nature neuroscience

Dorsal hippocampus contributes to model-based planning

Kevin J Miller[1], Matthew M Botvinick[1–3] & Carlos D Brody[1,4]

Planning can be defined as action selection that leverages an internal model of the outcomes likely to follow each possible action. Its neural mechanisms remain poorly understood. Here we adapt recent advances from human research for rats, presenting for the first time an animal task that produces many trials of planned behavior per session, making multitrial rodent experimental tools available to study planning. We use part of this toolkit to address a perennially controversial issue in planning: the role of the dorsal hippocampus. Although prospective hippocampal representations have been proposed to support planning, intact planning in animals with damaged hippocampi has been repeatedly observed. Combining formal algorithmic behavioral analysis with muscimol inactivation, we provide causal evidence directly linking dorsal hippocampus with planning behavior. Our results and methods open the door to new and more detailed investigations of the neural mechanisms of planning in the hippocampus and throughout the brain.

Imagine a game of chess. As the players think about their next moves, they consider the outcome each action would have on the board, as well as the opponent's likely reply. The players' knowledge of the board and the rules constitutes an internal model of chess, a knowledge structure that links actions to their likely outcomes. The process of using such an 'action–outcome' model to inform behavior is defined within reinforcement learning theory as the act of planning[1]. Planning, so defined, has been an object of scientific investigation for many decades, and this research has generated important insights into the planning abilities of both humans and other animals[2–5].

Despite this progress, the neural mechanisms that underlie planning remain frustratingly obscure. One important reason for this continuing uncertainty lies in the behavioral assays that have traditionally been employed. Until recently, research on planning has largely employed behavioral tests (for example, outcome devaluation) in which the subject is put through a sequence of training stages, then makes just one decision to demonstrate planning (or an absence thereof)[2,6,7]. While the same animal can be tested multiple times[8], at most one behavioral measure is obtained per session. Seminal studies using these assays have established the relevance of several neural structures[3,4], and they continue to be fundamental for many experimental purposes, but these assays are constrained by the small number of planned decisions they elicit. In an important recent breakthrough, new tasks have been developed that lift this constraint[9–12], allowing the collection of many repeated trials of planned behavior. These tasks provide an important complement to existing behavioral assays, promising to allow both a detailed evaluation of competing models as well as new opportunities for experiments investigating the neural mechanisms of planning. They have, however, so far been applied only to human subjects, limiting the range of experimental techniques available.

Here we have adapted one of these tasks (the 'two-step' task[9]) for rats, combining for the first time a multitrial decision task with the experimental toolkit available for rodents. First, we conducted a set of detailed computational analyses on a large behavioral dataset and confirmed that rats, like humans, employ model-based planning to solve the task. In a second experiment, we employed causal neural techniques not available in humans to address an important open question in the neuroscience of planning: the role of the dorsal hippocampus.

A long-standing theory of hippocampal function holds that it represents a 'cognitive map' of physical space used in support of navigational decision-making[13]. Classic experiments demonstrate hippocampal involvement in navigation tasks[14,15], as well as the existence of 'place cells', which both encode current location[16] and 'sweep out' potential future paths at multiple timescales[17,18]. These findings have given rise to computational accounts of hippocampal function that posit a key role for the region in model-based planning[19–21]. However, support for these theories from experiments employing causal manipulations has been equivocal. Studies of both spatial navigation and instrumental conditioning have shown intact action–outcome behaviors following hippocampal damage[22–27]. At the same time, tasks requiring relational memory do show intriguing impairments following hippocampal damage[28–30]. The latter tasks assay whether behavior is guided by knowledge of relationships between stimuli (stimulus–stimulus associations), which plausibly involves representations and structures similar to those for the action–outcome associations that underlie planning, but they do not focus on action–outcome associations. Here with the two-step task, we isolate these action–outcome associations specifically.

Using rats performing the two-step task, we performed reversible inactivation experiments in both dorsal hippocampus (dH) and in

[1]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey, USA. [2]Gatsby Computational Neuroscience Unit, University College London, London, UK. [3]Google DeepMind, London, UK. [4]Howard Hughes Medical Institute and Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA. Correspondence should be addressed to M.M.B. (botvinick@google.com) or C.D.B. (brody@princeton.edu).

Received 28 December 2016; accepted 20 June 2017; published online 31 July 2017; doi:10.1038/nn.4613

orbitofrontal cortex (OFC), a brain region widely implicated in model-based control (i.e., planning) in traditional assays[31–33]. The repeated-trials nature of the task allowed us to use computational modeling to identify a set of separable behavioral patterns that jointly explained observed behavior and to quantify the relative strength of each pattern. We found that the behavior of our animals was dominated by a pattern consistent with model-based planning, with important influences of novelty aversion, perseveration and bias. The model-based pattern was selectively impaired by inactivation of OFC or dH, while other patterns were unaffected.

Notably, model-based planning depends on a number of computations; behaviorally observed planning impairments might be caused by impairments to the planning process itself or instead by impairments to learning and memory processes upon which planning depends. Computational modeling analysis indicates that our effects are not well-described as an impairment in learning or memory in general but as a specific attenuation of planned behavior. We therefore conclude that these regions either perform computations integral to the planning process itself (i.e., use the action–outcome model to inform choice) or represent inputs used specifically by the planning process. This provides what is, to our knowledge, the first causal evidence that dH contributes to model-based planning.

RESULTS

We trained rats to perform a multitrial decision making task[9], adapted from the human literature, designed to distinguish model-based versus model-free behavioral strategies (the two-step task; **Fig. 1**). In the first step of the task, the rat chooses between two choice ports, each of which leads to one of two reward ports becoming available with probability 80% (common transition) and to the other reward port becoming available with probability 20% (uncommon transition). In the second step, the rat does not have a choice but is instead instructed as to which reward port has become available, enters it and either receives (reward) or does not receive (omission) a bolus of water. Reward ports differ in the probability with which they deliver reward, and reward probability changes at unpredictable intervals (Online Methods). Optimal performance requires learning which reward port currently has the higher reward probability and selecting the choice port more likely to lead to that port. This requires using knowledge of the likely outcomes that follow each possible chosen action—that is, it requires planning.

Rats performed the two-step task in a behavioral chamber outfitted with six nose ports arranged in two rows of three (**Fig. 1b**). Choice ports were the left and right side ports in the top row, and reward ports were the left and right side ports in the bottom row. Rats initiated each trial by entering the center port on the top row, and then indicated their choice by entering one of the choice ports. An auditory stimulus then indicated which of the two reward ports was about to become available. Before entering the reward port, however, the rat was required to enter the center port on the bottom row. This kept motor acts relatively uniform across common and uncommon trial types. For some animals, the common transition from each choice port led to the reward port on the same side (**Fig. 1a**; 'common-congruent' condition), while for others it led to the reward port on the opposite side ('common-incongruent'). These transition probabilities constituted stable relationships between actions (choice ports) and their likely outcomes (reward ports). Subjects therefore had the opportunity to incorporate these action–outcome relationships into an internal model and to use them in order to plan.

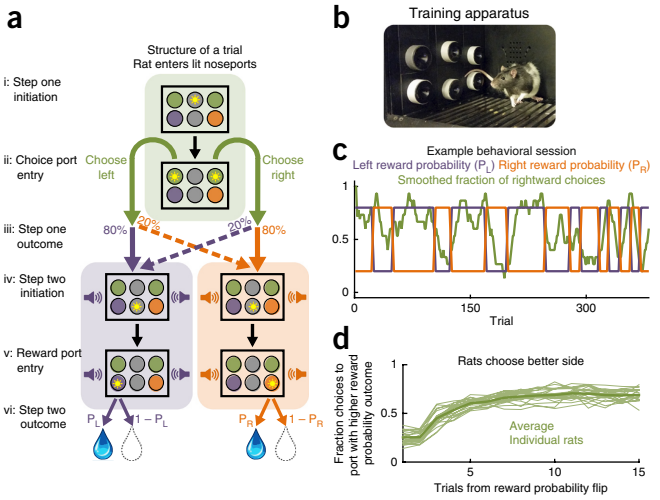

Figure 1 Two-step decision task for rats. (**a**) Structure of a single trial of the two-step task. (i) Top center port illuminates to indicate trial is ready, and the rat enters it to initiate the trial. (ii) Choice ports illuminate, and the rat indicates its decision by entering one of them. (iii) Probabilistic transition takes place, with probability depending on the choice of the rat. Sound begins to play, indicating the outcome of the transition. (iv) Center port in the bottom row illuminates, and the rat enters it. (v) The appropriate reward port illuminates, and the rat enters it. (vi) Reward is delivered with the appropriate probability. P_L, probability that the left port will provide a reward; P_R, probability that the right port will provide a reward. (**b**) Photograph of behavioral apparatus, consisting of six nose-ports with LEDs and infrared beams, as well as a speaker mounted in the rear wall. (**c**) Example behavioral session. Rightward choices are smoothed with a 10-trial boxcar filter. At unpredictable intervals, reward probabilities at the two ports flip synchronously between high and low. Rats adapt their choice behavior accordingly. (**d**) Choice data for all rats ($n = 21$). The fraction of trials on which the rat selected the choice port whose common (80%) transition led to the reward port with currently higher reward probability, as a function of the number of trials that have elapsed since the last reward probability flip.

Two analysis methods to characterize behavior and quantify planning

We trained 21 rats to perform the two-step task in daily behavioral sessions (n = 1,959 total sessions), using a semiautomated training pipeline that enabled us to run large numbers of animals in parallel with minimal human intervention (Online Methods). Although optimal performance in the two-step task requires planning, good performance can be achieved by both planning and model-free strategies (**Supplementary Fig. 1**). Critically, however, each type of strategy gives rise to different patterns of choices[9]. Model-free strategies tend to repeat choices that resulted in reward and avoid choices that led to omission, regardless of whether the transition after the choice was a common or an uncommon one. Planning strategies, in contrast, are by definition aware of these action–outcome probabilities. Thus, after an uncommon transition, planning strategies tend to avoid choices that led to a reward, because the best way to reach the rewarding port again is through the common transition that follows the opposite choice. Similarly, after an uncommon transition, planning strategies tend to repeat choices that led to a reward omission, because the best way to avoid the unrewarding port is through the common transition likely to occur after repeating the choice. Following this logic, Daw *et al.*[9] examined how humans' choices in a given trial depend on the immediately previous trial and concluded that humans appear to use a mixture of model-free strategies and model-based planning.

