR as a hypothesis for biological reinforcement learning. First, converging evidence from other domains suggests that the SR is explicitly represented in the brain: the SR defined over space can capture many properties of rodent hippocampal place cells[11], whereas in tasks with a more abstract sequential stimulus structure, the SR captures properties of functional magnetic resonance imaging pattern similarity in the hippocampal and prefrontal areas[12,13]. The SR is also closely related to the representation posited by the temporal context model of memory[5], which successfully explains numerous memory effects. Second, SR learning can be implemented using a version of the predominant neural theory of MF learning—the temporal difference learning theory of the dopamine response[14]. In particular, if the expected future visits stored in M are used as the state input for a temporal difference learner, temporal difference will learn the reward function R (ref. [4]). Together with the MB-like flexibility of the SR, this observation may help to explain several puzzling reports that dopamine affects putative behavioural signatures of not just MF but also MB learning[15,16]. These results are unexpected for standard MB algorithms, which do not share any aspects of their learning with temporal difference methods. Third, most existing empirical evidence in favour of MB algorithms in the brain is equally consistent with an SR-based account. In particular, SR and MB algorithms make identical predictions about reward revaluation experiments—a class that includes reward devaluation, latent learning[8,17], sensory preconditioning[18,19] and two-step Markov decision tasks, which have been widely used with humans[6,15]. The important empirical challenge of discriminating between these accounts was the focus of the present study.

In the present study, we provide new experimental designs that aim to tease apart behaviour using SR and MB computations, and in particular to investigate whether people learn and use representations that cache long-run expectancies about future state occupancy. Although the SR can flexibly adapt to distal changes in reward (as in reward revaluation), it cannot do so with distal changes in the transition structure (what we call transition revaluation). As the SR caches a predictive representation that effectively aggregates over the transition structure, it cannot be flexibly updated in response to changes in this structure, unlike an MB strategy. Instead, the SR can only learn about changes in the transition structure incrementally and through direct experience, much like the way MF algorithms learn about changes in the reward structure. We exploited this difference by comparing the effects of reward and transition revaluation manipulations on human behaviour. MF algorithms predict that participants will be equally insensitive to reward and transition revaluation, whereas MB algorithms predict that participants will be equally sensitive to both (and accordingly any linear combination of the two algorithms will predict equal sensitivity to both conditions). Crucially, any learning strategy that uses the SR (either the SR alone or a hybrid strategy that combines the SR with another strategy) predicts that participants will be more sensitive to reward than transition devaluation (Fig. 2).

To summarize, MF strategies do not store any representations of future states and do not compute state representations at the decision time (Figs. 1 and 2). In contrast, MB strategies store and

	Representation	Computation	Behavior
MF learner	Q: Cached value	Retrieve cached value Lowest cost	Habit, Fast
MB learner	R: Vector of all state rewards T: One-step state transitions matrix	Iteratively compute values Highest cost, resource-constrained	Fully flexible, Slow
SR learner	R: Vector of all state rewards M: Multi-step future state occupancy matrix (policy-dependent caching)	Combine cached future occupancies with rewards Intermediate costs	Semi-flexible, Fast
Hybrid SR	R & M (as above) SR output combined with T, or update SR with replay (on- or offline)	Combine SR with MB or replay Intermediate costs: mostly SR costs, at times MB or replay costs	Flexible but asymmetric, Fast (mostly)

Fig. 1 | Comparison of stored representations, computations at the decision time and behaviour across models. Both the MF cached value and the SR can be learned via simple temporal difference learning during the direct experience of trajectories in the environment. M, the SR or a 'rough' predictive map of each state's successor states; Q, value function (cached action values); R, reward function; T, full single-step transition matrix.

retrieve one-step representations, leading to high computational demand at the decision time. However, the SR caches a predictive map of states that the agent expects to visit in the future. Using these cached representations at the decision time, the SR can solve reward revaluation but not transition revaluation, while the MB strategy is equally successful at all revaluations and the MF strategy is equally unsuccessful. Another possibility is to have a blend of the SR with other strategies, which we refer to as hybrid SR strategies. Hybrid SR strategies could combine the predictive representation with MB computations or replay in order to either update the SR offline in a Dyna-like architecture[20] (which we call SR–Dyna) or augment the SR at the decision time (which we refer to as hybrid SR–MB or SR-replay strategies). As such, all hybrid SR strategies should perform better than a pure SR strategy on transition revaluation (but worse than the MB strategy). Specifically, hybrid SR strategies predict higher accuracy and faster response times for reward revaluation than transition revaluation (an asymmetry in performance that is not predicted by either the MF or MB strategy; see Figs. 1 and 2). Here, we experimentally test and confirm these predictions in two studies, providing direct evidence for the SR in human reinforcement learning.

Results

Experiment 1: differential sensitivity to reward and transition revaluation in a passive learning task. We designed a multi-step sequential learning task to compare human behaviour under reward revaluation and transition revaluation. Experiment 1 used a passive learning task, which permitted the simplest possible test of the theory, removing the need to model action selection. A schematic of the design is displayed in Fig. 3 and Supplementary Fig. 1. Participants played 20 games, each of which was made up of three phases. In phase 1 (the learning phase), participants first learned three-step trajectories leading to reward. These trajectories were deterministic and passively experienced (that is, the transition required no action from the participant). Participants were exposed to one stimulus at a time and were asked to indicate their preference for the middle state after every five stimuli. The learning phase ended if the participant indicated preference for the highest paying trajectory three times, or after 20 stimulus presentations. At the end of the learning phase, they were asked to indicate which starting state they believed led to greater future reward by reporting their relative preference using a continuous scale. Learning was assessed by the participant's preference for the starting state associated with the more rewarding trajectory.

In phase 2 (the re-learning phase), trajectories were initiated at the second or middle state of the trajectory, and the structure of the task was altered in one of two ways (within participants): in the reward revaluation condition, the rewards associated with the terminal states were swapped, whereas in the transition revaluation condition, the transitions between the step 2 and step 3 states were swapped (Figs. 2 and 3). Both conditions induced equivalent changes to the values of the first-stage states. In addition, we included a control condition in which no change occurred during phase 2 (the re-learning phase). As in phase 1, participants were probed for state 2 preferences after each five stimuli, and phase 2 ended if participants had indicated the middle state of the most rewarding trajectory three times or after 20 stimuli (see Methods and Supplementary Fig. 1). Finally, in phase 3 (the test phase), participants were again asked which starting state they preferred. They had 20 s to give a response. Revaluation was measured as the extent of preference change between phases 1 and 3 (Δpreference, signed so that positive values indicated a preference shift towards the newly optimal starting state; Fig. 3 and Supplementary Fig. 1; see Methods for a more detailed explanation of the experimental design).

Figure 4a displays the mean (±1 standard error of the mean (s.e.m.)) revaluation scores for the three conditions (reward revaluation: 0.5199±0.0203; transition revaluation: 0.4503±0.0229; control: 0.0310±0.0310; $n=58$ participants). We conducted a repeated-measures analysis of variance and post hoc t-tests (on participant means) using Bonferroni correction. A one-way within-participants

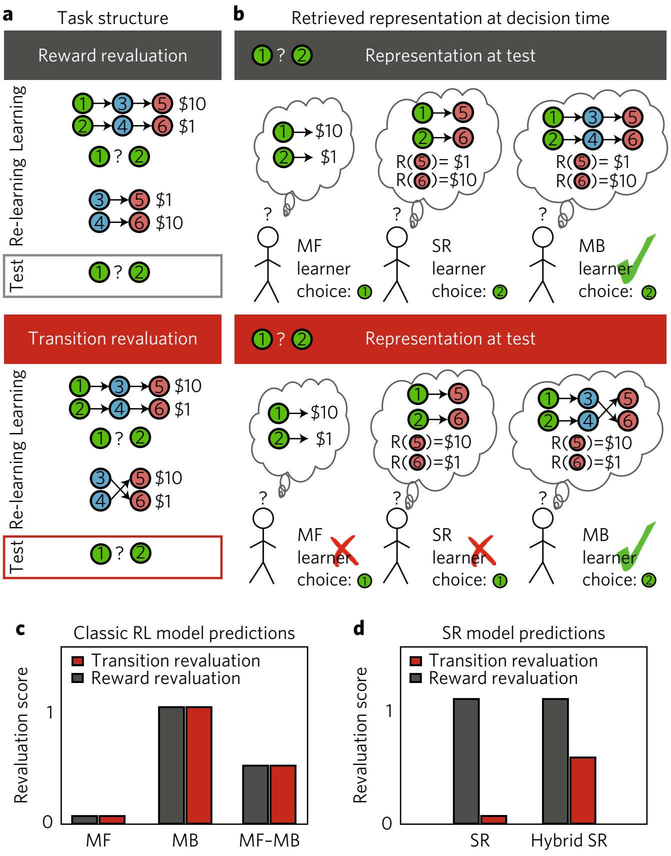

Fig. 2 | Schematic of retrieved representations at test and model predictions in reward and transition revaluation trials. a, Schematics for task structures in reward revaluation (top) and transition revaluation (bottom). **b**, Schematics of the representations retrieved at the decision time by different learners: an MB learner retrieves one-step transitions, and rolls out and computes full transitions (a costly computation), then combines them with the reward vector R to produce a decision. The MB learner is equally successful in both reward and transition revaluation. A purely SR learner retrieves the SR (of which only the relevant rows are displayed here) without further computation and readily combines it with the reward vector. **c**, Qualitative model predictions for reward (grey) and transition (red) revaluation. Predicted revaluation scores for MF, MB, MB–MF, purely SR and a hybrid SR learner (SR–MB or SR-replay). Classic reinforcement learning (RL) solutions all predict symmetrical responses to the retrospective revaluation problems. That is, while the MF strategy has no solution to reward or transition revaluations, MB and hybrid MF–MB learners predict symmetrical performance for all types of revaluation. **d**, SR strategies predict asymmetrical responses: the SR algorithm is sensitive to changes in reward. However, since SR stores a multi-step predictive map *M*, and not the step-by-step transition structure, it cannot update *M* in the absence of direct experience. That is, the SR effectively 'compiles' the transition structure into an aggregate predictive representation of future states and therefore cannot adapt to local changes in the transition structure in the environment without experiencing the new trajectories in full. While a pure SR algorithm cannot solve transition revaluation, a hybrid SR learner who is updated via simulated experience (for example, via MB representations or episodic replay) adjusts their decision for any revaluation, but performs best in reward revaluation.

