nature
neuroscience

Model-based choices involve prospective neural activity

Bradley B Doll[1,2], Katherine D Duncan[2], Dylan A Simon[3], Daphna Shohamy[2,4] & Nathaniel D Daw[1,3]

Decisions may arise via 'model-free' repetition of previously reinforced actions or by 'model-based' evaluation, which is widely thought to follow from prospective anticipation of action consequences using a learned map or model. While choices and neural correlates of decision variables sometimes reflect knowledge of their consequences, it remains unclear whether this actually arises from prospective evaluation. Using functional magnetic resonance imaging and a sequential reward-learning task in which paths contained decodable object categories, we found that humans' model-based choices were associated with neural signatures of future paths observed at decision time, suggesting a prospective mechanism for choice. Prospection also covaried with the degree of model-based influences on neural correlates of decision variables and was inversely related to prediction error signals thought to underlie model-free learning. These results dissociate separate mechanisms underlying model-based and model-free evaluation and support the hypothesis that model-based influences on choices and neural decision variables result from prospection.

The brain appears to employ two general strategies for decision-making, one relying on previous reinforcement and the other based on more flexible prospective reasoning about the consequences of actions. Under the first strategy, actions are valued by the rewards they have previously produced, as postulated in Thorndike's law of effect[1] and formalized in model-free reinforcement learning[2]. In contrast, under the second strategy, choices reflect knowledge of task contingencies or structure and also of the outcomes that might be realized, as demonstrated when navigating new paths in a spatial maze[3] or generalizing from known relationships to those that were never directly learned[4–6]. Such learning, formalized by model-based reinforcement learning theories, allows flexible evaluation of new or changing options[7,8].

Although there is much evidence that both choices and choice-related neural activity in reinforcement learning tasks can reflect knowledge of task contingencies beyond mere reward history[7–13], the nature of the computational process that actually gives rise to such model-based decisions and decision variables remains unclear. It is widely assumed that such behavior is produced by evaluation conducted prospectively at choice time through a sort of mental simulation, computing the value of potential actions over expected future trajectories. A possible substrate for such prospective computation is suggested by observations that hippocampal place cells report potential future paths during spatial navigation[14,15]; other prospective representations have been shown in humans using functional magnetic resonance imaging (fMRI)[16]. However, the evidence that choices and neural decision variables can reflect knowledge of task contingencies is separate from the reports of prospective neural representations, and it remains unknown whether the one underlies or even coincides with the other. Indeed, it is also possible that model-based choices result from some other mechanism altogether, since some choice algorithms can produce similar flexible behaviors through alternative means such as precomputing possible decisions when outcomes are received[17–19]. Consistent with these alternative mechanisms, some evidence suggests that flexible, apparently model-based choices in humans are driven at least in part by generalization that occurs during initial learning[4,5,20] or during rest periods[21]. Here we sought to directly test the hypothesis that model-based choices arise from forward-looking computations at the time of choice.

RESULTS

Behavior reflects both model-based and model-free learning

Twenty human subjects underwent fMRI while performing a two-stage sequential decision-making task[22] designed to distinguish model-based from model-free reinforcement learning strategies. Stages in the task were represented using visual stimuli from categories with specific neural correlates (faces, tools, body parts, scenes), allowing us to probe their prospective representations in category-specific regions of cortex at choice time (**Fig. 1**). Each trial began in one of two start states (faces or tools), determined pseudorandomly, where participants chose between two options. This initial choice deterministically controlled which of two more two-option choices (scene or body part states) they would encounter next. (This aspect of the task differs from those in previous studies of similar sequential decision tasks[11,12,23], which relied on the consequences of first-stage choices being stochastic.) Each second-stage option was rewarded with money or not rewarded, with a slowly and randomly drifting probability, such that subjects continuously learned by trial and error which sequence of choices was most likely to be rewarded.

The first-stage options were implicitly parallel between the two states: selecting one of the tools or one of the faces always led to the scenes, while the other tool or face led to the body parts. This structural equivalence allowed us to dissociate behavior consistent with model-based and model-free learning approaches because only

[1]Center for Neural Science, New York University, New York, New York, USA. [2]Department of Psychology, Columbia University, New York, New York, USA. [3]Department of Psychology, New York University, New York, New York, USA. [4]Kavli Institute for Brain Science, Columbia University, New York, New York, USA. Correspondence should be addressed to B.B.D. (bradley.doll@nyu.edu).

Received 5 December 2014; accepted 15 February 2015; published online 23 March 2015; doi:10.1038/nn.3981

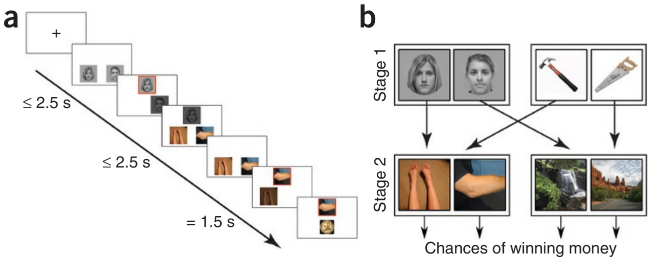

Figure 1 Task design. (**a**) Timeline of events. 272 trials begin on a randomly selected first-stage state (faces or tools, left/right presentation randomized). First-stage choices deterministically produce second-stage choices (body parts or scenes), which probabilistically produce reward. (**b**) State transition structure. Choices in the face state are equivalent to choices in the tool state. Faces and tools here depicted on the left always lead to the body part state, while those depicted on the right always lead to the scene state.