To assess the extent to which rat subjects were using a planning strategy, we extended the analysis of Daw *et al.*, which considered the influence of the immediately preceding trial on present-trial behavior[9], to use information from multiple trials in the past (**Supplementary Fig. 2**). We have shown separately that this many-trials-back approach is robust to some potential artifacts (for example, due to slow learning rates[34]). The many-trials-back approach consists of a logistic regression model that predicts the choice of the rat on each trial, given the history of recent trials and their outcomes. A trial that occurred τ trials ago can be one of four types: common-rewarded (CR), uncommon-rewarded (UR), common-omission (CO) and uncommon-omission (UO). For each τ, each of these trial types is assigned a weight ($\beta_{CR}(\tau)$, $\beta_{UR}(\tau)$, $\beta_{CO}(\tau)$ and $\beta_{UO}(\tau)$ respectively). Positive weights correspond to a greater likelihood that the rat will make the same choice that was made on a trial of that type that happened τ trials in the past, while negative weights correspond to a greater likelihood that the rat will make the other choice. The weighted sum of past trials' influence then dictates choice probabilities (Online Methods). Notably, because model-free strategies do not distinguish between common and uncommon transitions, model-free strategies will tend to have $\beta_{CR} \approx \beta_{UR}$ and $\beta_{CO} \approx \beta_{UO}$. In contrast, model-based strategies tend to change their behavior in different ways following common versus uncommon transitions and will therefore have $\beta_{CR} > \beta_{UR}$ and $\beta_{CO} < \beta_{UO}$.

Applying this approach to synthetic data from artificial reinforcement-learning agents using planning or model-free strategies (Online Methods) yields the expected patterns (**Fig. 2a**,**b**). For the planning agent (**Fig. 2a**), trials with common and uncommon transitions have opposite effects on the current choice. In contrast, for the model-free agent (**Fig. 2b**), common and uncommon transition trials have the same effect, and only reward versus omission is important. **Figure 2c** shows the result of fitting the regression model to data from an example rat. The behavioral patterns observed are broadly similar to those expected of a model-based agent (cf. **Fig. 2a**,**c**).

We next applied this approach to the behavior of each rat in our dataset (**Supplementary Fig. 3**) to reveal the nature of that animal's choice strategy. To quantify the overall extent to which each rat showed evidence of planning versus a model-free strategy, we defined a 'planning index' and a 'model-free index' by summing over the regression weights consistent with each pattern (**Fig. 2** and Online Methods). We have previously found that these measures provide a more reliable guide to behavioral strategy than standard measures, which consider only the immediately previous trial[34]. We found that trained rats overwhelmingly showed large positive planning indices (**Fig. 2**; mean over rats, 4.2; standard error, 0.3) and small positive model-free indices (mean, 0.6; standard error, 0.1), consistent with their having adopted a planning strategy. Similarly, we found that movement times from the bottom center port to the reward port were faster for common versus uncommon transition trials (average median movement time, 700 ms for common and 820 ms for uncommon, $P < 10^{-5}$; **Supplementary Fig. 4**), further indicating that rats used knowledge of the transition probabilities to inform their behavior. These results were similar between rats in the common-congruent condition (common outcome for each choice port is the reward port on the same side; **Fig. 1a**) and those in the common-incongruent condition (common outcome is the port on the opposite side; $P > 0.2$).

This regression analysis also revealed substantial rat-by-rat variability (**Fig. 3a**). Furthermore, there are noteworthy deviations from the predicted model-based pattern (**Fig. 3a** and **Supplementary Fig. 4**). For example, the rat in the top left panel of **Figure 3a** (same rat as in **Fig. 2c**) showed the overall pattern of regression weights

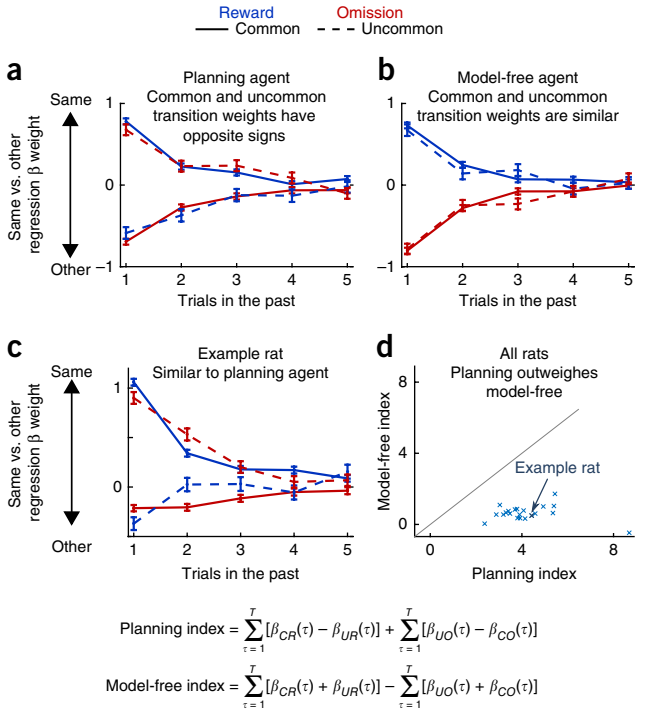

Figure 2 Multitrial history regression analysis. (**a**) Applying the analysis to simulated data from a model-based planning agent correctly shows that trials with common (solid lines) and uncommon (dashed lines) transitions produce opposite effects on the current choice. Error bars indicate standard errors of the fit regression weights. (**b**) Results of the same analysis applied to a model-free temporal difference learning agent. For this agent, there is no difference between solid and dashed lines, and current choice is driven purely by the history of rewards (blue) versus omissions (red). (**c**) Results of the analysis applied to data from an example rat. (**d**) Model-free and planning indices computed from the results of the regression analysis, shown for all rats (x marks) in the dataset ($n = 21$).

expected for a model-based strategy, but in addition, for this rat all weights are shifted in the positive direction (i.e., the 'repeat choice' direction). This particular rat's behavior can thus be succinctly described as a combination of a model-based strategy plus a tendency to repeat choices; we refer to the latter behavioral component as 'perseveration'. While the regression analysis' rich and relatively theory-neutral description of each rat's behavioral patterns is useful for identifying such deviations from a purely model-based strategy, it is limited in its ability to disentangle the extent to which each individual deviation is present in a dataset. The regression analysis suffers from several other disadvantages as well; it requires a relatively large number of parameters, and it is implausible as a generative account of the computations used by the rats to carry out the behavior (requiring an exact memory of the past five trials). We therefore turned to a complementary analytic approach: trial-by-trial model fitting using mixture-of-agents models.

Mixture-of-agents models provide both more parsimonious descriptions of each rat's dataset (involving fewer parameters) and more plausible hypotheses about the underlying generative mechanism. Each model comprises a set of agents, each deploying a different choice strategy. Rats' observed choices are modeled as reflecting the weighted influence of these agents, and fitting the model to the data means setting these weights, along with other parameters internal to the agents, so as to best match the observed behavior. We found that