analysis of variance revealed a significant effect of condition on the revaluation scores ($F_{2,\,1,154} = 38.77$, $P < 0.001$). As our experiment was designed with clear a priori hypotheses, we performed planned comparisons using paired sample *t*-tests with Bonferroni correction, which indicated that the mean revaluation score was significantly different in the reward revaluation condition than the transition revaluation condition ($t_{57} = 2.8928$, $P = 0.016$). In addition, the mean revaluation scores were higher for both the reward revaluation condition ($t_{57} = 10.148$, $P < 0.001$) and the transition revaluation condition ($t_{57} = 9.0543$, $P < 0.001$) compared with the control condition. In the control condition, as expected, revaluation scores were not significantly different from zero ($t_{57} = 1.2$, $P = 0.22$). This finding is important because it verifies that baseline forgetting or randomness cannot explain participants' behaviour in the non-control conditions.

We further analysed the data for any time-on-task effects on accuracy or differences in accuracy (that is, whether behaviour improved as a result of practice and whether these changes were significantly different in transition versus reward revaluation conditions). For the non-control trials, there was a significant effect of time on task (trial number) on the revaluation score ($F_{1,\,57.259} = 9.9171$, $P < 0.01$), indicating that participants' ability to perform the task improved over time. However, there was no significant interaction of this effect with revaluation condition ($F_{1,\,68.284} = 0.15436$, $P = 0.695$). Finally, we also found a significant main effect of revaluation condition on response times during the test phase ($F_{2,\,171} = 7.74$, $P < 0.001$; Fig. 4b). In particular, response times were slower in the transition revaluation condition compared with both the reward revaluation condition ($t_{57} = 2.08$, $P < 0.05$) and the control condition ($t_{57} = 4.04$, $P < 0.001$), and response times in the reward revaluation condition were significantly slower compared with the control condition where no changes had occurred ($t_{57} = 3.5646$, $P < 0.001$). Substantial individual differences in behaviour under different revaluation conditions were observed, along with the appearance of multi-modality (Fig. 4c). Previous research has suggested that the balance between MB and MF learning tracks important individual differences, such as symptoms of mental illness[21]. Future work exploring individual differences in subtler forms of the computation and representation presented here may provide valuable insight into the arbitration of different representation and control processes across populations, as well as their effects on the flexibility and pathologies of decision-making.

A hybrid SR model explains differential sensitivity to varieties of revaluation. The key signature of the SR's caching of multi-step future state occupancies (that is, caching how often the agent expects or needs to visit a successor state in the future) is differential sensitivity to reward versus transition revaluations. Participants' differential sensitivity to these manipulations argues against a pure MB or MF account (see Methods for a detailed description of all the models considered here). MF algorithms predict equivalent and total insensitivity to both revaluation conditions because participants are never given the opportunity to re-experience the start state following the revaluation phase. This effectively fools algorithms like temporal difference learning that rely on the chaining of trajectories of direct experience to incrementally update cached value estimates. In contrast, MB algorithms predict equal sensitivity to both conditions (Fig. 5a) so long as the revalued contingencies are themselves learned, because the updated internal model following the revaluation phase produces accurate action values for the start state in either case. Accordingly, any weighted combination of these two evaluation mechanisms—which is the hybrid reinforcement learning model often used to explain previous sequential decision tasks[15]—also does not predict differential sensitivity. This is because the combination simply scales the equal sensitivity of either algorithm up or down.

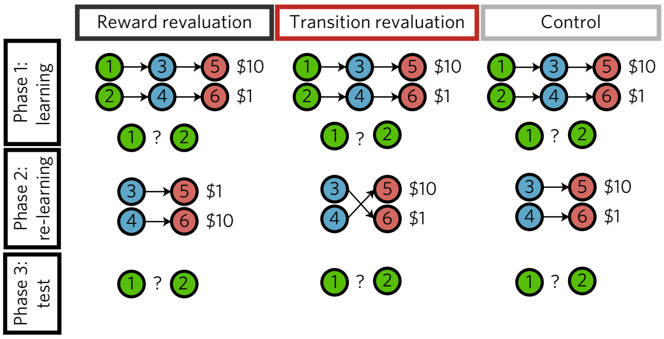

Fig. 3 | Schematic of the design of experiment 1. This schematic represents the structure of states, rewards and revaluation conditions. Participants never saw these graphs and experienced the task structure one stimulus at a time (as displayed in Supplementary Fig. 1). The experiment consisted of three phases: learning, re-learning/revaluation and test in extinction. At the end of phase 1 (learning) and phase 3 (test), participants provided a continuous valued rating indicating which of the two starting states they preferred. We computed a quantity, the revaluation score, by taking the difference between these two ratings.

SR-based algorithms fare better (Fig. 5b) in that they predict that an agent will be insensitive to transition revaluation but sensitive to reward revaluation. In particular, algorithms that update a cached estimate of the SR using a temporal difference-like learning rule require full trajectories through the state space in order to update the start state's SR (that is, the future state occupancies it predicts) following the revaluation phase. This mirrors the direct experience requirement of MF algorithms for value estimation. However, unlike MF algorithms, SR-based algorithms can instantly adapt to changes in reward structure because this only requires updating the immediate reward prediction, which then propagates through the entire state space when combined with the SR.

We did not hypothesize, nor do our results suggest, total reliance on an SR strategy; instead, we sought to investigate whether such a strategy is used at all by humans. A pure SR account does not by itself explain our data, because it predicts complete insensitivity to transition revaluation. In contrast, we see significantly greater revaluation in the transition revaluation condition compared with the control condition. This can be understood in terms of a hybrid SR–MB account analogous to the MB–MF hybrids[13,21] considered previously (Fig. 5c). Although (as we discuss later) there are several ways to realize such a hybrid, for simplicity we chose to linearly combine the ratings from MB and SR algorithms. This linear combination allows the hybrid model to show partial sensitivity to transition revaluation. The hybrid model may also provide insight into the response time differences; under the assumption that effortful MB re-computation is invoked preferentially following transition revaluation (when it is, in fact, most needed), this condition would slow the response time, consistent with our findings. Crucially, previous MB–MF hybrids[13,21] cannot explain the asymmetry between transition and reward revaluation conditions (Fig. 2c); among the theories we considered, this asymmetry is uniquely explained by a hybrid SR–MB account.

Experiment 2: differential sensitivity to revaluation types in a sequential decision task. In a second experiment, we sought to replicate and extend our results in two ways. First, experiment 1 used a passive learning task, in which participants were exposed to a sequence of images, and the dependent measure was the relative preference rating between different starting states. This is similar to previous Pavlovian experiments such as sensory preconditioning[22]. Since a key purpose of state evaluation is guiding action choice,