model-based learners automatically generalize experiences across the equivalent start state options. Specifically, they compute each option's value prospectively in terms of its expected second-stage consequences. For this reason, for a model-based learner, each outcome at the second stage should have equivalent effects on first-stage preference on the next trial, regardless of whether the new trial starts with the same state as its predecessor (for example, faces followed by faces) or a different one (for example, faces followed by tools; **Fig. 2a**). In contrast, a canonical model-free learner evaluates options in terms of the outcomes they have previously produced. Outcomes received following one start state will not affect subsequent choices from the other start state (**Fig. 2b**). Importantly, although generalization indicates that choices effectively take account of the sequential contingencies, such an effect might in principle be supported by some computation that does not involve the prospective contemplation of future paths at choice time. For instance, a reward might be credited, at the time of receipt, to both choices that might have produced it[4,7]. Similar nonprospective mechanisms for model-based evaluation can apply more generally to arbitrary tasks[19].

We analyzed how choices were influenced by preceding rewards to investigate model-based and model-free influences on choice. Consistent with previous reports using similar tasks[11–13,22], evidence of both strategies was observed in behavior. Significant generalization of rewards was observed between equivalent start states ($P = 2.4 \times 10^{-5}$), but still larger effects of rewards were observed when the start state remained the same ($P = 0.0036$; **Fig. 2c**, Online Methods and **Supplementary Table 1**), with notable individual variation in the relative balance of these effects.

Model-based behavior involves neural prospection

As mentioned, while the behavioral findings demonstrate that choices reflect the state transition structure, behavior alone does not establish whether this dependence involves prospective evaluation at choice time. To investigate this, we first took advantage of between-subject variability in behavioral strategies together with the patterns of blood-oxygen-level dependent (BOLD) responses to task stimuli, which recruit dissociable regions of the visual and temporal cortices[24] (**Fig. 3a**,**b**). We hypothesized that if model-based choices arise from prospective evaluation at choice time then subjects relying on this strategy should show neural correlates of states they planned to visit, with the degree of the neural effect reflecting variability in the behavioral strategy.

We assessed this prospective activation using BOLD activity in category-specific regions defined per-subject from an independent functional localizer (Online Methods). We included nuisance variables in the model to control for all categories of on-screen events and their durations, in order to account for residual visual effects from the previous trial that might contaminate estimation of prospective activity. Further, to avoid confounding any prospective activation related to the chosen option with activation related to the actual visual presentation of the end-stage stimuli that follow, this analysis was restricted to randomly interspersed catch trials. On these trials, the first stage proceeded as usual but was followed by the presentation of two gray boxes rather than the end-stage states.

We found that the strength of prospective activation of the chosen versus unchosen second-stage categories correlated positively with the relative strength of model-based choice behavior across subjects (estimate = 0.46, $\chi^2(1) = 4.4$, $P = 0.036$, **Fig. 3c**, **Supplementary Table 1**, Online Methods general linear model (GLM) 1). This result is consistent with model-based choices being computed by prospective evaluation at decision time, as has been widely speculated. To control for the possibility that this correlation was affected by differences across subjects in the patterns of visual experience arising from choices that were more or less model based (for example, because model-free choices produce more alternation in the encountered second-stage state), we reestimated prospective activation on the subset of catch trials in which subjects visited the same second stage as in the previous trial. In this more stringent analysis, which equates previous-trial visual experience across subjects regardless of their choice strategy, the positive relationship between prospection and model-based choice remained significant (estimate = 0.38, $\chi^2(1) = 4.94$, $P = 0.026$).

Neural prospection relates to other neural signatures of model-based decision variables

Next we considered whether neural evidence for prospection also related to neural signatures of trial-by-trial decision variables, which

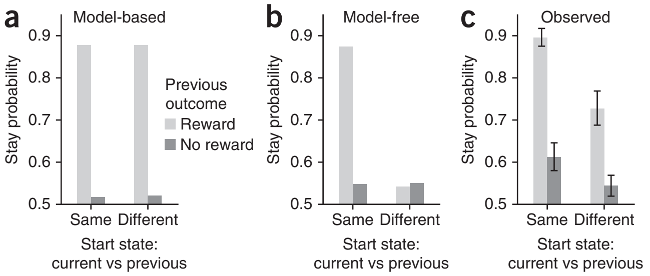

Figure 2 Model behavioral predictions and data. Plots depict the probability of staying with start-state choice made on previous trial. Choices binned by whether start state (tools or faces) was the same as or different from the start state on the previous trial and whether the previous trial ended in reward or not. (**a**) Model-based predictions. Start states are used in terms of the transitions they afford, so outcomes affect subsequent choice regardless of whether the previous start state matches or differs from the current one. (**b**) Model-free predictions. Separate values are learned for the actions at the different start states, so outcomes only affect subsequent choice on trials with the same start state. Panels **a** and **b** produced from generative reinforcement learning model task performance (Online Methods), with weighting parameter *w* specifying fully model-based ($w = 1$) and fully model-free learning ($w = 0$), respectively. (**c**) Observed data indicate the presence of both effects. The effect of previous reward on choice (estimate = 0.47, $Z = 4.6$, $P = 2.4 \times 10^{-5}$) is greater when the current start state matches the previous one (estimate = 0.48, $Z = 2.9$, $P = 0.0036$, Online Methods and **Supplementary Table 1**). Error bars reflect s.e.m.