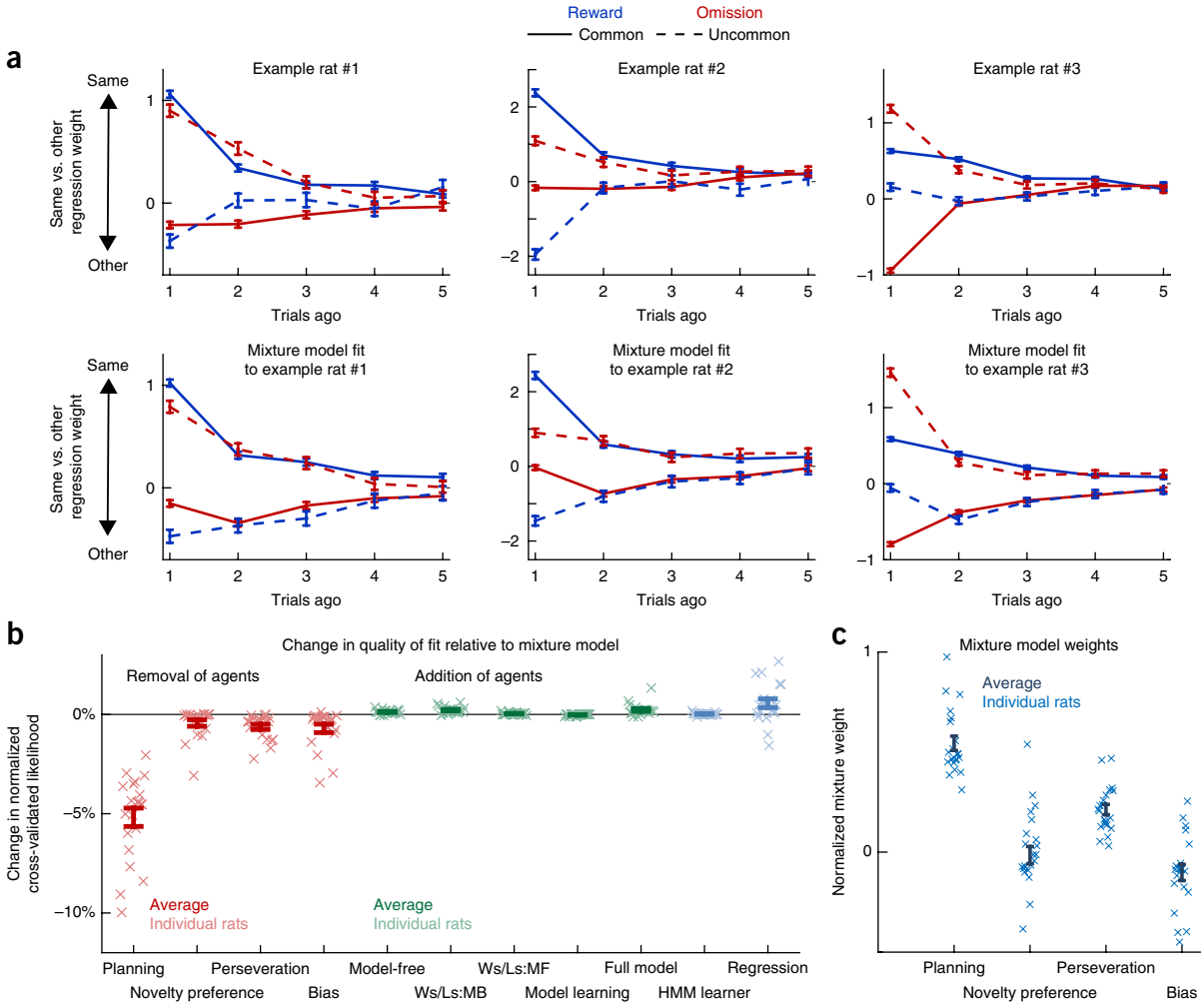

Figure 3 Model-fitting analysis. (**a**) Results of the trial-history regression analysis applied to data from three example rats (above) and simulated data produced by the agent model with parameters fit to the rats (below). Error bars indicate standard errors of the fit regression weights. (**b**) Changes in quality of fit resulting from removing (red) or adding (green) components to the reduced behavioral model ($n = 21$ rats); error bars indicate s.e.m. Ws/Ls, win–stay-vs.-lose–shift; HMM, hidden Markov model; MB, model-based; MF, model-free. (**c**) Normalized mixture weights resulting from fitting the model to rats' behavior; error bars indicate s.e.m.

a good qualitative match to rats' behavior could be achieved with a mixture of only four simple agents, representing four patterns. We call these patterns planning, choice perseveration, novelty preference and choice bias (**Fig. 3a** and **Supplementary Fig. 3**). The four agents implementing these four patterns were a model-based reinforcement learning agent (planning), an agent that repeated the previous trial's choice (perseveration), an agent that repeated or avoided choices that led to a common versus an uncommon transition (novelty preference) and an agent that preferred the left or the right choice port on each trial (choice bias; Online Methods). In all, this model contained five free parameters: four mixing weights, β_{plan}, β_{np}, β_{persev} and β_{bias}, each associated with one of the four agents, and a learning rate, α_{plan}, internal to the planning agent. We arrived at these four particular patterns as the necessary components because removing any of the four agents from the mixture resulted in a large decrease in quality of fit (assessed by cross-validated likelihood; **Fig. 3b** and Online Methods) and because adding a variety of other additional patterns (model-free reinforcement learning, model-based and model-free win–stay versus lose–switch, transition learning or all of the above; Online Methods) resulted in only negligible improvements (**Fig. 3b**), as did substituting an alternate learning mechanism based on hidden Markov models into the planning agent (**Fig. 3b** and Online Methods). We found that the mixture model performed similarly in terms of quality of fit to the regression-based model, for all but a minority of rats (**Fig. 3b**). The planning agent earned, on average, the largest mixing weights of any agent, indicating that model-based planning is the dominant component of behavior on our task (**Fig. 3c**). Taken together, these findings indicate that this mixture model is an effective tool for quantifying patterns present in our behavioral data and that well-trained rats on the two-step task exhibit perseveration, novelty preference and bias but predominantly exhibit model-based planning.

Pharmacological inactivation of hippocampus or OFC impairs planning

In the next phase of this work, we took advantage of both the regression analysis and the mixture-of-agents model to investigate the causal contribution of OFC and dH to planning behavior. We implanted six well-trained rats with infusion cannulae targeting each region bilaterally (**Supplementary Fig. 5**) and used these cannulae to perform reversible inactivation experiments. In these experiments, we infused the $GABA_A$ agonist muscimol into a target brain region bilaterally, then allowed the animals to recover for a short time before placing them in

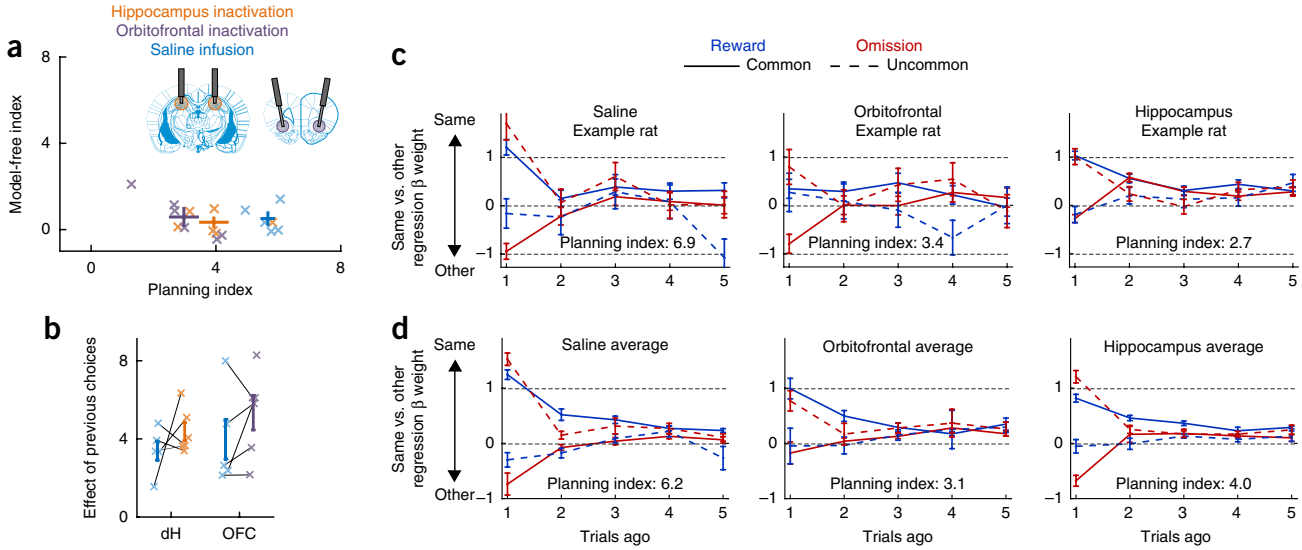

Figure 4 Effects of muscimol inactivation. (**a**) Planning index and model-free index for implanted rats ($n = 6$) performing the task on OFC inactivation sessions (purple), dH inactivation sessions (orange) and pooled saline infusions (blue; pooled for display). Inactivation of either region significantly decreases the planning index. X marks indicate individual rats; error bars show mean across rats and s.e.m. Insets show locations of infusions (dH, left; OFC, right). (**b**) Main effect of past choice on future choice during the same sessions (saline session, unpooled). Inactivation has no significant effect on this measure ($P = 0.4$ for OFC and $P = 0.7$ for dH). Error bars show mean across rats and s.e.m. (**c**) Results of the regression analysis of **Figure 2** applied to data from an example rat on saline sessions (left), OFC infusions (middle) and dH infusions (right). (**d**) Average over all rats of the regression analysis. In **c** and **d**, error bars indicate s.e.m.

the behavioral chamber to perform the task (Online Methods). We compared behavior during muscimol sessions to behavior during control sessions performed the day before and the day after inactivation, as well as to sessions before which we infused saline into the target region (**Supplementary Figs. 6–8**). We found that inactivation of either region substantially reduced the magnitude of the planning index relative to both each region's control sessions (**Fig. 4**; OFC, $P = 0.001$; dH, $P = 0.01$; Online Methods) and to pooled saline sessions (OFC, $P = 0.004$; dH, $P = 0.04$). We found no effect of inactivation on the model-free index (all $P > 0.5$). We also found that inactivation of dH resulted in decreases in task performance as measured by the fraction of times the rat chose choice ports whose common transition led to the reward port with larger reward probability ($P = 0.003$; **Supplementary Fig. 9**). For completeness, we also present results of the traditional one-trial-back analysis on the inactivation dataset (**Supplementary Figs. 10** and **11**). The impact of inactivation on model-based behavioral patterns was not simply due to an overall reduction in the modulation of current trial choices by past trials: we computed the aggregate main effect of past choices on future choices ($\beta_{CR} + \beta_{UR} + \beta_{CO} + \beta_{UO}$; Online Methods) for each rat for each type of session and found that this measure was insensitive to inactivation of either region (**Fig. 4b**; OFC, $P = 0.4$; dH, $P = 0.7$). Together, these results suggest that inactivation of OFC or dH reduced the extent to which behavior showed evidence of planning but did not affect evidence for perseveration or model-free patterns.

To determine the extent to which these muscimol-induced behavioral changes were specific to planning, we applied our mixture-of-agents model to the inactivation datasets (**Fig. 5a** and Online Methods). To make the most efficient use of our data, we adopted a hierarchical modeling approach, simultaneously estimating parameters for both each rat individually as well as for the population of rats as a whole. For each rat, we estimated the mixture-of-agents model parameters (β_{plan}, α_{plan}, β_{np}, β_{persev} and β_{bias}) for control and inactivation sessions. For the population, we estimated the distribution of each of the rat-level parameters across animals, as well as the effect of inactivation on each parameter. To perform Bayesian inference with this model, we conditioned it on the observed datasets and used Hamiltonian Markov chain Monte Carlo to estimate the posterior distribution jointly over all model parameters (Online Methods and **Supplementary Figs. 12** and **13**). We summarize this distribution by reporting the median over each parameter, taking this as our estimate for that parameter. Estimates for parameters governing behavior on control sessions were similar to those produced by fitting the model to unimplanted rats (compare **Fig. 5b** and **Table 1**). Estimates for parameters governing the effect of inactivation on performance suggested large and consistent effects on the planning parameter β_{plan}, with weak and/or inconsistent effects on other parameters. To test whether inactivation affected behavior at the population level, we computed for each population-level parameter the fraction of the posterior in which that parameter has the opposite sign as its median: the Bayesian analog of a P value. We found that this value was small only for the parameter corresponding to the planning weight (β_{plan}; OFC, $P = 0.01$; dH, $P = 0.01$) and large for all other parameters (all $P > 0.1$). To determine whether this was robust to tradeoff in parameter estimates between β_{plan} and other parameters, we inspected plots of the density of posterior samples as a function of several parameters at once. **Figure 5c** shows a projection of this multidimensional density onto axes that represent the change in β_{plan} (planning agent's weight) and the change in α_{plan} (planning agent's learning rate) due to the infusion. We found that no infusion-induced change in α_{plan} would allow a good fit to the data without a substantial reduction in the β_{plan} parameter (all of the significant density is below the 'effect on $\beta_{plan} = 0$' axis). We found similar robustness with respect to the other population-level parameters (**Supplementary Fig. 14**).