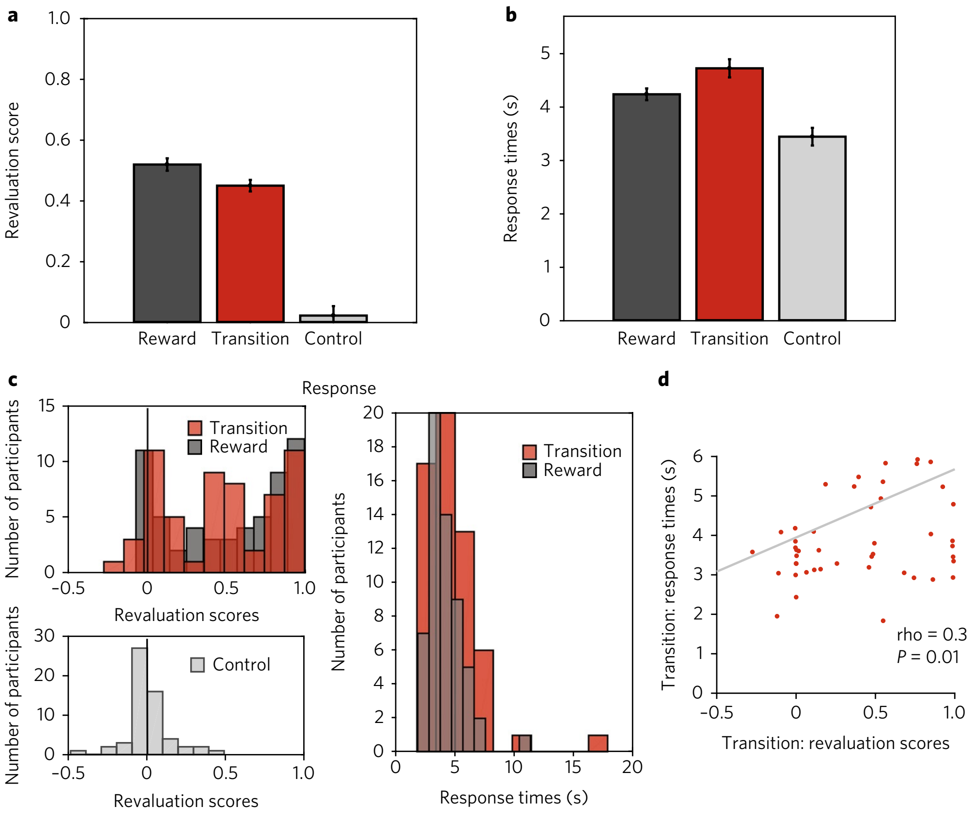

Fig. 4 | Behavioural performance in a passive sequential learning task. Human performance was measured as the change in preference ratings for the starting states. Revaluation scores for each game denote the change in a given participant's relative preference rating after versus before the re-learning phase. **a**, Mean revaluation scores are plotted for the three conditions (reward revaluation, transition revaluation and control games). There was a significant main effect of condition. **b**, Mean response times to the final preference decision in phase 3 under reward and transition revaluation. Behavioural responses to the preference rating were significantly slower during transition revaluation. The error bars indicate s.e.m. **c**, Histograms reveal the distribution of revaluation scores and response times across the main conditions. **d**, There was a significant correlation between the accuracy of transition revaluation responses and response times: more accurate transition revaluation took longer, suggesting that successful transition revaluation might have involved more computation at the decision time. Together with significantly faster response times compared with reward revaluation (**b**), this positive correlation lends further evidence to the possibility that, compared with reward revaluation, transition revaluation required more cycles of computation at the decision time, relying less on cached representations.

we sought to examine the same questions in terms of decisions in a multi-step instrumental task. This framing also allowed us to include an additional condition, which we call 'policy revaluation'. The SR-based algorithms predict the same patterns of behaviour for this condition as they do for transition revaluation, but the actual sequence of participants' experiences is much more closely matched to the reward revaluation condition. In particular, this condition turns on a change in reward amounts rather than state transition contingencies during phase 2 re-learning; no changes in the transition function occur in policy revaluation.

Participants completed four games, each of which corresponded to a different experimental condition (Fig. 6). In each trial of each game, participants navigated through a three-stage decision tree (represented as rooms in a castle; see Methods for experiment details; Supplementary Fig. 2). From the first stage (state 1), participants made a choice that took them deterministically to one of two second-stage states (states 2 and 3). Each second-stage state contained two available actions (and one unavailable action), and each action led deterministically to one of three reward-containing terminal states.

As in the previous experiment, these trials were grouped into 3 phases for each game (Fig. 6). In phase 1 (the learning phase), participants were trained on a specific reward and transition structure. If, for any condition, participants failed to perform the correct action from each non-terminal state on three of their last four visits to that state during phase 1, they were removed from the analysis. In phase 2 (the re-learning phase where revaluation could happen), a change in either the reward structure or the set of available actions occurred (the latter causing a change in the state–action–state transition function). Participants learned about the changed structure in nine trials, such that they were exposed to the change at least three times. Importantly, as in experiment 1, participants did not revisit the starting state in phase 2, and hence never experienced any of the new contingencies following an action taken from the starting state. In phase 3, participants performed a single test trial beginning from the starting state. For each condition, we defined the revaluation score as a binary variable indicating whether a participant switched their action in state 1 between the end of phase 1 and the single probe trial in phase 3.

Our results replicate those from experiment 1, extending them to a new policy revaluation condition. The proportion of changed choices in the phase 3 test, by condition, is shown in Fig. 7 (mean ± s.e.m.: reward revaluation: 0.6591 ± 0.0505; transition revaluation: 0.4659 ± 0.0532; policy revaluation: 0.5000 ± 0.0533; control: 0.0795 ± 0.0288). Logistic regression verified that more participants successfully switched their stage 1 action choice following the reward revaluation than the transition revaluation (contrast estimate: −0.7958, Wald test Z-score (Z) = −2.85, P = 0.0034) and

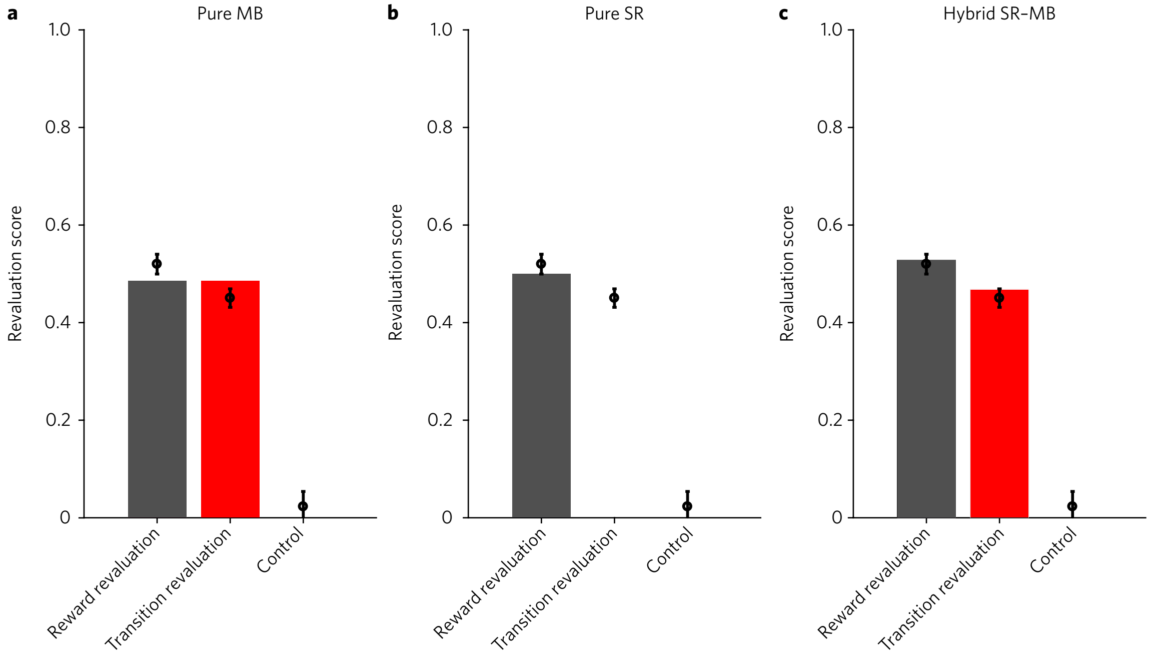

Fig. 5 | Model fits to the phase 3 test data from the passive learning task. a–c, We compared model performance (solid bars) against human data (error bars) using a pure MB learner (**a**), a pure SR learner (**b**) and a hybrid SR–MB learner (**c**). Human behaviour is best explained by the hybrid account.

also the policy revaluation (contrast estimate: −0.6592, $Z=-2.61$, $P=0.0043$). In contrast, there was no significant difference between the proportions of participants who changed preference following policy revaluation compared with transition revaluation conditions (contrast estimate: 0.14, $Z=0.56$, $P=0.5771$). All three of the revaluation manipulations produced more switching than the control condition with no revaluation (reward > control contrast estimate: 3.1078, $Z=6.95$, $P<0.0001$; transition > control contrast estimate: 2.31, $Z=5.11$, $P<0.0001$; policy > control contrast estimate: 2.45, $Z=5.20$, $P<0.0001$), verifying that these results were due to a shift in preferences rather than non-specific effects such as forgetting. There was no significant effect of time on task (trial number) on revaluation score ($F_{1,\ 190.6}$, $P=0.076$). There was also no significant interaction of time on task with revaluation condition ($F_{2,\ 69.49}$, $P=0.367$).

A hybrid model explains lower sensitivity to policy revaluation. The logic of the policy revaluation (Fig. 6) is that the introduction of a large new reward at state 4 in the re-learning phase should cause a change in the preferred action at state 2. The effect of this is to change which terminal state can be expected to follow the top-stage action that leads to state 2. Like transition revaluation, this manipulation should produce a change in top-stage preferences due to a change in the terminal state transition expectancies, but crucially it does so due to learning about reward amounts rather than the actual transition links in the graph. As the SR caches predictions about which terminal state follows either state 1 action, it cannot update its decision policy without experiencing the newly preferred state along a trajectory initiated by the state 1 action leading to state 2. The MB and MF models (and the various hybrids) also treat this condition the same as the transition revaluation: in particular, the MB model should correctly re-compute the new stage 1 action choice given learning about the new reward, whereas the MF model's stage 1 preferences should be blind to the change.