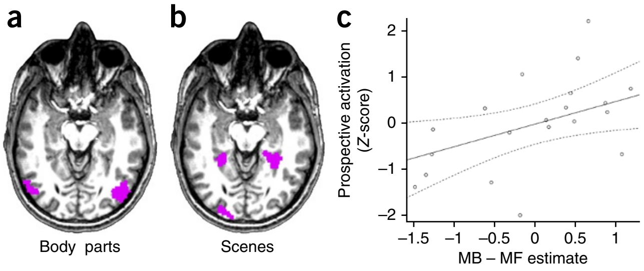

Figure 3 Neural evidence of prospective activation correlates with model-based behavior. (**a**,**b**) Example subject regions of interest (ROIs), derived from independent functional localizer, for body parts (**a**) and scenes (**b**) (see **Supplementary Fig. 2** for group-level depiction of these ROIs). (**c**) Correlation of averaged ROI BOLD response and model-based relative to model-free choice behavior. Prospective activation estimated at task start stage on catch trials (68 randomly interspersed trials in which no second stage occurred) as a contrast of chosen second-stage state relative to the unchosen one in the relevant ROI (body parts or scenes; individual scores are averages of the two contrasts). Prospective activation correlates positively with tendency to make model-based choices (estimate = 0.46, $\chi^2(1) = 4.4$, $P = 0.036$). Lines depict group-level linear effects and 95% confidence curves.

have also previously been shown to reflect evidence of model-based influences[11,12,23]. Neural correlates of the expected value of chosen options have been reported throughout frontal cortex, often reflecting the difference in value between chosen and unchosen options[23,25,26], which we express in units of probability as the likelihood assigned by the reinforcement learning model to the chosen (versus unchosen) option[12,27]. In some areas, such correlations have a negative sign (that is, activity correlates positively with unchosen minus chosen value), perhaps reflecting the relative value of alternatives or switching[26–29]. Since model-based and model-free methods differ in how they compute expected value (for example, via prospection or reward history, generalizing or not between start states), these influences on trial-by-trial neural value signals are dissociable, analogously to choice behavior. We looked for correlates of these different first-stage choice values in the brain. Specifically, we modeled the first stage of the task with two parametric regressors (Online Methods, GLM2). The first regressor encoded the time series of reinforcement learning model–estimated probabilities of selecting the chosen stimulus on each trial for a model-free learner. The second regressor encoded the difference in these choice probabilities between a model-based and model-free learner, allowing us to recognize, by any positive loading on this regressor, the extent to which neural signals reflect any specifically model-based influences over any model-free ones[12].

This analysis identified a number of regions in the frontal cortex that correlated negatively with the first-stage choice probabilities—that is, activity correlating positively with the relative value of the unchosen option (**Fig. 4a**,**b** and **Supplementary Tables 2** and **3**)—for both