To test the hypothesis that the effects of inactivation were specific to planning, we constructed several variants of our model and compared them to one another using cross-validation. The first of these was designed to simulate a global effect of inactivation on memory

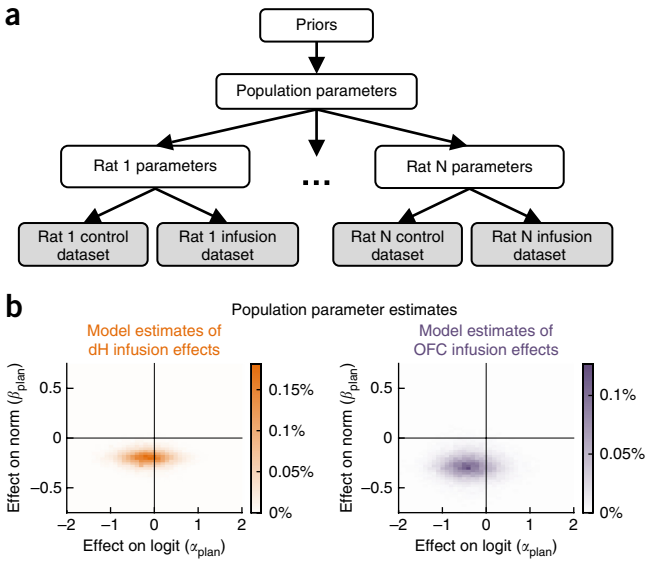

Figure 5 Effects of muscimol inactivation on mixture model fits. (**a**) Schematic showing hierarchical Bayesian framework for using the agent model for parameter estimation. Each rat is characterized by a set of control parameters governing performance in saline sessions, as well as a set of infusion effect parameters governing the change in behavior following infusion. The population of rats is characterized by the means and s.d. of each of the rat-level parameters. These population parameters are subject to weakly informative priors. (**b**) Posterior belief distributions produced by the model over the parameters governing the effect of inactivation on planning weight (β_{plan}) and learning rate (α_{plan}).

and constrained any effect on β_{plan}, β_{np} and β_{persev} to be equal across all three weights. The second was designed to simulate an effect specifically on memory for more remote past events and allowed inactivation to affect only the influence of outcomes that occurred two or more trials in the past. The third was a combination of these two, allowing inactivation to have different effects on the recent and the remote past but constraining it to affect all agents equally. We found that in all cases, model comparison strongly dispreferred these alternative models, favoring a model in which inactivation has different effects on different components of behavior (log posterior predictive ratios of 42, 56 and 47 for OFC in the first, second and third alternative models, respectively, and log posterior predictive ratios of 26, 43 and 26 for dH; Online Methods). Taken together, these findings indicate that both OFC and dH play particular roles in supporting particular behavioral patterns and that both play a specific role in model-based planning behavior. We find no evidence that either region plays a consistent role in supporting any behavioral component other than planning.

DISCUSSION

We report the first successful adaptation of the two-step task—a repeated-trial, multistep decision task widely used in human research—to rats. This development, along with parallel efforts in other labs (**Supplementary Discussion**) provides a broadly applicable tool for investigating the neural mechanisms of planning. While existing planning tasks for rodents are well-suited to identifying the neural structures involved and expose the process of model learning for study, the two-step task provides important complementary advantages. By eliciting many planned decisions in each behavioral session, it opens the door to a wide variety of new experimental designs, including those employing neural recordings to characterize

Table 1 Parameter estimates produced by the hierarchical Bayesian model for population parameters

	Saline	OFC effects	dH effects
Normalized planning (β_{plan})	0.73	–0.28*	–0.19*
Normalized novelty preference (β_{np})	0.09	–0.13	0.02
Normalized perseveration (β_{persev})	0.21	–0.02	–0.04
Bias (β_{bias})	0.09	0.17	0.05
Logit learning rate (α_{plan})	–0.38	–0.39	–0.34

Column one shows parameters governing behavior on saline sessions. Columns two and three show parameters governing the change in performance due to OFC or dH inactivation. In columns two and three, asterisks indicate parameters for which 95% or more of the posterior distribution shares the same sign.

the neural correlates of planning, as well as those, like ours, employing trial-by-trial analysis to quantify the relative influence of planning versus other behavioral strategies.

Analysis of choice behavior on our task reveals a dominant role for model-based planning. Notably, our analysis reveals little or no role for model-free reinforcement learning, in contrast with the performance of human subjects on the same task[9]. One possible reason for this is the extensive experience our rat subjects have with the task; human subjects given several sessions of training tend, like our rats, to adopt a predominantly model-based strategy[35]. These data stand in tension with theoretical accounts suggesting that model-based control is a slower, more costly or less reliable alternative to model-free control and should be avoided when it does not lead to a meaningful increase in reward rates[5,36]. However, they are in accord with data showing that human subjects adopt model-based strategies even when this does not result in an increase in reward rate[37]. Together, these data suggest that model-based control may be a default decision-making strategy adopted in the face of complex environments. Notably, rats also revealed knowledge of action–outcome contingencies in their movement times (**Supplementary Fig. 3**), making it unlikely that they were using any model-free strategy, including one that might use an alternative state space to allow it to mimic model-based choice[38] (**Supplementary Discussion**).

We found that reversible inactivation of OFC selectively impaired model-based choice, consistent with previous work indicating causal roles for this region in model-based control[31–33], as well as theoretical accounts positing a role for this structure in model-based processing and economic choice[39–41]. That we observed similar effects in the rat two-step task is an important validation of this behavior as an assay of planning in the rat. Not all accounts of OFC's role in model-based processing are consistent with a causal role in instrumental choice[42]. Our findings here are therefore not merely confirmatory but also help adjudicate between competing accounts of OFC function.

Inactivation of dH also selectively impaired model-based control, leaving other behavioral patterns unchanged. This finding offers the first causal demonstration, using a well-controlled task in which planning can be clearly identified, of a long-hypothesized role in planning for hippocampus. Long-standing theories of hippocampal function[13] hold that it represents a cognitive map of physical space and that this map is used in navigational planning. Classic causal data indicate that hippocampus is necessary for tasks that require navigation[14,15] but do not speak to the question of its involvement specifically in planning. Such data are consistent with theoretical accounts in which hippocampus provides access to abstract spatial state information (i.e., location) as well as abstract spatial actions (for example, 'run south',

independent of present orientation)[43]. This information might be used by a strategy based on either action–outcome associations (i.e., a planning strategy) or on stimulus–response associations (a model-free strategy). An example of this comes from experiments using the elevated plus maze[14], in which a rat with an intact hippocampus might adopt a strategy of running south at the intersection, independent of starting location, either because it knows that this action will lead to a particular location in the maze (planning) or because it has learned a stimulus–response mapping between this location and this spatial action. A related literature argues that the hippocampus is important for working memory, citing hippocampal impairments in tasks such as delayed alternation and foraging in radial arm mazes[44,45], in which decisions must be made on the basis of recent past events. Impairments on these tasks are consistent both with accounts in which information about the recent past is used in model-free learning (i.e., generalized stimulus–response learning in which the 'stimulus' might be a memory) as well as with accounts in which it supports action–outcome planning in particular. We find that our data are less well explained by models in which inactivation impairs memory in general. This indicates that, if the role of the hippocampus in our task is to support memory, this is a particular type of memory that is specifically accessible for the purposes of planning.

Our results are in accord with theoretical accounts that posit a role for the hippocampus in planning[19–21] but stand in tension with data from classic causal experiments. These experiments have demonstrated intact action–outcome behaviors following hippocampal damage in a variety of spatial and nonspatial assays. One prominent example is latent learning, in which an animal that has previously been exposed to a maze learns to navigate a particular path through that maze more quickly than a naive animal, whether or not it has an intact hippocampus[22,23,27]. Hippocampal damage also has no impact on classic assays of an animal's ability to infer causal structure in the world, including contingency degradation, outcome devaluation and sensory preconditioning[24–26]. A comparison of these assays to our behavior reveals one potentially key difference: only the two-step task requires the chaining together of multiple action–outcome associations. Outcome devaluation, for example, requires one action–outcome association (for example, lever–food), as well as the evaluation of an outcome (food–utility). Our task requires two action–outcome associations (for example, top-left poke–bottom-right port lights; bottom-right poke–water) as well an evaluation (water–utility). This difference suggests a possible resolution: perhaps the hippocampus is necessary specifically in cases where planning requires linking actions to outcomes over multiple steps. This function may be related to the known causal role of hippocampus in relational memory tasks[28,29], which require chaining together multiple stimulus–stimulus associations. It may also be related to data indicating a role in second-order classical conditioning[46], as well as in trace conditioning[47]. Future work should investigate whether it is indeed the multistep nature of the two-step task, rather than some other feature, that renders it hippocampus-dependent.

Another contentious question about the role of hippocampus regards the extent to which it is specialized for spatial navigation[48], as opposed to playing some more general role in cognition[49,50]. While performing the two-step task does require moving through space, the key relationships necessary for planning on this task are nonspatial, namely the causal relationships linking the first-step choice to the second-step outcome. Once the first-step choice was made, lights in each subsequent port guided the animal through the remainder of the trial; apart from the single initial left–right choice, no navigation or knowledge of spatial relationships was necessary. Taken together with the literature, our results suggest that multistep planning specifically may depend on the hippocampus, in the service of both navigation and other behaviors.