The similarity in performance on transition and policy revaluations suggests that the differences we observed, in both experiments, between transition and reward revaluation cannot be due to different learning rates for adapting to transition and reward changes. That is, the difference between transition and reward revaluation cannot be explained by an MB learner with a much slower learning rate for the transition matrix T than the learning rate R. This is because the policy revaluation fools the SR in the same way as transition revaluation, but it does so by requiring participants to learn about a change in rewards rather than transitions. If participants were using a pure MB algorithm but were differentially skilled at transition and reward learning, we would expect policy revaluation to look more like reward revaluation than transition revaluation, which was not the case.

We developed action-based variants of the models described in the previous section and fit them to the behavioural data (see Methods for details). Consistent with the results from the passive learning task, only the hybrid SR–MB model was able to adequately capture the pattern of differential sensitivity across conditions (Fig. 8).

Testing alternative possibilities. *Possibility 1: difference in MF–MB arbitration between conditions.* It has been suggested that the relative balance between 'state prediction errors' and 'reward prediction errors' may be used for arbitration in an MB–MF hybrid learner[23]. In the policy revaluation condition, participants experience greater reward prediction error than in the reward revaluation condition, and thus the arbitration account predicts that participants should use more MB strategies and thus be more successful at revaluation on policy revaluation trials compared with reward revaluation trials. This is the opposite of the SR prediction, and was contradicted by our experimental observations.

To examine strategy changes further, we analysed whether participants changed strategy over the course of the experiment. Because only test trials can be used to ascertain strategy, and there was only one test trial per block, we could not determine whether participants' strategies changed within the course of a single block. If, however, participants' strategies changed between blocks, over the course of the task, we would expect their revaluation performance to change as well. We therefore used repeated-measures analysis of variance to investigate whether there was an effect of accumulated

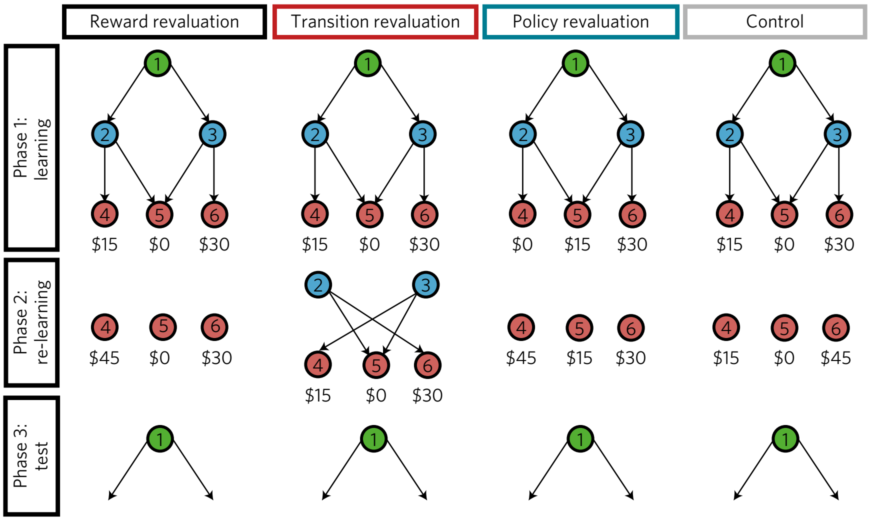

Fig. 6 | Schematic of the active sequential learning task. The underlying structure of each condition in experiment 2 is represented. Numbered circles denote different states (rooms in a castle), and arrows denote unidirectional actions available upon entering that state and the deterministic transition associated with those actions (that flow always from states with lower numbers to higher numbers; top to bottom in the schematic). Unavailable actions in states 2 and 3 are not shown. On each trial, participants were placed in one of the six states (castle rooms) and were required to make choices between upcoming states until they arrived at a terminal state and collected its reward. For a given phase of a given condition, trials began only in the states that are displayed in the figure for that condition and phase. For example, trials in phase 2 of the reward revaluation condition began only from states 4, 5 and 6. In all conditions, state 1 contained two actions, both of which were always available. States 2 and 3 each contained three actions; however, at any given time only two were available. Upon arriving in either state 2 or state 3, the participant observed which actions were available and which were unavailable. For each condition, we measured whether participants changed their state 1 action choice between the end of phase 1 and the single probe trial in phase 3 from the action leading to state 3 to the action leading to state 2.

time on the task (trial number) on the revaluation score. Because a change in model choice would only affect revaluation scores of non-control trials, we eliminated control trials from this analysis. For experiment 1, this analysis revealed a significant effect of time on task on revaluation score ($F_{1,\,57.259}=9.9171$, $P<0.01$) indicating that participants' ability to perform the task improved over time. If such a change in strategy over the task were responsible for the difference in revaluation score between reward and transition revaluation conditions, we would expect the effect of trial number on revaluation score to interact with the revaluation condition. However, there was no significant interaction of this effect with revaluation condition ($F_{1,\,68.284}=0.15436$, $P=0.695$). For experiment 2, there was no significant effect of time on task (trial number) on revaluation score ($F_{1,\,190.6}$, $P=0.076$). There was also no significant interaction of time on task with revaluation condition ($F_{2,\,69.49}$, $P=0.367$). Thus, time on task cannot explain the difference between the conditions.

Possibility 2: differences in learning and updating *T*, *M* and *R*. For simplicity, we assumed that during the re-learning phase, the MB learner fully updates the experienced transitions in the transition matrix (T) and the SR learner fully updates the successor matrix M. However, performance on reward revaluation is not perfect. Could this be due to different learning rates for R and T? The re-learning phase of experiment 1 included a probe trial once every five trials in which participants were required to choose between the second-level state in either sequence. Participants were unable to progress to the test phase until they chose the correct (highest value) state three times in a row. A repeated-measures analysis of variance was conducted to investigate the effect of revaluation condition (limited to reward revaluation trials and transition revaluation trials) on the number of trials required to meet the criterion for moving to the test phase. There was no significant effect of revaluation condition ($F_{1,\,924}=0.549$, $P=0.359$). Thus, we found no evidence for learning rate differences between conditions. Furthermore, the results from the policy revaluation condition in experiment 2 show that the findings are not merely limited to a difference between reward versus transition learning, since in policy revaluation there are no changes in the transition probabilities (not updating T or SR) and yet the behaviour is more similar to transition revaluation rather than reward revaluation. Taken together, these findings are consistent with the idea that the differences between conditions are not merely due to differences in reward learning versus transition learning.

Discussion

The brain must trade off the computational costs of solving complex, dynamic decision tasks against the costs of making suboptimal decisions due to employing computational shortcuts. It has, accordingly, been argued that compared with MB solutions, simple MF learning saves time and computation at the decision time at the cost of occasionally producing maladaptive choices in particular circumstances, such as rats working for devalued food. Here, we consider a third strategy based on the SR, which is noteworthy for two reasons. First, the SR caches temporal abstractions of future states. At the decision time, while MB relies on forward search to evaluate actions, the SR simply retrieves cached representations of successor states and produces rapid, flexible behaviours, which in many circumstances were previously taken as signatures of the more costly MB deliberation. Second, the SR predicts (and our experiments confirmed) a novel asymmetric pattern of errors across different types of revaluation task. While MB performs equally well on all revaluation tests and MF solves none, the SR can use its cached representations to readily solve reward revaluation, but not transition or policy revaluation.

Previously, revaluation tasks—mostly reward revaluation—have been useful in distinguishing MB from MF predictions. However, MB and SR-based algorithms make similar predictions for standard reward revaluation tasks, which account for the bulk of evidence previously argued to support MB learning. By exploring other variants of revaluation (transition and policy revaluation), we were able to provide direct empirical support for SR-based algorithms in human behaviour. The crucial prediction made by the SR account, confirmed in two experiments, was that human participants would be more sensitive to changes in reward structure than to changes in transition and policy structures. Notably, even in the absence of any changes in the transition structure in the policy revaluation condition, experiment 2 showed that participants were also less sensitive to a shift in the optimal policy at intermediate states compared with the reward revaluation condition. This is consistent with SR-based algorithms but inconsistent with either MB algorithms or accounts of different MF–MB arbitration strategies[23] following reward versus state prediction errors.

It is important to stress that the SR is only one of a number of candidates for exact or approximate value computation mechanisms, and our study aimed to find affirmative evidence for its use rather than to argue that it can explain all choice behaviour on its own. Studies using tasks with detour and shortcut manipulations[24], particularly in the spatial domain, are conceptually similar to our transition revaluation. As in our study, some previous research suggests that organisms can in some circumstances also solve these tasks[25]. These results (together with more explicit evidence for step-by-step planning in tasks like chess or in evaluating truly novel compound concepts like tea jelly[2,26]) suggest some residual role for fully MB computation—or, alternatively, that the brain employs additional mechanisms, such as replay-based learning that would achieve the same effect[20].