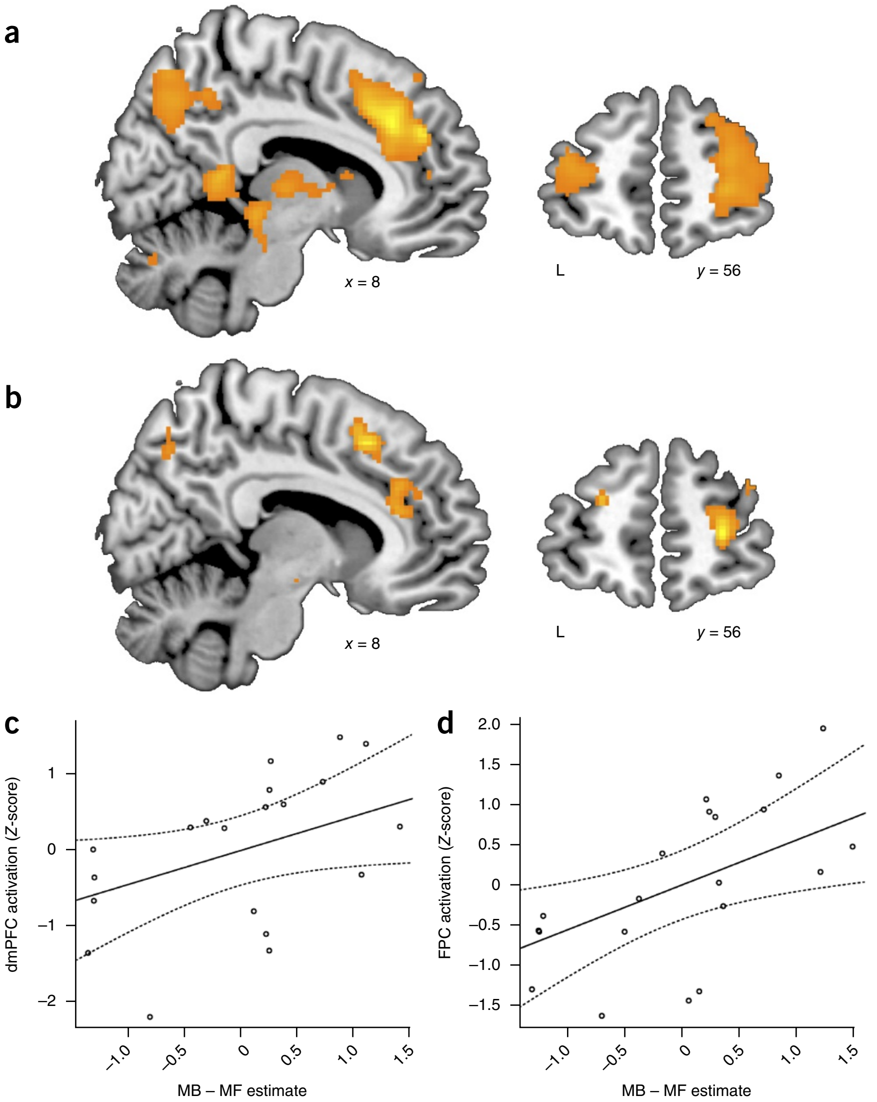

Figure 4 Correlates of choice probabilities derived from chosen minus unchosen values estimated by model-free and model-based learning at the task's first stage. Highlighted regions show negative correlation (that is, activity correlates positively with unchosen minus chosen value) for (**a**) model-free choice probabilities and (**b**) the difference between model-based and model-free choice probabilities, the latter difference isolating activity significantly related to model-based rather than model-free learning. Bold response in dmPFC (left in **a** and **b**; model-free peak: 8 30 34, $P = 0$; model-based – model-free peak: 12 34 16, $P = 3.9 \times 10^{-7}$) and FPC (right in **a** and **b**; model-free peak: −40 44 12, $P = 2.5 \times 10^{-11}$; model-based – model-free peak: 26 56 6, $P = 2.9 \times 10^{-8}$) correlates negatively with both regressors (**Supplementary Tables 2** and **3**). Cluster P values corrected for family-wise error for whole-brain comparisons. Maps thresholded at $P < 0.001$, uncorrected for display purposes. L, left. Color maps reflect t statistics, ranging from 3.6 to 8.7. (**c**,**d**) Effect size (from **b**) of model-based (MB) (minus model-free (MF)) unchosen choice probabilities in dmPFC and FPC correlate with the tendency to make model-based choices across subjects. (**c**) dmPFC (estimate = 0.45, $\chi^2(1) = 3.89$, $P = 0.049$). (**d**) FPC (estimate = 0.55, $\chi^2(1) = 6.4$, $P = 0.011$). Lines depict group-level linear effects and 95% confidence curves.

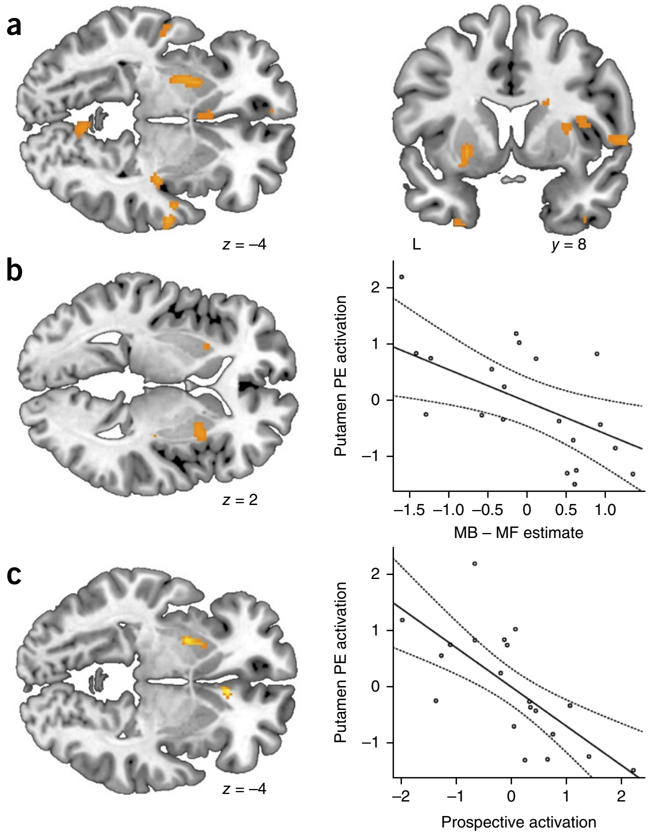

Figure 5 Neural evidence of model-free prediction errors and correlates of prediction error with model-free behavior. (**a**) Putamen BOLD response correlates with model-free prediction errors (PEs) that accompany state transitions (peak: –24 8 –4, $P = 0.0005$, cluster-corrected for family-wise error for whole-brain comparisons; **Supplementary Table 4**). (**b**) Effect size of model-free PEs in putamen covaries negatively across subjects with the tendency to make model-based relative to model-free choices. Left: peaks 28 10 2; 32 12 0 ($P = 0.019$), –26 12 –8 ($P = 0.0826$). Coordinates and P values here and in **c** (left panel) reflect small-volume correction for clusters in anatomical mask of striatum. Right: correlation estimated from average activity in significant clusters depicted in **a**, restricted to striatum (estimate = –0.57, $\chi^2(1) = 6.93$, $P = 0.008$). (**c**) Neural measures of model-based prospection and model-free PE negatively correlate. Left: clusters showing across subject negative correlation of model-based prospection and model-free PE. Peaks: –26 2 –4 ($P = 0.032$), 28 10 10 ($P = 0.001$). Right: correlation estimated from average activity in significant striatal clusters depicted in **a** ($r = –0.73$, $P = 0.0003$). Lines in **a** and **c** depict group-level linear effects and 95% confidence curves. Maps thresholded at $P < 0.001$, uncorrected for display purposes. Color maps reflect t statistics, ranging from 3.6 to 9.3.

model-free and model-based (minus model-free) regressors. Consistent with previous reports[26–30], this activation was seen in the BOLD response in dorsomedial prefrontal cortex (dmPFC; including anterior cingulate and mid-cingulate cortex (model-free peak: 8 30 34, $P = 0$; model-based minus model-free peak: 12 34 16, $P = 3.9 \times 10^{-7}$) and lateral frontopolar cortex (FPC; model-free peak: −40 44 12, $P = 2.5 \times 10^{-11}$; model-based minus model-free peak: 26 56 6, $P = 2.9 \times 10^{-8}$). Focusing on these activations, we asked whether the size of the model-based relative valuation effect (unchosen minus chosen) related to the degree to which subjects' choices were model-based relative to model-free. Across subjects, the size of the model-based effect in dmPFC and in lateral FPC correlated significantly with the behavioral tendency to make model-based choices (**Fig. 4c**,**d**; dmPFC: estimate = 0.45, $\chi^2(1) = 3.89$, $P = 0.049$; FPC: estimate = 0.55, $\chi^2(1) = 6.4$, $P = 0.011$). These results link the neural correlates of decision variables used by a model-based learner with model-based choice in the task.