Model-based planning is a process that requires multiple computations. Notably, our results do not reveal the particular causal role within the model-based system that is played by either hippocampus or OFC. A remaining open question is whether these regions perform computations involved in the planning process *per se* (i.e., actively using an action–outcome model to inform choice) or instead perform computations that are specifically necessary to support planning (for example, planning-specific forms of learning or memory). It is our hope that future studies employing the rat two-step task, perhaps in concert with electrophysiology and/or optogenetics, will be able to shed light on these and other important questions about the neural mechanisms of planning.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Erlich, C. Kopec, C.A. Duan, T. Hanks and A. Begelfer for training K.J.M. in the techniques necessary to carry out these experiments, as well as for comments and advice on the project. We thank N. Daw, I. Witten, Y. Niv, B. Wilson, T. Akam, A. Akrami and A. Solway for comments and advice on the project, and we thank J. Teran, K. Osorio, A. Sirko, R. LaTourette, L. Teachen and S. Stein for assistance in carrying out behavioral experiments. We especially thank T. Akam for suggestions on the physical layout of the behavior box and other experimental details. We thank A. Bornstein, B. Scott, A. Piet and L. Hunter for comments on the manuscript. K.J.M. was supported by training grant NIH T-32 MH065214 and by a Harold W. Dodds fellowship from Princeton University.

AUTHOR CONTRIBUTIONS

K.J.M., M.M.B. and C.D.B. conceived the project. K.J.M. designed and carried out the experiments and the data analysis, with supervision from M.M.B. and C.D.B. K.J.M., M.M.B. and C.D.B. wrote the paper, starting from an initial draft by K.J.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: an Introduction* (MIT Press, 1998).
2. Tolman, E.C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
3. Dolan, R.J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
4. Balleine, B.W. & O'Doherty, J.P. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69 (2010).
5. 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).
6. Brogden, W.J. Sensory pre-conditioning. *J. Exp. Psychol.* **25**, 323–332 (1939).
7. Adams, C.D. & Dickinson, A. Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. B* **33**, 109–121 (1981).
8. Hilário, M.R.F., Clouse, E., Yin, H.H. & Costa, R.M. Endocannabinoid signaling is critical for habit formation. *Front. Integr. Neurosci.* **1**, 6 (2007).
9. 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).
10. Simon, D.A. & Daw, N.D. Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* **31**, 5526–5539 (2011).
11. 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).

12. Huys, Q.J.M. *et al.* Interplay of approximate planning strategies. *Proc. Natl. Acad. Sci. USA* **112**, 3098–3103 (2015).
13. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon Press Oxford, 1978).
14. Packard, M.G. & McGaugh, J.L. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.* **65**, 65–72 (1996).
15. Morris, R.G., Garrud, P., Rawlins, J.N. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
16. 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).
17. Wikenheiser, A.M. & Redish, A.D. Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294 (2015).
18. Pfeiffer, B.E. & Foster, D.J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
19. Koene, R.A., Gorchetchnikov, A., Cannon, R.C. & Hasselmo, M.E. Modeling goal-directed spatial navigation in the rat based on physiological data from the hippocampal formation. *Neural Netw.* **16**, 577–584 (2003).
20. Foster, D.J. & Knierim, J.J. Sequence learning and the role of the hippocampus in rodent navigation. *Curr. Opin. Neurobiol.* **22**, 294–300 (2012).
21. Pezzulo, G., van der Meer, M.A.A., Lansink, C.S. & Pennartz, C.M.A. Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.* **18**, 647–657 (2014).
22. Kimble, D.P. & BreMiller, R. Latent learning in hippocampal-lesioned rats. *Physiol. Behav.* **26**, 1055–1059 (1981).
23. Kimble, D.P., Jordan, W.P. & BreMiller, R. Further evidence for latent learning in hippocampal-lesioned rats. *Physiol. Behav.* **29**, 401–407 (1982).
24. Corbit, L.H. & Balleine, B.W. The role of the hippocampus in instrumental conditioning. *J. Neurosci.* **20**, 4233–4239 (2000).
25. Corbit, L.H., Ostlund, S.B. & Balleine, B.W. Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *J. Neurosci.* **22**, 10976–10984 (2002).
26. Ward-Robinson, J. *et al.* Excitotoxic lesions of the hippocampus leave sensory preconditioning intact: implications for models of hippocampal function. *Behav. Neurosci.* **115**, 1357–1362 (2001).
27. Gaskin, S., Chai, S.-C. & White, N.M. Inactivation of the dorsal hippocampus does not affect learning during exploration of a novel environment. *Hippocampus* **15**, 1085–1093 (2005).
28. Bunsey, M. & Eichenbaum, H. Conservation of hippocampal memory function in rats and humans. *Nature* **379**, 255–257 (1996).
29. Dusek, J.A. & Eichenbaum, H. The hippocampus and memory for orderly stimulus relations. *Proc. Natl. Acad. Sci. USA* **94**, 7109–7114 (1997).
30. Devito, L.M. & Eichenbaum, H. Memory for the order of events in specific sequences: contributions of the hippocampus and medial prefrontal cortex. *J. Neurosci.* **31**, 3169–3175 (2011).
31. Jones, J.L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953–956 (2012).
32. McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y. & Schoenbaum, G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705 (2011).
33. Gremel, C.M. & Costa, R.M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* **4**, 2264 (2013).
34. Miller, K.J., Brody, C.D. & Botvinick, M.M. Identifying model-based and model-free patterns in behavior on multi-step tasks. Preprint at http://www.biorxiv.org/content/early/2016/12/24/096339 (2016).
35. Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R.J. Model-based reasoning in humans becomes automatic with training. *PLOS Comput. Biol.* **11**, e1004463 (2015).
36. Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Comput. Biol.* **7**, e1002055 (2011).
37. Kool, W., Cushman, F.A. & Gershman, S.J. When does model-based control pay off? *PLOS Comput. Biol.* **12**, e1005090 (2016).
38. Akam, T., Costa, R. & Dayan, P. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLOS Comput. Biol.* **11**, e1004648 (2015).
39. Padoa-Schioppa, C. Neurobiology of economic choice: a good-based model. *Annu. Rev. Neurosci.* **34**, 333–359 (2011).
40. Wilson, R.C., Takahashi, Y.K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
41. Stalnaker, T.A., Cooch, N.K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nat. Neurosci.* **18**, 620–627 (2015).
42. Ostlund, S.B. & Balleine, B.W. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. *J. Neurosci.* **27**, 4819–4825 (2007).
43. Foster, D.J., Morris, R.G. & Dayan, P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).
44. Olton, D.S., Becker, J.T. & Handelmann, G.E. Hippocampus, space, and memory. *Behav. Brain Sci.* **2**, 313–322 (1979).
45. Racine, R.J. & Kimble, D.P. Hippocampal lesions and delayed alternation in the rat. *Psychon. Sci.* **3**, 285–286 (1965).
46. Gilboa, A., Sekeres, M., Moscovitch, M. & Winocur, G. Higher-order conditioning is impaired by hippocampal lesions. *Curr. Biol.* **24**, 2202–2207 (2014).
47. Solomon, P.R., Vander Schaaf, E.R., Thompson, R.F. & Weisz, D.J. Hippocampus and trace conditioning of the rabbit's classically conditioned nictitating membrane response. *Behav. Neurosci.* **100**, 729–744 (1986).
48. Hartley, T., Lever, C., Burgess, N. & O'Keefe, J. Space in the brain: how the hippocampal formation supports spatial cognition. *Phil. Trans. R. Soc. Lond. B* **369**, 20120510 (2013).
49. Hassabis, D., & Maguire, E.A. Deconstructing episodic memory with construction. *Trends in Cog. Sci.*, **11**, 299–306 (2007).
50. Eichenbaum, H. & Cohen, N.J. Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron* **83**, 764–770 (2014).

ONLINE METHODS

Subjects. All subjects were adult male Long-Evans rats (Taconic Biosciences, NY), placed on a restricted water schedule to motivate them to work for water rewards. Some rats were housed on a reverse 12-h light cycle and others on a normal light cycle; in all cases, rats were trained during the dark phase of their cycle. Rats were pair-housed during behavioral training and then singly housed after being implanted with cannula. Animal use procedures were approved by the Princeton University Institutional Animal Care and Use Committee and carried out in accordance with NIH standards. One infusion rat was removed from the study before completion due to health reasons; this rat did not complete any saline sessions.

The number of animals used in the inactivation experiment was determined informally by comparison to similar previous studies and by resources available. Particular animals were selected for inclusion informally: they were the first three in each transition probability condition to complete training on the present version of the task, with high trial counts per session. Example animals (**Figs. 2c**, **3a** and **4c**) were selected on the basis of cleanly demonstrating effects that were consistent in the population. Corresponding plots for all animals can be found in **Supplementary Figures 4** and **6**.

Behavioral apparatus. Rats performed the task in custom behavioral chambers (Island Motion, NY) located inside sound- and light-attenuated boxes (Coulborn Instruments, PA). Each chamber was outfitted with six nose ports, arranged in two rows of three, and with a pair of speakers for delivering auditory stimuli. Each nose port contained a white light emitting diode (LED) for delivering visual stimuli, as well as an infrared LED and infrared phototransistor for detecting rats' entries into the port. The left and right ports in the bottom row also contained sipper tubes for delivering water rewards. Rats were placed into and removed from training chambers by technicians blind to the experiment being run.

Training pipeline. Here we outline a procedure suitable for efficiently training naive rats on the two-step task. Automated code for training rats using this pipeline via the bControl behavioral control system can be downloaded from the Brody lab website (http://brodylab.org/code/two-step-planning-task-code/). This formalization of our training procedure into a software pipeline should also facilitate efforts to replicate our task in other labs, as the pipeline can readily be downloaded and identically re-run.

Phase I: sipper tube familiarization. In this phase, rats become familiar with the experimental apparatus and learn to approach the reward ports when they illuminate. Trials begin with the illumination of the LED in one of the two reward ports, and reward is delivered upon port entry. Training in this phase continues until the rat is completing an average of 200 or more trials per day.