To reiterate, although our findings argue against a pure MB account (which would handle all our revaluation conditions with equal ease, or symmetrically), they also argue against a pure SR account, which predicts complete insensitivity to transition and policy revaluation (see Figs. 2 and 8). Our data show that people display significant revaluation behaviour even in these conditions, although less than in the reward revaluation condition. Such results are expected under a hybrid SR–MB model in which decision policies reflect a combination of value estimates from the MB strategy and the SR. We demonstrate that this hybrid theory provides a close fit to our data. It is best to think of the combination as a rough proxy for multi-system interactions, which are probably more complex[22] than what we have sketched here. For instance, although we did not formally include or estimate purely MF learning in our modelling here, this is only because it predicts equally bad performance across all of our experimental revaluation conditions. We do not mean to deny the substantial evidence in favour of MF learning in certain circumstances, such as after overtraining. Indeed, MF learning may contribute to our finding that participants do not achieve 100% revaluation performance in any of our conditions, accounting for the slight difference between unnecessary switching in the control condition (which should measure non-specific sloppiness, such as forgetting or choice randomness) and failure to fully adjust in the reward revaluation condition (see Figs. 3 and 6).

Insofar as our results suggest that participants rely on a number of different evaluation strategies, they highlight the question of how the brain determines when to rely on each strategy (an arbitration problem). One general possibility is that humans use a form of meta-decision-making, weighing the costs and benefits of extra deliberation to determine when to invoke MB computation[27–29]. This basic approach might fruitfully be extended to MB versus SR as well as MB versus MF arbitration. A meta-rational agent would be expected to mostly use the computationally cheap SR for flexible, goal-directed behaviour (or the even simpler MF strategy for automaticity in stable environments), but to sometimes employ the more computationally intensive MB strategy to correct the SR-based estimate when needed (for example, when transition structure changes). Given finite computational resources (and the problem that perfectly recognizing the circumstances when MB is required is potentially as hard as MB planning itself) this correction could be insufficient, leaving a residual trace of the biases induced by the SR. Our results on response times in the first experiment may provide a hint of such a hybrid strategy, since the MB system should take longer and might be more likely invoked in the transition revaluation condition (where it is actually needed).

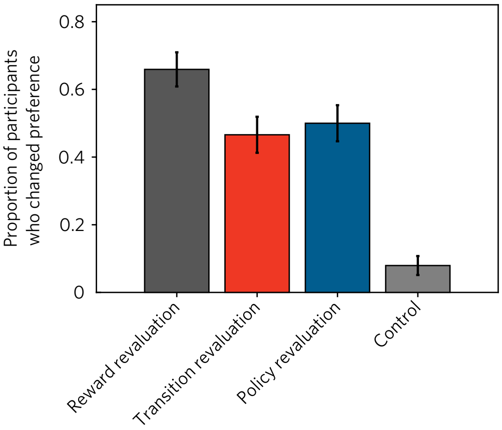

Fig. 7 | Behavioural performance in a sequential decision task. Proportion of participants ($n = 88$) who changed preference following the re-learning phase for reward, transition and policy revaluation as well as the no revaluation control condition. The error bars represent 1 standard error of the proportion estimate.

Another form of SR hybrid could be realized using the MB system (a cognitive map), or episodic memory replay, as a simulator to generate data for training the SR. This resembles the family of Dyna algorithms[20]. Evidence from rodents and human studies showing that offline replay of sequences during rest and sleep enhances memory consolidation[30] and learning new trajectories[31,32]. Because the SR is updated via the simulations of the MB system or episodic memory offline, this Dyna-like hybrid model retains the SR's advantage of fast action evaluation at the decision time (Fig. 8). Updating predictive representations via replay is in line with recent attention to the role of memory systems in planning and decision-making[22,33]. These different realizations of an SR–MB hybrid are essentially speculative in the absence of direct evidence. Further work is required to adjudicate between them.

All these models highlight the fact that the SR is itself a sort of world model, not entirely unlike the sorts of cognitive maps usually associated with the hippocampus. The learned representation is a predictive model, which allows the mental simulation of distal future events rapidly, at least in the aggregate. It differs from the one-step model representations learned and used in standard MB learning, mainly because it aggregates these predictions over many future time steps. This aggregation introduces a new free parameter: the timescale over which future events are aggregated. In theory, the prediction timescale (known as the 'planning horizon') is controlled by the discount factor over future state occupancies in equation (1) (see Methods), and need not, in general, be the same as the agent's time discount preference over delayed rewards[34]. Instead, we predict (and leave to future work to investigate) that the planning horizon should rationally be influenced by the statistical structure of experience, such as the stability or volatility of transitions and rewards in the environment. In other words, the structures of the environment should be reflected in the representations that are learned and stored in memory[35]. For instance, in more stable environments, it may be rational to cache representations with multi-step contingencies over longer planning horizons, compared with volatile environments, where transition contingencies change frequently. In the unstable case, it would be counterproductive to cache contingencies beyond

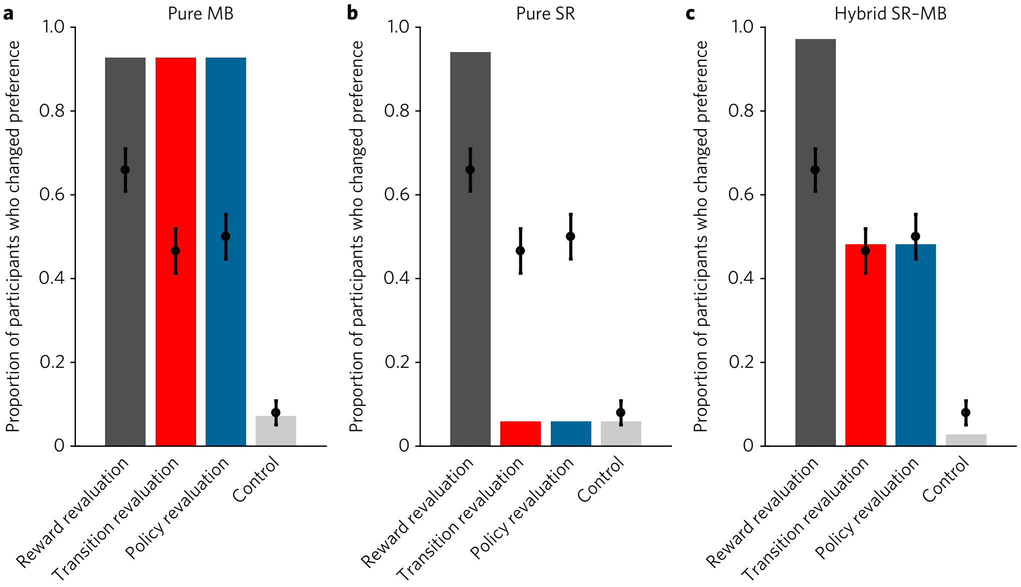

Fig. 8 | Model fits to the data from the sequential decision task. a–c, We compared model performance (solid bars) with human data (error bars) using a pure MB learner (**a**), a pure SR learner (**b**) and a hybrid SR–MB learner (**c**). The proportion of switches predicted by each algorithm under each condition are shown. Consistent with the results from the passive learning task, an algorithm using hybrid representations best captures human behaviour.

the hazard rate of the environment. This idea expands on previous suggestions that environmental volatility should influence the use of MB revaluation versus MF reward caching[36]. A further possibility, which remains to be tested, is that the brain might learn models at multiple timescales simultaneously (or build them up using offline replay) and later adaptively use representations for flexible planning at different scales[37–39].

The SR hypothesis generates clear predictions about the neural representations underlying varieties of revaluation behaviour, which could be tested in future functional neuroimaging studies. At least two major brain structures may underlie the SR: the medial temporal lobe (in particular, the hippocampus) and the prefrontal cortex (PFC). The hippocampus is implicated in the representation of both spatial[40] and non-spatial[41,42] cognitive maps[12] (consistent with Tolman's classic notion[8]), predictive representations of prospective goals[43], as well as associative[44], sequential[45] and statistical learning[12,46]. Hippocampal replay processes help capture the topological structure of novel environments[32] and sequentially simulate and construct paths to future goals[47]—beyond the animal's direct experience—via forward and reverse replay[48]. This is consistent with recent functional magnetic resonance imaging and neural network modelling suggesting a potential role for the SR in complimentary learning systems, especially in the medial PFC and the hippocampus[12,49]. A recent modelling study[11] suggested that the SR could explain the underlying design principles of place cells as studied in rodent electrophysiology. Taken together, these findings lend evidence to the hypothesis that the hippocampus may be involved in building and updating representation of the SR's predictive maps.

The second brain structure that may underlie predictive representations is the PFC. A number of human studies have demonstrated the PFC's role in the representation of prospective goals[50,51]. Lesions to a ventromedial region of the rat PFC impair learning of transition structures (contingencies), but not incentive learning[25]. The ventromedial PFC is well connected to the hippocampus[44] and is thought to mediate sampling information from episodic memory with the goal of decision-making[52] and consolidation[53,54], as well as the comparison and integration of value, abstract state-based inference[55] and latent causes[56]. Furthermore, it has been suggested that the orbitofrontal cortex may also be involved in 'cognitive map'-like representations of task spaces[57] and state spaces[58,59]. A recent finding suggests that the ventromedial PFC and the hippocampus encode proximity to a goal state[60]. Together, the hippocampus and the orbitofrontal cortex may be involved in forming and updating the SR (that is, a rough predictive map of multi-step state transitions, according to simulated experience). Optimal decision-making may rely on the integration of orbitofrontal cortex, ventromedial PFC and hippocampal cognitive maps consistent with our proposed hypothesis of hybrid predictive representations involved in decision-making. Testing the specific role of the PFC and hippocampal contributions to the SR offers an exciting avenue for future functional neuroimaging studies.