We next assessed whether the neural markers of prospection and model-based valuation covaried on a trial-by-trial basis. We reasoned that if these model-based decision variables themselves arise from prospective computation then these two neural signatures should also relate to one another. Trials with greater prospective activation, for example, should be accompanied by a stronger representation of model-based choice probability. We estimated neural prospection on each catch trial and entered this time series into subsequent models (Online Methods, GLM3) of dmPFC and FPC BOLD response, using functional regions of interest defined from the parametric effect of unchosen minus chosen choice probability (**Fig. 4b**). This analysis revealed a significant interaction in the dmPFC of model-based choice probability on catch trials and prospection (model-based × prospection mean = 4.35, $t(19) = 2.19$, $P = 0.041$; model-free × prospection mean = −0.23, $t(19) = −0.13$, $P = 0.9$). The effect in FPC trended in the same direction but did not reach significance (model-based × prospection mean = 3.05, $t(19) = 1.79$, $P = 0.088$; model-free × prospection mean = 3.28, $t(19) = 1.27$, $P = 0.22$). These results are consistent with the hypothesis that prospective computations also underlie the computation of neural decision variables of the sort that have previously been taken as a signature of model-based computations.

Prediction error signals are related to model-free behavior and reduced prospective neural activity

Having established a relationship between prospective neural activation, model-based valuation and model-based choice behavior, we next sought to investigate how all these quantities relate to previously identified trial-by-trial neural signatures of model-free learning. Learning of model-free values is widely believed to be driven by reward prediction errors, whose correlates are often observed in striatum[31]. We hypothesized that prospection might relate negatively to any such signatures since model-free learning represents an alternative approach to solving the task.

Although prediction error signals have been shown in other contexts to reflect contributions of both model-based and model-free value expectations[12], some prediction errors in the current task are expected to be purely model free. This is because a model-based learner would base reward expectations on the deterministic transition structure of the task and therefore would experience no prediction error (change in reward expectation) when transitioning between task stages (in contrast to results previously reported in a task with stochastic transitions[12]). Conversely, for a model-free learner, a prediction error is encountered as the difference in the expected values of the chosen options in the second and first task stages (Online Methods equation (4)), which generally will not match because of a failure to generalize between start states. Thus, prediction error activation accompanying stage transitions is expected, in the current task, to be a unique and affirmative signature of the extent to which subjects employ model-free learning.

We observed such activation in several locations (Online Methods, GLM2, **Supplementary Table 4**), notably the left putamen (**Fig. 5a**, peak: −24 8 −4, $P = 0.0005$, cluster corrected for family-wise error for whole-brain multiple comparisons), an area previously implicated in habitual, extensively trained and model-free actions[23,32–34]. Further, the size of this effect in putamen across subjects correlated negatively with the tendency to make model-based relative to model-free choices (**Fig. 5b**, estimate = −0.57, $\chi^2(1) = 6.93$, $P = 0.008$), confirming the theoretical prediction that these signals should correspond specifically to model-free choice. Together, these results show affirmative neural correlates of each behavioral strategy. Neural prospection effect size correlated with the degree to which behavior was model-based, and prediction error effect size in putamen correlated with the degree to which it was model-free. This affirmative demonstration of neural correlates of each strategy helps to rule out the possibility that positive covariation between model-based learning and prospective activity in stimulus-sensitive areas was driven by more generic differences in motivation or attention. A second affirmative neural signature of model-free learning was also observed in the inferior frontal gyrus using a simpler contrast between trials in which the first-stage state changed rather than staying the same (Online Methods, GLM4, **Supplementary Fig. 1**).

Finally, we sought to determine the relationship between the neural signatures of prospection and prediction error. We did this by examining how they covary across subjects. (These measures could not be compared across trials within a subject because the prospection index is only defined on catch trials and the prediction error only on non-catch trials, when the second stage actually appears.) Supporting their dissociability, the neural signatures of the two strategies were inversely related: the size of the neural prospection effect correlated negatively with that of striatal model-free prediction error activity ($r = -0.73$, $P = 0.0003$, **Fig. 5c**). These results support the hypothesis that learning from prediction errors and prospective anticipation of future states represent alternative mechanisms for evaluating candidate actions that underlie model-free and model-based influences on choice.

DISCUSSION

These results demonstrate a relationship between behavioral and neural signatures of model-based evaluation and prospective neural computation at the time of choice. While it is often assumed that model-based reinforcement learning arises from forward-looking computations, direct evidence has been lacking, and indeed some evidence suggests that alternative, nonprospective computations may also support these phenomena[4,5]. Here we demonstrate that neural evidence of prospection correlates with the degree of model-based influences on both choices and neural decision variables and, furthermore, show evidence that this prospective evaluation mechanism trades off against prediction error signals for model-free learning.