Phase II: trial structure familiarization. In this phase, rats must complete all four actions of the complete task, with rewards delivered on each trial. Trials begin with the illumination of the LED in the top center port, which the rat must enter. Upon entry, one of the side ports (chosen randomly by the computer) will illuminate, and the rat must enter it. Once the rat does this, the LED in the bottom center port illuminates, and a sound begins to play indicating which of the bottom side ports will ultimately be illuminated (according to the 80%/20% transition probabilities for that rat). The rat must enter the lit bottom center port, which causes the appropriate bottom side port to illuminate. Upon entry into this side port, the rat receives a reward on every trial. For rats in the congruent condition, the reward port available will be on the same side as the choice port selected 80% of the time, while for rats in the incongruent condition, ports will match in this way 20% of the time. 'Violation trials' occur whenever the rat enters a port that is not illuminated, which results in a 5-s timeout and an aversive white noise sound. Training in this phase continues until the rat is completing an average of 200 or more trials per day with a rate of violation trials less than 5%.

Phase IIIa: performance-triggered flips. In this phase, probabilistic dynamic rewards are introduced, and rats must learn to choose the choice port that is associated with the reward port that currently has higher reward probability. Trial structure is as in phase II, except that in 90% of trials both choice ports illuminate after the rat enters the top center port, and the rat must decide which choice port to enter. The rat then receives an auditory cue and LED instructions to enter the bottom center port and one of the reward ports, as above. This phase consists of blocks, and in each block, one of the reward ports is 'good' and the other is 'bad'. If the good reward port is illuminated, the rat will receive a water reward for entering it 100% of the time. If the bad reward port is illuminated, the rat must enter it to move on to the next trial, but no water will be delivered. Which reward port is good and which is bad changes in blocks, and the change in blocks is enabled by the rat's performance. Each block lasts a minimum of 50 trials, after which the block switch is 'enabled' if the rat has selected the choice port that leads most often to the good reward port on 80% of free choices in the last 50 trials. On each trial after the end is enabled, there is a 10% chance per trial that the block will actually switch, and the reward ports will flip their roles. Phase IIIa lasts until rats achieve an average of three to four block switches per session for several sessions in a row. Rats that show a decrease in trial count during this phase can often be remotivated by using small rewards (~10% of the usual reward volume) in place of reward omissions at the bad port.

Phases IIIb and **IIIc.** These phases are the same as phase IIIa, except that the good and bad reward ports are rewarded 90% and 10% of the time, respectively, in phase IIIb and 80% and 20% of the time in phase IIIc. Block flips are triggered by the rat's performance, as above. Each of these phases lasts until the rat achieves an average of two to three block changes per session for several sessions in a row.

Phase IV: final task. The final task is the same as phase IIIc, except that changes in block are no longer triggered by the performance of the rat but occur stochastically. Each block has a minimum length of 10 trials, after which the block has a 2% chance of switching on each trial. In our experience, approximately 90% of rats succeed in reaching the final task.

Behavioral analysis. We quantified the effect of past trials and their outcomes on future decisions using a logistic regression analysis based on previous trials and their outcomes[51]. We define vectors for each of the four possible trial outcomes: common-reward (CR), common-omission (CO), uncommon-reward (UR) and uncommon-omission (UO), each taking on a value of +1 for trials of their type in which the rat selected the left choice port, a value of −1 for trials of their type in which the rat selected the right choice port and a value of 0 for trials of other types. We define the following regression model:

$$\log\left(\frac{P_{\text{left}}(t)}{P_{\text{right}}(t)}\right) = \sum_{\tau=1}^{T} \beta_{\text{CR}}(\tau)*\text{CR}(t-\tau) + \sum_{\tau=1}^{T} \beta_{\text{CO}}(\tau)*\text{CO}(t-\tau) + \sum_{\tau=1}^{T} \beta_{\text{UR}}(\tau)*\text{UR}(t-\tau) + \sum_{\tau=1}^{T} \beta_{\text{UO}}(\tau)*\text{UO}(t-\tau) \quad (1)$$

where β_{CR}, β_{CO}, β_{UR} and β_{UO} are vectors of regression weights that quantify the tendency to repeat on the next trial a choice that was made τ trials ago and resulted in the outcome of their type, and T is a hyperparameter governing the number of past trials used by the model to predict upcoming choice. Unless otherwise specified, T was set to 5 for all analyses (**Supplementary Fig. 15**).

We expect model-free agents to show a pattern of repeating choices that lead to reward and switching away from those that lead to omissions, so we define a model-free index for a dataset as the sum of the appropriate weights from a regression model fit to that dataset:

$$\text{Model-free Index} = \sum_{\tau=1}^{T} \left[\beta_{\text{CR}}(\tau) + \beta_{\text{UR}}(\tau)\right] - \sum_{\tau=1}^{T} \left[\beta_{\text{UO}}(\tau) + \beta_{\text{CO}}(\tau)\right] \quad (2)$$

We expect that planning agents will show the opposite pattern after uncommon transition trials, since the uncommon transition from one choice is the common transition from the other choice. We define a planning index:

$$\text{Planning Index} = \sum_{\tau=1}^{T} \left[\beta_{\text{CR}}(\tau) - \beta_{\text{UR}}(\tau)\right] + \sum_{\tau=1}^{T} \left[\beta_{\text{UO}}(\tau) - \beta_{\text{CO}}(\tau)\right] \quad (3)$$

We test for significant model-free and planning indices using a one-sample t test across rats. We test for significant differences between rats in the common-congruent and the common-incongruent conditions using a two-sample t test.

Behavior models. We model our rats' behavior using a mixture-of-agents approach, in which each rat's behavior is described as resulting from the influence of a weighted average of several different 'agents' implementing different behavioral strategies to solve the task. On each trial, each agent A computes a value, $Q_A(a)$, for each of the two available actions a, and the combined model

makes a decision according to a weighted average of the various strategies' values, $Q_{total}(a)$:

$$Q_{\text{total}}(a) = \sum_{A \in \{\text{agents}\}} \beta_A Q_A(a) \quad \text{and} \quad \pi(a) = \frac{e^{Q_{\text{total}}(a)}}{\sum_{a'} e^{Q_{\text{total}}(a')}} \tag{4}$$

where each β is a weighting parameter determining the influence of each agent, and $Q(a)$ is the probability that the mixture-of-agents will select action a on that trial. We considered models consisting of subsets of the seven following agents: model-based temporal difference learning, model-free temporal difference learning, model-based win–stay/lose–switch, model-free win–stay/lose–switch, common–stay/uncommon–switch, perseveration and bias. The 'full model' consists of all of these agents, while the 'reduced model' consists of four agents, which were found to be sufficient to provide a good match to rat behavior. These four were model-based temporal difference learning (without transition updating), novelty preference, perseveration and bias.

Model-based temporal difference learning. Model-based temporal difference learning is a planning strategy that maintains separate estimates of the probability with which each action (selecting the left or the right choice port) will lead to each outcome (the left or the right reward port becoming available), $T(a,o)$, as well as the probability, $R_{\text{plan}}(o)$, with which each outcome will lead to reward. This strategy assigns values to the actions by combining these probabilities to compute the expected probability with which selection of each action will ultimately lead to reward:

$$Q_{\text{plan}}(a) = \sum_o R_{\text{plan}}(o) * T(a,o) \tag{5}$$

At the beginning of each session, the reward estimate $R_{\text{plan}}(o)$ is initialized to 0.5 for both outcomes, and the transition estimate $T(a,o)$ is initialized to the true transition function for the rat being modeled (0.8 for common and 0.2 for uncommon transitions). After each trial, the reward estimate for both outcomes is updated according to

$$R_{\text{plan}}(o) \leftarrow \begin{cases} R_{\text{plan}}(o) + \alpha_{\text{plan}}\left(r_t - R_{\text{plan}}(o)\right), & o = o_t \\ R_{\text{plan}}(o) + \alpha_{\text{plan}}\left(1 - r_t - R_{\text{plan}}(o)\right), & o \neq o_t \end{cases} \tag{6}$$

where o_t is the outcome that was observed on that trial, r_t is a binary variable indicating reward delivery, and α_{plan} is a learning-rate parameter. The full model (but not the reduced model) also included transition learning, in which the function $T(a,o)$ is updated after each outcome according to

$$T(a_t,o) \leftarrow \begin{cases} T(a_t,o) + \alpha_T\left(1 - T(a_t,o)\right), & o = o_t \\ T(a_t,o) + \alpha_T\left(0 - T(a_t,o)\right), & o \neq o_t \end{cases} \tag{7}$$

where a_t is the action taken, and α_T is a learning rate parameter.

Model-free temporal difference learning. Model-free temporal difference learning is a nonplanning reward-based strategy. It maintains an estimate of the value of the choice ports, $Q_{\text{mf}}(a)$, as well as an estimate of the values of the reward ports, $R_{\text{mf}}(o)$. After each trial, these quantities are updated according to

$$Q_{\text{mf}}(a_t) \leftarrow Q_{\text{mf}}(a_t) + \alpha_{\text{mf}}\left(R_{\text{mf}}(o_t) - Q_{\text{mf}}(a_t)\right) + \alpha_{\text{mf}}\lambda\left(r_t - R_{\text{mf}}(o_t)\right)$$

$$R_{\text{mf}}(o) \leftarrow \begin{cases} R_{\text{mf}}(o) + \alpha_{\text{mf}}\left(r_t - R_{\text{mf}}(o)\right), & o = o_t \\ R_{\text{mf}}(o) + \alpha_{\text{mf}}\left(1 - r_t - R_{\text{mf}}(o)\right), & o \neq o_t \end{cases} \tag{8}$$

where α_{mf} and λ are learning-rate and eligibility-trace parameters affecting the update process.