In short, we have shown that human behaviour reveals the contribution of a particular sort of internal model of outcome predictions—the SR. We designed varieties of revaluation tasks in which different algorithms for representation learning predict different planning and decision-making behaviour. In contrast to learning only one-step representations, as in classic MB learning, the SR stores multi-step predictive representations of future states. These predictive representations can be learned via mechanisms such as temporal difference learning and can be updated via multiple routes, including direct experience, interaction with rolled-out MB predictions, and simulated experience or offline replay. We have shown that human behaviour under different varieties of revaluation reveals the contribution of such predictive representations. Future studies could explore the individual differences observed here to study the flexibility and pathologies in arbitration of different representation learning approaches in planning and decision-making. We anticipate these findings to open up avenues for computational, electrophysiological and neuroimaging studies investigating the neural underpinnings of this evaluation mechanism.

Methods

Task 1: sequential learning task. A total of 69 participants (mean ± standard deviation (s.d.) age: 22.2 ± 4.6; 42 female) were recruited for the passive learning task, of which four were excluded as they did not learn the task and could not finish the study within the allotted 1.5 h. Seven were removed from the final analysis due to accuracies below 80% in the categorization task (described below)—a threshold used as a measure of attention to (or engagement with) the experiment, leaving 58 participants. The study design and the collection of data complied with all relevant ethical regulations. Princeton University's ethics committee approved the study. All participants signed an informed consent form, reported no history of mental illness, and had normal or corrected-to-normal vision.

Participants played 20 games, each corresponding to one of three conditions: reward devaluation (eight games), transition devaluation (eight games) and a control condition (four games). Each game had three phases: (1) a learning phase, (2) a revaluation phase and (3) a test phase. Games of various conditions were randomly interleaved for each participant. In Fig. 2, schematics of all the phases and two experimental conditions (reward and transition revaluation) are shown as state transition diagrams. 'States' are represented as numbered circles and arrows specify one-step deterministic transitions. Each state was uniquely tagged in each game with a distinct image (of either a face, scene or object; Fig. 2). The stage of the current state within a multi-stage trajectory was indicated by the distinct background colour of that state (for example, state 1 had a green background, state 2 blue and state 3 red; Fig. 2 and Supplementary Fig. 1).

During phase 1 (the learning phase) of each game, participants first experienced all states and their associated reward. They passively traversed six states and learned the transition structure that divided them into two trajectories. To ensure the participants attended to each state, they were asked to perform a category judgement on the images associated with each state (face, scene or object). Phase 1 was concluded once the participant reached a learning criterion, which was reached if they preferred the middle state of the most rewarding trajectory (a preference between state 3 versus 4 in Fig. 2 and Supplementary Fig. 1). The criterion was tested every five trials: participants were shown the middle states of each trajectory (blue background in Fig. 2) and asked which one they preferred. For each trajectory, the learning phase criterion was reached once their preference indicated the middle state of the optimally rewarding trajectory, or after 20 stimulus presentations. Trials in which participants did not reach the learning criterion within the allotted 20 stimulus presentations were excluded from further analysis. During the final test phase, participants were once again shown the starting state of the two trajectories and asked to indicate which one they preferred (that is, which one led to greater reward) on a continuous scale.

During phase 2 (the revaluation phase) participants passively viewed all states except the starting states of each trajectory (states 1 and 2 in Fig. 1); trajectories were always initiated in one of the second-stage states. As in phase 1, participants performed a category judgment on the images of the states they visited. This category task served as a measure of attention to the states during both phases. In the control condition, there were no changes to the task structure. In the reward revaluation condition, the rewards associated with the terminal states of the trajectories were swapped. In the transition revaluation condition, the connectivity between the second- and third-stage states was altered, such that the middle state of a given trajectory now led to the final state of the other trajectory (Fig. 2). As in phase 1, participants were probed for their preference of the middle states every five stimuli, and phase 2 concluded once they met the learning criterion (three correct decisions about the middle states) or after 20 stimuli. During phase 3, participants were instructed to once again rate their preference for the start states.

Experiment 1: computational models. We compared the performance of the following three models to human behaviour: (1) MB learning (computing values using knowledge of the transition and reward functions; Fig. 5a); (2) pure SR learning (computing values using estimates of the reward function and SR; Fig. 5b); and (3) a hybrid SR–MB model that linearly combines the ratings of the two learners (Fig. 5c).

The SR learner model uses two structures to compute state value: a vector R and a matrix M. R_s stores the immediate reward expected upon encountering state s. $M_{s,s'}$ stores the expected number of (future discounted) visits to states′ along a sequence of states starting in state s:

$$M_{s,s'} = E\left[\sum_{t=0}^{\infty} \gamma^t I(s_t = s') | s_0 = s\right] \quad (1)$$

where E denotes expectation, $I(\cdot) = 1$ if its argument is true, and 0 otherwise, s_t is the state encountered t time-steps following the visit of state s_0 and γ is a discount parameter. Since in our task the terminal states are absorbing, we set $\gamma = 1$ (that is, no discounting). The SR learner combines these two structures to compute the value of a state s, $V(s)$, by taking the inner product of R and the row of M corresponding to that state:

$$V(s) = \sum_{s'} M_{s,s'} \times R_{s'} \quad (2)$$

The MB learner computes state values by iterating the Bellman equation over all states until convergence[61]:

$$V(s) = R_s + \gamma V(s') \quad (3)$$

where s' is the immediate successor of state s.

We assumed that preference ratings were generated by a scaled function of the state values:

$$Rating = b \times (V(2) - V(1))$$

where b is a free parameter. For the SR–MB hybrid model, we assumed that the preference rating was a linear combination of the ratings generated by the two component models:

$$Rating_{hybrid} = w \times Rating_{MB} + (1-w) \times Rating_{SR}$$

where w is a free parameter.

Note that our strategy here is to model such ratings predicted by the different algorithms' representations, rather than the trial-by-trial learning process that produced these representations. This is because the structure of the task does not provide enough variability in participants' experience, or monitoring of participants' ongoing beliefs, to constrain trial-by-trial learning within the acquisition phases. In particular, because the experienced rewards and transitions are deterministic within a given phase of each game, variables like learning rates, which would govern the rate at which model representations reach their asymptotic values, are under-constrained. Furthermore, the task is passive; participants' beliefs are tested only sporadically and indirectly with relative preference judgements.

We therefore assume that by the end of each phase, each model representation has reached its asymptotic value, consistent with the information presented and experiences permitted during that phase, and with the usual learning rules for these algorithms. Specifically, we assume that at the end of phase 1 and again following phase 2, the MB learner has appropriately updated the transition function (providing which s' follows which s) to the most recently experienced contingencies, the SR learner has appropriately updated $M(s, s')$ and both learners have appropriately updated $R(s)$. Importantly—to capture what would be the endpoint of trial-by-trial learning of the different representations—in each case we assume the various representations are only updated for all states s visited during a phase; representations for states not visited in phase 2 remain unchanged. Using these updated representations, we compute $V(s)$ at the end of phase 1 and again at the end of phase 2 using equation (2) for the SR learner and equation (3) for the MB learner. $V(s)$ is used to derive ratings at the end of phase 1 and the beginning of phase 3. Revaluation scores are then computed by subtracting the phase 1 rating from the phase 3 rating:

$$Score = Rating_{phase3} - Rating_{phase1}$$

A participant's revaluation score for a given trial was modelled as being drawn from a Gaussian distribution, centred on the model's predicted score, for that revaluation condition:

$$f(score = s) = Normal(s; \mu = Score, \sigma^2)$$

where f is the density of a score s. Here, μ and σ^2 are the mean and variance parameters of a normal distribution. σ^2 is a free parameter.

Together, SR and MB each had two free parameters (b and σ^2) and the hybrid model had three (b, w and σ^2). For each participant, we estimated parameters b and w by maximizing the likelihood of their revaluation scores, jointly with the group-level distributions over the entire population using an expectation maximization procedure[62]. For each participant, the maximum likelihood estimate of σ^2 was directly computed based on the residuals between the model estimates (determined by the other parameters) and the data:

$$\sigma^2 = \frac{\sum_t \left(y_t - \hat{y}_t\right)^2}{n}$$

where y_t is the participant's revaluation score on trial t, $\hat{y}_t$ is the model's predicted revaluation score for trial t and n is the number of trials. To constrain w to be between 0 and 1, the likelihood function passed the input to parameter w, w_{in} through the transformation

$$w = 0.5 + 0.5 \times \mathrm{erf}\left(\frac{w_{in}}{sqrt(2)}\right)$$

Aggregate log-likelihoods for each model were computed using a leave-one-out, participant-level, cross validation procedure. This involved, for a given participant j, using data from every other participant $i \neq j$ to fit group-level parameter distributions. The log likelihood of participant j's ratings was then computed averaging the log-likelihood of 10,000 samples taken from the group distribution. We repeated this procedure for each participant and computed aggregate log-likelihoods by summing over participants.