Across species, both choices and accompanying neural activity have repeatedly been shown to reflect knowledge from a cognitive map or model. However, such effects need not in principle arise from prospective computation at the time of choice. Alterative computational accounts show that model-based behavior can arise, for example, by precomputing the results of a tree search[17,18]. Some evidence for such precomputation has been reported in humans[21]. A related approach is to generalize feedback when an outcome is received to the multiple actions that could have produced that outcome. Indeed, several studies have shown that apparently model-based choices are related to neural measures at the time of feedback rather than at the time of choice[4,5,20,35]. In the current task, for example, assigning credit from outcomes to both equivalent start stage actions would produce model-based behavior (but not neural prospection) through model-free updating methods. Though such generalization is not mutually exclusive and could also contribute, our results provide affirmative evidence for an alternative, prospective mechanism supporting model-based behavior and neural decision variables. These results have close parallels with the rodent literature on spatial navigation, where animals can solve navigational problems such as latent learning that require planning routes using a cognitive map[36]. A candidate mechanism to support this ability is prospective representations of potential future spatial locations in hippocampal place cells[14,15], but since such neural recordings have not yet been paired with behavioral manipulations demonstrating map-based choice, this connection remains to be directly established.

Our results establish a relationship between prospective neural activity and model-based choices, but owing to temporal limitations of fMRI they do not speak to the detailed dynamics by which this prospection occurs. Although some computational models envision a state-by-state traversal of the future decision tree[37], our data are compatible with many variant algorithms that share the key feature of predicting future states, but in various different orders or groups[19]. One promising candidate approach is to incorporate a predictive world model at the representational stage, representing states or actions in terms of the states or actions to which they are expected to lead[38,39]. Indeed, evidence suggestive of such representational shifts has been observed in the brain during incidental statistical learning tasks[16,40], though this has not previously been connected to model-based choice behavior. In the current task, this would amount to learning to represent the first-stage options by retrieving a representation of the second-stage options they produce. While such a predictive representation is consistent with our data—and with our conclusion that model-based choice is supported by retrieving future states from a learned predictive model—other superficially similar but nonpredictive representational schemes are not. For instance, equivalent start actions might be represented in terms of one another[41], which would support generalization between them behaviorally but would not produce neural activity selective for their shared future state at choice time. In general, representing actions in terms of their future consequences is closely related to another set of questions that have been studied in behavioral neuroscience, representation of hierarchical structure or temporal abstraction—for example, 'chunking' of state or action sequences into a unit. Taking account of such structures simplifies computation[39,42,43] by capturing aspects of the predictive regularities of events, much as in a world model.

In addition to the neural signals of prospection, we observed neural correlates of model-based decision variables in the dmPFC and FPC, both areas associated with representation values for unchosen or alternative actions. While the specific computational role of the dmPFC is debated[29,30], this region has been repeatedly implicated in the relative valuation of different courses of action, and, as such, signals there are appropriate for dissecting model-based versus model-free components of these valuations. (In the present data set, we did not observe strong value-related activity at the first choice stage in the ventromedial prefrontal cortex, another frontal area that has been associated with model-based valuation of chosen actions[23].) Like choice behavior, model-based influences on neural decision variables of this sort have been previously taken as a signature of model-based evaluation[12,23,44], and, as for choice behavior, the question arises whether these effects actually result from prospection. Consistent with this hypothesis, the strength of this effect in the dmPFC covaried significantly with neural prospection.

A second widely observed correlate of neural decision variables is reward prediction errors, which, although they can be influenced by model-based valuations as inputs[12], are believed to be involved in the learning of model-free evaluations. This leads to the expectation that such signals should trade off against prospection. The correlation of putamen activation and model-free prediction errors in the current study is consistent with the involvement of the putamen in habitual, extensively trained and model-free actions[23,32–34]. Activation in the putamen was indeed positively associated with model-free behavior, and it also was negatively associated with the neural prospection measure that accompanied model-based behavior, suggesting that these two behavioral strategies rely on two distinct neural signals.

These findings reveal a neural signature of model-based evaluation, which directly supports the widespread supposition that model-based preferences are computed by prospective evaluation. This signature bears the expected relationships to previously reported behavioral and neural correlates of model-based learning and, notably, relates negatively with neural prediction error signals that are thought to drive an alternative, model-free strategy for evaluation. By characterizing the distinct mechanisms driving model-based versus model-free decisions, these findings also have implications for maladaptive decision-making. Disorders such as addiction are characterized by an apparent failure to account for the future consequences of behavior, which may relate to an over-reliance on model-free evaluation[45–47]. A greater understanding of when and how decisions are made prospectively may guide interventions to shift the balance in favor of future-oriented processes.

METHODS

Methods and any associated 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 S.M. Fleming and L.Y. Atlas for helpful discussions. This work was supported by NINDS grant R01NS078784.