Model-free win–stay/lose–switch. Win–stay/lose–switch is a pattern that tends to repeat choices that led to rewards on the previous trial and switch away from choices that led to omissions. It calculates its values on each trial according to the following:

$$Q_{\text{wsls–mf}}(a_t) \leftarrow r_t \quad \text{and} \quad Q_{\text{wsls–mf}}(a \neq a_t) \leftarrow 1 - r_t \tag{9}$$

Model-based win–stay/lose–switch. Model-based win–stay/lose–switch follows the win–stay/lose–switch pattern after common transition trials but inverts it after uncommon transition trials.

$$Q_{\text{wsls-mb}}(a_t) \leftarrow \begin{cases} r_t, \text{common transition trials} \\ 1 - r_t, \text{uncommon transition trials} \end{cases}$$

$$Q_{\text{wsls-mb}}(a \neq a_t) \leftarrow 1 - Q_{\text{wsls-mb}}(a_t) \tag{10}$$

Novelty preference. The novelty preference agent follows an uncommon–stay/common–switch pattern, which tends to repeat choices when they lead to uncommon transitions on the previous trial and to switch away from them when they lead to common transitions. Note that some rats have positive values of the β_{np} parameter weighting this agent (novelty preferring) while others have negative values (novelty averse; **Fig. 3c**):

$$Q_{\text{np}}(a_t) \leftarrow \begin{cases} 1, \text{ common transition trials} \\ 0, \text{ uncommon transition trials} \end{cases}$$

$$Q_{\text{np}}(a \neq a_t) \leftarrow 1 - Q_{np}(a_t) \tag{11}$$

Perseveration. Perseveration is a pattern that tends to repeat the choice that was made on the previous trial, regardless of whether it led to a common or an uncommon transition and regardless of whether or not it led to reward.

$$Q_{\text{persev}}(a_t) \leftarrow 1$$

$$Q_{\text{persev}}(a \neq a_t) \leftarrow 0 \tag{12}$$

Bias. Bias is a pattern that tends to select the same choice port on every trial. Its value function is therefore static, with the extent and direction of the bias being governed by the magnitude and sign of this strategy's weighting parameter, β_{bias}.

$$Q_{\text{bias}}(\text{left}) \leftarrow 1$$

$$Q_{\text{bias}}(\text{right}) \leftarrow -1 \tag{13}$$

Model comparison and parameter estimation: unimplanted rats. We implemented the model described above using the probabilistic programming language Stan[52,53] and performed maximum a posteriori fits using weakly informative priors on all parameters[54]. The prior over the weighting parameters β was normal, with mean 0 and sd 0.5, and the prior over α_{mf}, α_{mb} and λ was a beta distribution with $a = b = 3$.

To perform model comparisons, we used two-fold cross-validation, dividing our dataset for each rat into even- and odd-numbered sessions and computing the log-likelihood of each partial dataset using parameters fit to the other. For each model for each rat, we computed the normalized cross-validated likelihood by summing the log-likelihoods for the even- and odd-numbered sessions, dividing by the total number of trials and exponentiating. This value can be interpreted as the average per-trial likelihood with which the model would have selected the action that the rat actually selected. We define the reduced model as the full model defined above, with the parameters β_{mf}, $\beta_{\text{wsls – mf}}$, $\beta_{\text{wsls – mb}}$ and α_T all set to zero, leaving as free parameters β_{plan}, α_{plan}, β_{np}, β_{persev} and β_{bias} (note that α_{mf} and λ become undefined when $\beta_{\text{mf}} = 0$). We compared this reduced model to nine alternative models: four in which we allowed one of the fixed parameters to vary freely, four in which we fixed one of the free parameters β_{plan}, β_{np}, β_{persev} or β_{bias} to zero, and the full model, in which all parameters are allowed to vary.

We estimated parameters by fitting the reduced model to the entire dataset generated by each rat (as opposed to the even/odd split used for model comparison), using maximum a posteriori fits under the same priors. For ease of comparison, we normalize the weighting parameters β_{plan}, β_{np} and β_{persev}, dividing each by the s.d. of its agent's associated values (Q_{plan}, Q_{np} and Q_{persev}) taken

doi:10.1038/nn.4613

across trials. Since each weighting parameter affects behavior only by scaling the value output by its agent, this technique brings the weights into a common scale and facilitates interpretation of their relative magnitudes, analogous to the use of standardized coefficients in regression models.

Synthetic behavioral datasets: unimplanted rats. To generate synthetic behavioral datasets, we took the maximum a posteriori estimates parameter estimates for each rat and used the reduced model in generative mode. The model matched to each rat received the same number of trials as that rat, as well as the same sequence of reward probabilities. We used these synthetic datasets for qualitative model-checking: if the reduced model does a good job capturing patterns in behavior, applying the regression analysis to both real and synthetic datasets should yield similar results.

Surgery. We implanted 6 rats with infusion cannula targeting dH, OFC and prelimbic cortex (PL), using standard stereotaxic techniques (data from PL are not reported in this paper). Anesthesia was induced using isoflurane, along with injections of ketamine and buprenorphine. The head was shaved, and the rat was placed in a stereotaxic frame (Kopf Instruments) using nonpuncture ear bars. Lidocaine was injected subcutaneously under the scalp for local anesthesia and to reduce bleeding. An incision was made in the scalp, the skull was cleaned of tissue and bleeding was stopped. Injection cannula were mounted into guide cannula held in stereotaxic arms (dH and OFC: 22-gauge guide, 28-gauge injector; PL: 26-gauge guide, 28-gauge injector; Plastics One, VA), while a separate arm held a fresh sharp needle. The locations of bregma and interaural zero were measured with the tip of each injector and with the needle tip. Craniotomies were performed at each target site, and a small durotomy was made by piercing the dura with the needle. The skull was covered with a thin layer of C&B Metabond (Parkell Inc., NY), and the cannula were lowered into position one at a time. Target locations relative to bregma were AP −3.8, ML ±2.5 and DV −3.1 for dH; AP +3.2, ML ±0.7 and DV −3.2 for PL; and AP +3.5, ML ±2.5 and DV −5 for OFC. Orbitofrontal cannula were implanted at a 10° lateral angle to make room for the prelimbic implant. Cannula were fixed to the skull using Absolute Dentin (Parkell Inc., NY), and each craniotomy was sealed with Kwik-Sil elastomer (World Precision Instruments, FL). Once all cannula were in place, Duralay dental acrylic (Reliance Dental, IL) was applied to secure the implant. The injector was removed from each guide cannula and replaced with a dummy cannula. Rats were treated with Ketofen 24 and 48 h postoperative and allowed to recover for at least 7 d before returning to water restriction and behavioral training.

Inactivation experiments. Each day of infusions, an injection system was prepared with the injection cannula for one brain region. The injection cannula was attached to a silicone tube, and both were filled with light mineral oil. A small amount of distilled water was injected into the other end of the tube to create a visible water–oil interface, and this end was attached to a Hamilton syringe (Hamilton Company, NV) filled with distilled water. This system was used to draw up and let out small volumes of muscimol solution, and we inspected it to ensure that it was free of air bubbles.

Rats were placed under light isoflurane anesthesia, and the dummy cannula were removed from the appropriate guide cannula. The injector was placed into the guide and used to deliver 0.3 µL of 0.25 mg/mL muscimol[55,56] solution over the course of 90 s. The injector was left in place for 4 min for the solution to diffuse, and then the procedure was repeated in the other hemisphere. For saline control sessions, the same procedure was used, but sterile saline was infused in place of muscimol solution. The experimenter was not blind to the region (OFC, dH or PL) or substance (muscimol or saline) being infused. After the completion of the bilateral infusion, rats were taken off of isoflurane, placed back in their home cages and allowed to recover for 30–60 min before being placed in the behavioral chamber to perform the task.

Analysis of inactivation data. For each rat, we considered five types of sessions: OFC muscimol, dH muscimol, OFC control, dH control and saline. Control sessions were performed the day before and the day after each infusion session, and saline sessions were pooled across OFC saline infusions and dH saline infusions (OFC muscimol, 18 sessions; OFC control, 36 sessions; OFC saline, 6 sessions; dH muscimol, 33 sessions; dH control, 64 sessions; dH saline, 10 sessions). Our dataset for each session consisted of up to the first 400 trials of each session in which at least 50 trials were performed. We performed regression analysis (equation (1)) and computed the model-free index and planning index (equations (2) and (3)) for each dataset. To compute *P* values, we performed a paired *t* test across rats on the difference between muscimol and control datasets for each region and on the difference between muscimol infusion in each region and the pooled saline infusion datasets.

Modeling inactivation data. We constructed a hierarchical Bayesian version of our reduced model, using the probabilistic programming language Stan[52,53,57,58]. This model considered two datasets from each rat simultaneously: an inactivation and a control dataset. Each of these datasets is modeled as the output of the reduced model (see the "Behavioral models" section, above), which takes the five parameters β_{plan}, α_{plan}, β_{np}, β_{persev} and β_{bias}, giving each rat ten parameters: five for the control dataset and five for the infusion dataset. For the hierarchical model, we reparameterize these, characterizing each rat *R* by ten parameters organized into two vectors, $\theta_{\mathbf{R}} = \theta_{\text{R}}^1 \ldots \theta_{\text{R}}^5$ and $\Delta_{\mathbf{R}} = \Delta_{\text{R}}^1 \ldots \Delta_{\text{R}}^5$, according to the following mapping:

- For rat *R* dataset:
  - $\text{Norm}(\beta_{\text{plan}}) = \theta_{\text{R}}^1$
  - $\text{Logit}(\alpha_{\text{plan}}) = \theta_{\text{R}}^2$
  - $\text{Norm}(\beta_{\text{np}}) = \theta_{\text{R}}^3$
  - $\text{Norm}(\beta_{\text{persev}}) = \theta_{\text{R}}^4$
  - $\beta_{\text{bias}} = \theta_{\text{R}}^5$
- For rat *R* infusion dataset:
  - $\text{Norm}(\beta_{\text{plan}}) = \theta_{\text{R}}^1 + \Delta_{\text{R}}^1$
  - $\text{Logit}(\alpha_{\text{plan}}) = \theta_{\text{R}}^2 + \Delta_{\text{R}}^2$
  - $\text{Norm}(\beta_{\text{np}}) = \theta_{\text{R}}^3 + \Delta_{\text{R}}^3$
  - $\text{Norm}(\beta_{\text{persev}}) = \theta_{\text{R}}^4 + \Delta_{\text{R}}^4$
  - $\beta_{\text{bias}} = \theta_{\text{R}}^5 + \Delta_{\text{R}}^5$

where 'norm' indicates normalization of the weight (see the "Parameter estimation" section, above), and 'logit' indicates the inverse-sigmoid logit function, which transforms a parameter bounded at 0 and 1 into a parameter with support over all real numbers.