Experiment 2: sequential decision task. This task was run on Amazon's Mechanical Turk using psiTurk software[63]. The study was approved by New

York University's human subject committee. We set out to collect data from 100 participants, and 112 participants (mean ± s.d. age: 33.6 ± 10.5; 53 females) were recruited to complete the experiment. All participants were required to achieve 100% accuracy on a nine-item instruction comprehension task before beginning the task. In total, 24 participants were excluded for failing to learn the appropriate decision policy at the end of the phase 1 training (preference 1) in at least one of the conditions (see below). After collecting data from 100 participants, we removed those who failed to pass the exclusion criteria and then continued to collect until we had an equal number of non-removed participants in each of the four experimental conditions (88 participants; age: 33.8 ± 10.8; 45 females), each of whom corresponded to an order of revaluation trial types. Participants received a bonus proportional to the total value of the reward collected.

Participants made choices in order to collect rewards by navigating an avatar through the rooms of a castle. The underlying structure of each condition of the task is displayed schematically in Fig. 6. In each trial, participants were placed in one of the six states (castle rooms) and were required to make choices until they arrived at a terminal state and collected the associated reward. States were displayed as coloured shapes on a screen. The spatial position and colour of each state was randomized across blocks, yet remained fixed within a block.

Each participant performed four blocks of trials. Each block corresponded to a different condition. The block order was counterbalanced across participants according to a Latin square design. Each block consisted of three phases (Fig. 6). In the learning phase (phase 1), participants were trained on a specific reward and transition structure. Training involved completing 39 trials. The starting state for each trial was randomized so that at least 14 trials began from state 1, at least 7 trials began from each of states 2 and 3, and at least 2 trials began from each terminal state. In each condition, the reward and available actions for phase 1 were arranged so that state 6 contained the highest reward and was exclusively accessible from state 3. Thus, by the end of phase 1, participants should have learned to select the state 1 action leading to state 3. The other terminal states, respectively, contained low and medium-sized rewards. One of the other terminal states was accessible from both states 2 and 3 and one was accessible exclusively from state 3. This arrangement ensured that there was a 'correct' action from each non-terminal state that would lead to a higher reward. If, for any condition, participants failed to perform the correct action from each non-terminal state on three of their last four visits to that state, they were removed from the analysis.

In phase 2, a change in either the reward associated with one of the terminal states or the set of available actions in states 2 and 3 occurred. In the reward revaluation condition, the amount of reward in state 4, which previously contained the highest reward accessible from state 2, was increased so that this state was now the most rewarding terminal state. This change thus altered the reward of the state that the participant had previously experienced as following the state 1 action leading to state 2. In the transition revaluation condition, the set of available actions in states 2 and 3 was changed so that state 6—the terminal state containing the highest reward—could be reached exclusively from state 2. This change thus altered which terminal state could follow either state 1 action and was thus comparable to the transition revaluation in passive learning task. In the policy revaluation condition, the amount of reward in the terminal state containing the smallest reward was increased so that this state now contained the highest reward. Because this would alter which action was preferred from state 2, it changed which terminal state would be expected to follow the action leading to state 2. Thus, despite involving a reward change, this change would have similar effects on SR models as the transition revaluation. Finally, in the control condition, the amount of reward in the state containing the highest amount of reward increased. In each condition, phase 2 consisted of nine trials. In the reward revaluation, policy revaluation and control conditions, these trials started three times from each terminal state. Phase 2 trials in the transition revaluation started three times from both state 2 and state 3 to allow participants to observe the change in available actions, and once from each terminal state. Crucially, participants did not visit the start state (state 1) during phase 2, and hence never experienced any changes in reward following an action taken from the start state. In the test phase (phase 3), participants performed a single trial beginning from the start state. In phase 3, participants completed a single trial starting from state 1. We defined the revaluation score as 1 if they switched to the now better action leading to state 2 and 0 if they stayed with the action leading to state 3.

Logistic regression analysis. All of our descriptive analyses involved performing pairwise comparisons between the proportions of participants who switched action preference following different revaluation conditions. To perform such pairwise comparisons while correctly accounting for the repeated-measures structure of the experiment, we fit a logistic regression model where the dependent variable was a binary indicator of whether a given participant changed action preference in state 1 between phases 1 and 3. The model had four independent variables: a binary indicator variable for each condition that was set to 1 when the given response was from that condition. This model provided a coefficient estimate for each condition indicating the logit-transformed probability that participants switched their state 1 action preference in phase 3 of that condition. To obtain standard errors on coefficient estimates that accounted for participant-level clustering due to the repeated measures, we employed a cluster-robust Huber–White estimator (using the robcov function from the R package rms; ref. [64]). Contrasts between coefficients were computed by fitting the model once for each condition and substituting that condition in as the intercept so that coefficient estimates for the other three conditions represented contrasts from it.

Experiment 2: computational models. *ε-greedy model.* All learners convert action values, Q, to choice probabilities using an ε-greedy rule. This rule chooses the available action with the max Q value with probability $1-\varepsilon$ and chooses a random available action with probability ε. Thus, for available actions sa in state s:

$$P(sa|s) = \begin{cases} 1 - \dfrac{(N-1)\varepsilon}{N} & \text{if } \arg\max_{j} Q(j) = sa; \\ \dfrac{\varepsilon}{N} & \text{otherwise}. \end{cases} \tag{4}$$

where N is the number of actions available in state s'. We consider terminal states to have a single available action. We set $P(sa|s) = 0$ for all actions not available in state s.

As in our modelling of the passive learning task, the SR learner uses two structures to compute value: a reward vector, R, and a matrix of expected future occupancies, M. The only change here is that the elements of M are indexed by actions, sa. Likewise, R_{sa} stores the reward associated with taking action a from state s. $M(sa,s'\,a')$ stores the expected future (cumulative, discounted) number of times action $s'\,a'$ will be performed on a trial following action a from state s, sa. The SR learner combines these two structures to compute the value of an action by taking the inner product of the reward vector R and the row of M that corresponds to that action in the current state:

$$Q(sa) = \sum\nolimits_{s'} M_{sa,s'a'} \times R_{s'a'}. \tag{5}$$

The MB learner computes value estimates by combining knowledge of the transition and reward structure, iterating the following Bellman equation until convergence:

$$Q(sa) = R_{sa} + \max_{s'a^{*} \in A_{s'}} \gamma Q(s'a^{*}) \tag{6}$$

where s' is the state to which $s'\,a'$ transitions and $A_{s'}$ is the set of actions, $s'a'$, available in state s'.

The SR–MB hybrid learner forms action probabilities by combining action probabilities from both SR and MB learners $P_{sr}(sa|s)$ and $P_{mb}(sa|s)$. The model assumes that the two action probabilities are combined according to a weighted average:

$$P_{hybrid}(sa|s) = w \times P_{mb}(sa|s) + (1-w) \times P_{sr}(sa|s)$$

where w is a free parameter.

As with the passive learning task, because parameters like learning rates are under-constrained, we assume that by the end of each phase, each model representation has reached its asymptotic value, appropriately updated according to the information presented and the experiences permitted during that phase. For the active learning task, this means that at the end of phase 1 and also phase 2, the MB learner has appropriately updated A_s, the SR learner has updated $M(sa,s'a')$ and both learners have adjusted $R(sa)$, but again in each case only for all states s visited and actions sa performed during that phase.

Using these updated representations, we compute $Q(sa)$ at the end of phase 1 and again at the end of phase 2 using equation (5) for the SR learner and equation (6) for the MB learner. $Q(sa)$ is used to derive action for each action at the end of phase 1 and also at the beginning of phase 3.

As in experiment 1, we computed an aggregate log-likelihood of each participant's phase 3 choice under each model using a leave-one-out, participant-level cross-validation procedure. Parameters w and ε were constrained to be between 0 and 1 by being passed through the same transformation used to constrain w in experiment 1. For each participant, we fit a distribution of group level parameters to participants' choices using data from every other participant. To constrain noise parameters (ε), we included the participant's last four choices in phase 1 from each decision state in addition to the participant's phase 3 choice in the likelihood function used to estimate parameters. We chose these choices because our exclusion criterion assumed that participants' learning would have reached asymptote by this point (and they were excluded if they did not perform the correct action in three of their last four visits to each of these states). We then computed the log likelihood of participant j's phase 3 choices by averaging the log-likelihood of this choice using 10,000 samples taken from the group distribution. The aggregate log-likelihood of the phase 3 choices under the model was computed by summing over the individual likelihoods computed for each participant.

Replication and randomization. While the two experiments were conceptual replications of one another, using different designs and choice approaches to compare similar conditions, we did not conduct individual replications of each experimental design. We employed within-participant designs: each participant underwent an equal number of all conditions in a randomized fashion. The order of experimental conditions was randomized by the experiment's code; therefore, investigators were blind to the specific order for each participant.

Code availability. The costume codes used for generating the models (Figs. 5 and 8) are available from the corresponding author on request.

Data availability. The data that support the findings of this study (the source data in Figs. 4 and 7) are available from the corresponding author on request.