AUTHOR CONTRIBUTIONS

All authors designed the experiment and analyses. B.B.D. and K.D.D. performed the experiment. B.B.D. analyzed the data. B.B.D., N.D.D. and D.S. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Thorndike, E.L. *Animal Intelligence: Experimental Studies* (Macmillan, New York, 1911).
2. Sutton, R.S. & Barto, A.G. *Introduction to Reinforcement Learning* 〈http://dl.acm.org/citation.cfm?id=551283〉 (MIT Press, 1998).
3. Tolman, E.C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
4. Shohamy, D. & Wagner, A.D. Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* **60**, 378–389 (2008).
5. Wimmer, G.E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
6. 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).
7. Doll, B.B., Simon, D.A. & Daw, N.D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
8. Dolan, R.J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
9. Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974 (1999).
10. Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J. & Doya, K. Evidence for model-based action planning in a sequential finger movement task. *J. Mot. Behav.* **42**, 371–379 (2010).
11. 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).
12. 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).
13. Eppinger, B., Walter, M., Heekeren, H.R. & Li, S.-C. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Front. Neurosci.* **7**, 253 (2013).
14. Pfeiffer, B.E. & Foster, D.J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
15. Johnson, A. & Redish, A.D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12176–12189 (2007).
16. Schapiro, A.C., Kustner, L.V. & Turk-Browne, N.B. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr. Biol.* **22**, 1622–1627 (2012).
17. Moore, A.W. & Atkeson, C.G. Prioritized sweeping: reinforcement learning with less data and less time. *Mach. Learn.* **13**, 103–130 (1993).
18. Sutton, R.S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine Learning: Proc. Seventh Int. Conf. on Machine Learning* (eds. Porter, B.W. & Mooney, R.J.) 216–224 (Morgan Kaufmann, Palo Alto, California, USA, 1990).
19. Daw, N.D. & Dayan, P. The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130478 (2014).
20. Zeithamova, D., Dominick, A.L. & Preston, A.R. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168–179 (2012).
21. Gershman, S.J., Markman, A.B. & Otto, A.R. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
22. Doll, B.B., Shohamy, D. & Daw, N.D. Multiple memory systems as substrates for multiple decision systems. *Neurobiol. Learn. Mem.* **117**, 4–13 (2015).
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. Reddy, L. & Kanwisher, N. Coding of visual objects in the ventral stream. *Curr. Opin. Neurobiol.* **16**, 408–414 (2006).
25. FitzGerald, T.H.B., Seymour, B. & Dolan, R.J. The role of human orbitofrontal cortex in value comparison for incommensurable objects. *J. Neurosci.* **29**, 8388–8395 (2009).
26. Boorman, E.D., Behrens, T.E.J., Woolrich, M.W. & Rushworth, M.F.S. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743 (2009).
27. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
28. Boorman, E.D., Behrens, T.E. & Rushworth, M.F. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol.* **9**, e1001093 (2011).
29. Kolling, N., Behrens, T.E.J., Mars, R.B. & Rushworth, M.F.S. Neural mechanisms of foraging. *Science* **336**, 95–98 (2012).
30. Shenhav, A., Straccia, M.A., Cohen, J.D. & Botvinick, M.M. Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nat. Neurosci.* **17**, 1249–1254 (2014).
31. Garrison, J., Erdeniz, B. & Done, J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* **37**, 1297–1310 (2013).
32. Foerde, K., Knowlton, B.J. & Poldrack, R.A. Modulation of competing memory systems by distraction. *Proc. Natl. Acad. Sci. USA* **103**, 11778–11783 (2006).
33. Tricomi, E., Balleine, B.W. & O'Doherty, J.P. A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* **29**, 2225–2232 (2009).
34. 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).
35. Kurth-Nelson, Z., Barnes, G., Sejdinovic, D., Dolan, R. & Dayan, P. Temporal structure in associative retrieval. *Elife* **4**, e04919 (2015).
36. Tolman, E.C. & Honzik, C.H. Introduction and removal of reward, and maze performance in rats. *Univ. Calif. Publ. Psychol.* **4**, 257–275 (1930).
37. 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).
38. Dayan, P. Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624 (1993).
39. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130480 (2014).
40. 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).
41. Gluck, M.A. & Myers, C.E. Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* **3**, 491–516 (1993).
42. Badre, D., Kayser, A.S. & D'Esposito, M. Frontal cortex and the discovery of abstract action rules. *Neuron* **66**, 315–326 (2010).
43. Botvinick, M.M., Niv, Y. & Barto, A.C. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
44. Simon, D.A. & Daw, N.D. Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* **31**, 5526–5539 (2011).
45. Everitt, B.J. & Robbins, T.W. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* **8**, 1481–1489 (2005).
46. Redish, A.D. Addiction as a computational process gone awry. *Science* **306**, 1944–1947 (2004).
47. Voon, V. *et al.* Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron* **65**, 135–142 (2010).

ONLINE METHODS

Subjects. 23 subjects (13 female, mean age = 23.8 years, s.d. = 4.6) were recruited from the NYU community. No statistical tests were used to predetermine the sample size, but this sample size is within the standard range in the field. One subject aborted the experiment, and behavior of two more was better fit by a null baseline model (see below) than the task model, which hybridizes model-based and model-free solutions, and so they were excluded from all analyses. Subjects were paid $30 in addition to their earnings on 20 randomly selected trials on the two-stage decision task. Subjects provided informed consent to participate in the study, which was approved by New York University's human subjects committee.

Tasks. *Functional localizer*. We used blocked functional localizer runs to identify subject-specific regions of interest (ROIs) that responded differentially to each task state (faces, body parts, tools, scenes). Two runs before and two after the two-stage decision task were completed. In each localizer run, 14-s image blocks were followed by 14-s rest blocks. In each image block, 20 images (either faces, tools, scenes, body parts or scrambled tools) were presented in randomized order for 300 ms, with a 400-ms inter-stimulus interval. Each category of image block recurred three times per run. During the functional localizer, subjects were instructed to attend to the images on the screen and respond by button press when a small black dot appeared somewhere on an image (10% of images, randomly interleaved). The images in the first two (pre-task) localizer runs were the 8 stimuli used in the subsequent two-step task. The images in the final two (post-task) runs were 20 novel exemplars from the same categories.

Two-stage decision task. Subjects completed 272 trials of a two-stage decision task in four runs of 68 trials (**Fig. 1**). On each trial, the first stage began on one of two randomly chosen start states (either the face or tool state). Selection of one of the two first stage stimuli resulted in transition to the second task stage, which also consisted of two possible states (body parts or scenes). Transition from the first stage to the second stage was deterministic: selection of one face or the equivalent tool always led to the body part state in the second stage, while selection of the other face or equivalent tool always led to the scene state in the second stage.