The values in $\boldsymbol{\theta}_{\mathbf{R}}$ and $\boldsymbol{\Delta}_{\mathbf{R}}$ adopted by a particular rat are modeled as draws from a Gaussian distribution governed by population-level parameter vectors $\boldsymbol{\theta}_{\boldsymbol{\mu}}$, $\boldsymbol{\theta}_{\boldsymbol{\sigma}}$, $\boldsymbol{\Delta}_{\boldsymbol{\mu}}$ and $\boldsymbol{\Delta}_{\boldsymbol{\sigma}}$, giving the mean and s.d. of the distribution of each of the rat-level parameters in the population:

$$\theta_{\text{R}}^m \sim \text{Normal}\left(\theta_{\mu}^m, \theta_{\sigma}^m\right) \quad \text{and} \quad \Delta_{\text{R}}^m \sim \text{Normal}\left(\Delta_{\mu}^m, \Delta_{\sigma}^m\right)$$

for each rat *R*, for each value of *m* indexing the various parameter vectors.

These population-level parameters are themselves modeled as draws from weakly informative prior distributions[54] chosen to enforce reasonable scaling and ensure that all posteriors were proper:

$$\theta_{\mu} \sim \text{Normal}(0,1) \quad \Delta_{\mu} \sim \text{Normal}(0,1)$$

$$\theta_{\sigma} \sim \text{Cauchy}(0,1) \quad \Delta\sigma \sim \text{Cauchy}(0,1)$$

Having established this generative model, we perform inference by conditioning it on the observed datasets (control and inactivation) for each rat and approximating the joint posterior over all parameters by drawing samples using Hamiltonian Markov chain Monte Carlo (H-MCMC)[54,59]. To obtain estimated values for each parameter, we took the median of these samples with respect to that parameter. To test whether inactivation produced effects on behavior

that were consistent at the population level, we computed a '*P* value' for each parameter in $\boldsymbol{\Delta}_{\boldsymbol{\mu}}$ given by the fraction of samples having the opposite sign as the median sample.

Inactivation model comparison. We performed a series of model comparisons between models like the above and alternative models in which inactivation affected memory in general, memory for distant past trials specifically, or a combination of these. In the first alternative model, inactivation was constrained to affect equally all of the agents that depend on the history of previous trials (planning, perseveration and novelty preference). This alternative model contains a new parameter, the 'memory multiplier', m, which scales the weights of these agents, in this revised version of equation (4):

$$Q_{\text{total}}(a) = \beta_{\text{bias}} Q_{\text{bias}} + \sum_{A \in \{\text{plan, persev, np}\}} m\beta_A Q_A(a)$$

This memory multiplier is fixed to 1 for control sessions but allowed to vary freely for each rat in infusion sessions. It has its own population-level mean and variance parameters, which are given weakly informative priors (see "Modeling inactivation data" section, above). In the alternative version of the model, the β_A parameters are fixed between control and inactivation sessions. Since bias does not require memory, β_{bias} is allowed to vary. We implement this by fixing the parameters Δ_{R}^{1} through Δ_{R}^{4} to zero for each rat R (see above), and allowing the effects of inactivation to be described by Δ_{R}^{5} and the new parameter m_R.

We compare this model to the model above using twofold cross validation of H-MCMC fits. To compare these models quantitatively, we compute the log posterior predictive ratio (lppr):

$$\log \frac{P(\text{testdata} \mid M1, \text{traindata})}{P(\text{testdata} \mid M2\ \text{traindata})}$$

In the next model comparisons, we separate the influence of the most recent trial's outcome from the influence of all trials further back in time. We implement this by replacing the model-based reinforcement learning agent (equations (5) and (6)) with both a model-based win–stay/lose–switch agent (equation (10), and a new 'lagged model-based' agent constructed by taking the value of Q_{plan} from one trial in the past and using it to guide decision-making on the current trial, so that the value of $Q_{\text{lagged - mb}}$ used on each trial contains information about the outcomes of all past trials except the most recent one. Fits of this model therefore contain two parameters to quantify planning: $\beta_{\text{wsls - mb}}$ for the influence of the most recent outcome and $\beta_{\text{lagged - mb}}$ for the influence of all trials further into the past.

For the second model comparison, we limit the influence of inactivation to only affect $\beta_{\text{lagged - mb}}$ and β_{bias}, that is, to affect the influence of distant past trials only on choice behavior and choice bias. For this model comparison, we also allow inactivation to affect the memory multiplier m, allowing it to have separate effects on memory for distant past trials and on memory for the immediately previous trial. We compare both of these models to a model in which inactivation can have separate effects on the each of the components of behavior. We compute the log posterior predictive ratio using leave-one-out cross-validation over sessions (i.e., we compute posteriors based on all of the dataset except for one session and compute the lppr for that session using those posteriors, then repeat for all sessions).

Synthetic behavioral datasets: inactivation data. To generate synthetic behavioral datasets, we took the parameter estimates produced by the hierarchical model for each rat for orbitofrontal, hippocampus and saline infusions. Parameters used for synthetic saline datasets were the average of the saline parameters produced by fitting the model to saline/hippocampus data and to saline/orbitofrontal (note that rat #6 did not complete any saline sessions; parameter estimates for this rat are still possible in the hierarchical model since they can be filled in based on infusion data and data from other rats). We used the reduced model in generative mode with these parameters, applying each parameter set to a simulated behavioral session consisting of 10,000 trials. We then applied the trial-history regression analysis to these synthetic datasets and used the results for qualitative model checking, comparing them to the results of the same analysis run on the actual data.

Code and data availability. All software used for behavioral training is available on the Brody lab website at http://brodylab.org/code/two-step-planning-task-code/. Software used for data analysis, as well as raw and processed data, are available from the authors upon reasonable request.

A Life Sciences Reporting Summary for this article is available.

51. Lau, B. & Glimcher, P.W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
52. Stan Development Team. MatlabStan: the MATLAB interface to Stan. *Stan.org.* http://mc-stan.org/users/interfaces/matlab-stan (2016).
53. Carpenter, C. *et al.* Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
54. Gelman, A. *et al. Bayesian Data Analysis, Third Edition* (CRC Press, 2013).
55. Krupa, D.J., Ghazanfar, A.A. & Nicolelis, M.A. Immediate thalamic sensory plasticity depends on corticothalamic feedback. *Proc. Natl. Acad. Sci. USA* **96**, 8200–8205 (1999).
56. Martin, J.H. Autoradiographic estimation of the extent of reversible inactivation produced by microinjection of lidocaine and muscimol in the rat. *Neurosci. Lett.* **127**, 160–164 (1991).
57. Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V. & van der Sluis, S. A solution to dependency: using multilevel analysis to accommodate nested data. *Nat. Neurosci.* **17**, 491–496 (2014).
58. Daw, N.D. in *Decision Making, Affect, and Learning* (eds. Delgado, M.R., Phelps, E.A. & Robbins, T.W.) 3–38 (Oxford University Press, 2011).
59. Duane, S., Kennedy, A.D., Pendleton, B.J. & Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222 (1987).

doi:10.1038/nn.4613

nature research

Corresponding author(s): Matthew Botvinick, Carlos Brody

☐ 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.

No formal power analysis was carried out. In all cases, we aimed for a sample size typical of similar studies in the field. In the case of behavior-only rats, we aimed for a sample size sufficient to characterize rat-by-rat variability in behavioral strategy. In the case of inactivation rats, we aimed for a sample size large enough to demonstrate the consistency of behavioral effects

2. Data exclusions

Describe any data exclusions.

We limited our analysis of inactivation data to the first 400 trials performed by the rat in each inactivation session. This criteria was consistent with previous practice in the lab.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Data were not divided formally into separate experiments for replication. We do report results for each subject separately in supplemental figures. For the major claims of our paper (rats show model-based index > 0; model-based index is decreased by hippocampus inactivation), the sign of the effect is the same in all subjects, and many of them are significant individually within-subject.

4. Randomization

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

The relevant comparisons in our paper are all within-subject, so formal randomization of subjects was not necessary or employed. Rats were were divided into common-congruent and common-incongruent conditions upon the beginning of training with the aim of collecting roughly equal numbers of subjects in each group. All results were similar between these two conditions. Within-subject randomization of trial types comes from the structure of the task, and is described in Methods.

5. Blinding

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

During behavior experiments, the rat was placed into and removed from the box by technicians blind to the experiment being run. Methods, Behavioral Apparatus. During infusion experiments, the experimenter was not blind to the brain region (OFC, dH, PL) or substance (muscimol or saline) being infused. Methods, Inactivation Experiments

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

Nature Neuroscience: doi:10.1038/nn.4613

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.	Analysis was performed using custom scripts created using MATLAB 2015, Stan 2.12, and R 3.3 . Software and data are available from the corresponding author upon request.

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.	All materials used are commercially available, and vendors are listed in the relevant Methods sections

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 in this work

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used in this work
b. Describe the method of cell line authentication used.	No cell lines were used in this work
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used in this work
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 cell lines were used in this work

Nature Neuroscience: doi:10.1038/nn.4613

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.

All subjects were adult male Long-Evans rats (Taconic Biosciences, NY), placed on a restricted water schedule to motivate them to work for water rewards. Some rats were housed on a reverse 12-hour light cycle, and others on a normal light cycle – in all cases, rats were trained during the dark phase of their cycle. Rats were pair housed during behavioral training and then single housed after being implanted with cannula. Animal use procedures were approved by the Princeton University Institutional Animal Care and Use Committee and carried out in accordance with NIH standards.

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.

No human subjects were used in this work

Nature Neuroscience: doi:10.1038/nn.4613