Received: 17 October 2016; Accepted: 7 July 2017;
Published online: 28 August 2017

References

1. Dayan, P. Twenty-five lessons from computational neuromodulation. *Neuron* **76**, 240–256 (2012).
2. Daw, N. D. & Dayan, P. The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B* **369**, 20130478 (2014).
3. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130480 (2014).
4. Dayan, P. Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624 (1993).
5. Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A. & Sederberg, P. B. The successor representation and temporal context. *Neural Comput.* **24**, 1553–1568 (2012).
6. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
7. Dickinson, A. Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. B Biol. Sci.* **308**, 67–78 (1985).
8. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
9. Lengyel, M. & Dayan, P. *Hippocampal Contributions to Control: The Third Way* in *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2007).
10. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).
11. Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. The hippocampus as a predictive map. Preprint at http://www.biorxiv.org/content/early/2017/07/27/097170 (2017).
12. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nat. Neurosci.* **16**, 486–492 (2013).
13. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* **6**, e17086 (2017).
14. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. Preprint at http://www.biorxiv.org/content/early/2016/10/27/083857 (2017).
15. 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).
16. Sadacca, B. F., Jones, J. L. & Schoenbaum, G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife* **5**, e13665 (2016).
17. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
18. Brogden, W. J. Sensory pre-conditioning. *J. Exp. Psychol.* **25**, 323 (1939).
19. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
20. Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin* **2**, 160–163 (1991).
21. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
22. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
23. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
24. Spiers, H. J. & Gilbert, S. J. Solving the detour problem in navigation: a model of prefrontal and hippocampal interactions. *Front. Hum. Neurosci.* **9**, 125 (2015).
25. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).
26. Shohamy, D. & Daw, N. D. Integrating memories to guide decisions. *Curr. Opin. Behav. Sci.* **5**, 85–90 (2015).
27. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).
28. Boureau, Y.-L., Sokol-Hessner, P. & Daw, N. D. Deciding how to decide: self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015).
29. Kool, W., Cushman, F. A. & Gershman, S. J. When does model-based control pay uff? *PloS Comput. Biol.* **12**, e1005090 (2016).
30. Karlsson, M. P. & Frank, L. M. Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* **12**, 913–918 (2009).
31. Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. Hippocampal place cells construct reward related sequences through unexplored space. *eLife* **4**, e06063 (2015).
32. Wu, X. & Foster, D. J. Hippocampal replay captures the unique topological structure of a novel environment. *J. Neurosci.* **34**, 6459–6469 (2014).
33. Doll, B. B., Shohamy, D. & Daw, N. D. Multiple memory systems as substrates for multiple decision systems. *Neurobiol. Learn. Mem.* **117**, 4–13 (2015).
34. Jiang, N., Kulesza, A., Singh, S. & Lewis, R. *The Dependence of Effective Planning Horizon on Model Accuracy* in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (IFAAMAS, 2015).
35. Anderson, J. R. & Schooler, L. J. Reflections of the environment in memory. *Psychol. Sci.* **2**, 396–408 (1991).
36. Simon, D. A. & Daw, N. D. *Environmental Statistics and the Trade-off Between Model-Based and TD Learning in Humans* in *Proceedings of the 24th International Conference on Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2011).
37. Sutton, R. S. *TD Models: Modeling the World at a Mixture of Time Scales*. (University of Massachusetts, Amherst, MA, 1995).
38. Tanaka, S. C. et al. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* **7**, 887–893 (2004).
39. Kurth-Nelson, Z. & Redish, A. D. Temporal-difference reinforcement learning with distributed representations. *PLoS ONE* **4**, e7362 (2009).
40. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
41. Barron, H. C., Dolan, R. J. & Behrens, T. E. J. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* **16**, 1492–1498 (2013).
42. Tavares, R. M. et al. A map for social navigation in the human brain. *Neuron* **87**, 231–243 (2015).
43. Brown, T. I. et al. Prospective representation of navigational goals in the human hippocampus. *Science* **352**, 1323–1326 (2016).
44. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* **23**, R764–R773 (2013).
45. Foster, D. J. & Knierim, J. J. Sequence learning and the role of the hippocampus in rodent navigation. *Curr. Opin. Neurobiol.* **22**, 294–300 (2012).
46. Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M. & Turk-Browne, N. B. The necessity of the medial temporal lobe for statistical learning. *J. Cogn. Neurosci.* **26**, 1736–1747 (2014).
47. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Hippocampal replay is not a simple function of experience. *Neuron* **65**, 695–705 (2010).
48. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
49. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160049 (2017).
50. Momennejad, I. & Haynes, J.-D. Human anterior prefrontal cortex encodes the 'what' and 'when' of future intentions. *NeuroImage* **61**, 139–148 (2012).
51. Momennejad, I. & Haynes, J.-D. Encoding of prospective tasks in the human prefrontal cortex under varying task loads. *J. Neurosci.* **33**, 17342–17349 (2013).
52. Euston, D. R., Gruber, A. J. & McNaughton, B. L. The role of medial prefrontal cortex in memory and decision making. *Neuron* **76**, 1057–1070 (2012).
53. Maguire, E. A. Memory consolidation in humans: new evidence and opportunities. *Exp. Physiol.* **99**, 471–486 (2014).
54. Nieuwenhuis, I. L. C. & Takashima, A. The role of the ventromedial prefrontal cortex in memory consolidation. *Behav. Brain Res.* **218**, 325–334 (2011).
55. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).

56. Wunderlich, K., Dayan, P. & Dolan, R. J. Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* **15**, 786–791 (2012).
57. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
58. Wikenheiser, A. M. & Schoenbaum, G. Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nat. Rev. Neurosci.* **17**, 513–523 (2016).
59. Ramus, S. J. & Eichenbaum, H. Neural correlates of olfactory recognition memory in the rat orbitofrontal cortex. *J. Neurosci.* **20**, 8199–8208 (2000).
60. Balaguer, J., Spiers, H., Hassabis, D. & Summerfield, C. Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron* **90**, 893–903 (2016).
61 Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* Vol. 1 (MIT Press, Cambridge, MA, 1998).
62. Huys, Q. J. M. et al. Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLOS Comput. Biol.* **7**, e1002028 (2011).
63. Gureckis, T. M. et al. psiTurk: an open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* **48**, 829–842 (2015).
64. Huber, P. *The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. (Univ. California Press, Berkeley, CA, 1967).

Acknowledgements

This project was made possible by grant support through the National Institutes of Health Collaborative Research in Computational Neuroscience award 1R01MH109177, the National Institutes of Health under R. L. Kirschstein National Research Service Award 1F31MH110111-01 and the John Templeton Foundation. The authors acknowledge K. Norman and R. Otto for helpful conversations, and A. Rich and S. Tubridy for assistance with psiTurk. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies. No funders had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

I.M., M.M.B. and S.J.G. designed experiment 1. J.H.C. conducted and collected the data. E.M.R. and N.D.D. designed and conducted experiment 2. I.M., E.M.R. and S.J.G. analysed the data and ran model simulations. I.M., E.M.R., M.M.B., N.D.D. and S.J.G. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41562-017-0180-8).

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to I.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Ida Momennejad

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Both experiments were inspired by the procedure described in ershman, S.J., Markman, A.B., Otto, A.R. J. Exp. Psychol. Gen. 143, 182-194 (2014), which used a sample size of 60, for which we recruited 69 participants and arrived at n=58 in experiment 1, after excluding participants who did not satisfy minimal learning criteria or did not finish the task. We increased this n by at least 30% in experiment 2, given the choice nature of the task. After recruiting 112 participants and excluding participants who did not satisfy our learning criteria in all conditions, we arrived at n=88 for experiment 2.

2. Data exclusions

Describe any data exclusions.

Experiment 1:
4 out of 69 participants were excluded as they did not learn the task and could not finish the study within the allotted 1.5 hours. 7 were removed from final analysis due to accuracies below 80% in the categorization task (described below), a threshold used as a measure of attention to (engagement with) the experiment, leaving 58 participants.

Experiment 2:
115 participants were recruited to complete the experiment. All participants were required to achieve 100% accuracy on a 9-item instruction comprehension task before beginning the task. 27 participants were excluded for failing to learn the appropriate decision policy at the end of phase 1 training (Preference 1) in at least one of the conditions.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The two experiments take different approaches to get at the same effect, thus they are conceptual replications of one another. However, we did not individually replicate each experimental paradigm.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

We employed within-subject designs, where each participant underwent equal number of all conditions in within-subject randomized order.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The order of the conditions was randomized within participants, and determined by the code that produced the experimental task and stimuli during each session.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Matlab, Psiturk software55, Julia, R.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No laboratory animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Experiment 1
Human participants were recruited via the Princeton University subject recruitment interface (SONA). 69 (mean age = 22.2, STD = 4.6, 42 female) participants were recruited for the passive learning task, of which 4 participants were excluded as they did not learn the task and could not finish the study within the allotted 1.5 hours. 7 were removed from final analysis due to accuracies below 80% in the categorization task (described in the text) a threshold used as a measure of attention to (engagement with) the experiment, leaving 58 participants.

Experiment 2
This task was run on Amazon's Mechanical Turk (AMT) using Psiturk software. 112 participants (mean age = 33.6, STD = 10.5, 53 female) were recruited to complete the experiment. All participants were required to achieve 100% accuracy on a 9-item instruction comprehension task before beginning the task. 24 participants were excluded for failing to learn the appropriate decision policy at the end of phase 1 training (Preference 1) in at least one of the conditions. Data from 88 participants (mean age =33.8, STD = 10.8, 45 female) was used for further analysis.