Selection between stimuli in the second stage (body part or scene state) produced a reward ($1 or $0) with a randomly and slowly diffusing probability between 0.25 and 0.75. To encourage visitation of both second-stage states over the course of the task, expected values of each were made equal over the length of the experiment: the Gaussian random walks that controlled the reward probabilities for each stimulus in one state were temporally reversed to set the reward probabilities for each stimulus in the other state. Left/right presentation of the stimuli in each stage was randomized from trial to trial.

To avoid confounding prospective activity related to the chosen option with activity related to the actual visual presentation of the end-stage stimuli that followed (**Fig. 1a**), 68 of the 272 task trials were selected as catch trials (17 randomly selected per run, excluding the first 5 trials of the first run, and with the condition that catch trials be separated by at least one non-catch trial). On these trials, first-stage action selection was followed by the presentation of two gray boxes instead of the two stimuli in the chosen second-stage state.

Subjects had 2.5 s to respond at each stage. After stimulus selection, the chosen stimulus immediately rose to the top of the screen, while the unchosen stimulus dimmed and disappeared. On catch trials, the chosen first stage stimulus also disappeared. Failure to respond within the response window aborted the trial (red Xs appeared over all stimuli). Outcomes ($1 or $0) appeared on screen for 1.5 s. Trials were separated by a jittered intertrial interval (durations drawn from an exponential distribution with mean 3 s). Remaining time in the response window of each stage (2.5 s – reaction time) was added to the subsequent intertrial interval.

Before entering the scanner, subjects were instructed on the rules of the task and practiced 50 trials with different stimuli (Tibetan characters). Subjects were informed that the catch trials were randomly interspersed in the task and could not be predicted, and that choices would not influence catch trial occurrence in any way.

Behavioral analysis. Following previous work[12,48], we fit the choices from the decision task in two complementary ways: by using a full reinforcement learning model that nests both model-based and model-free approaches and by using a simplified logistic regression model that captures the essence of both learning approaches more qualitatively by examining only the effect of the most recent trial's outcome on choice.

Multilevel logistic regression. We fit multilevel logistic regression models to the choices from the decision task using the lme4 package (http://cran.r-project.org/web/packages/lme4/index.html) in the R statistical language (http://www.r-project.org/). All coefficients (exclusive of fMRI covariates described below) were taken as random effects—that is, varying from subject to subject around a group mean.

For each subject, choice of a first-stage action on each trial was regressed onto the choices and rewards from the previous trial. The pattern of these dependencies distinguishes signatures of model-based from model-free strategies (**Fig. 2**). Specifically, the binary dependent variable was choice of actions that produce the body parts state versus choice of actions that produce the scenes state. Explanatory variables for each trial i were, in addition to an intercept, the previous reward r_{i-1}, and the previous choice c_{i-1} to capture any tendency to repeat or switch actions regardless of reward. Critically, to assess whether the effect of experience in one start state is carried over to the equivalent action in the other state, a binary explanatory variable ($same_i$) was included. The variable $same_i$, which indicates whether the current start state was the same or different from the start state on the previous trial, both was entered alone and interacted with r_{i-1} and c_{i-1}.

In this regression, the main effect of the previous reward r_{i-1} on choice indexes model-based choice because it carries over to the next trial even when start states change ($same_i = 0$), whereas the interaction term $r_{i-1} * same_i$ captures any reward effects are that specific to the state in which they were received ($same_i = 1$) and thus indexes model-free choice. We used the difference between the model-based and model-free coefficients to assess the relative weighting of each strategy in each subject's choices (**Fig. 2**) (this is analogous to the weighting parameter w in the computational model; see "Computational model" below and **Supplementary Table 1**).

To test whether per-subject measures of BOLD activation related to model-based or model-free learning, we used fMRI activity to create additional explanatory variables for the above regression model of choices, so as to estimate whether the relevant choice effects changed as a function of (that is, interacted with) the BOLD-derived covariate. (This is analogous to estimating parameters separately from the choices and the fMRI activity, then testing whether they correlate with one another, but by testing these effects within the behavioral model, we explicitly account for uncertainty about the estimated behavioral coefficients rather than treating them as point estimates.) Specifically, one fMRI-derived estimate for each subject (Z-scored effects from the models of the fMRI time series described below) was entered into the regression model, fully crossed with the rest of the factors to produce a set of additional explanatory variables measuring the interaction between the BOLD measure and each of the other explanatory variables.

Since the difference between the coefficients for r_{i-1} and $r_{i-1} * same_i$ is our index of the relative weighting of model-based and model-free learning, we analogously measure the extent to which this weighting changes with the neural covariate by testing the difference from zero between the coefficients measuring the interactions of the BOLD covariate with the model-based ($r_{i-1} * covariate$) and model-free ($r_{i-1} * same_i * covariate$) terms. This contrast deviates positively from zero for covariates that are increasing as net model-based choice increases, and negatively from zero for covariates that are decreasing.

We estimated five such models relating BOLD activation to first-stage decision task choices. The BOLD covariates in these models were, first, prospective activation (**Fig. 3c**, see "fMRI analysis" below, GLM1); second and third, dmPFC and FPC model-based relative valuation of first stage choices, respectively (**Fig. 4c**,**d**; see "fMRI analysis," GLM2); fourth, putamen PE activation (**Fig. 5b**,**c** right panels; see "fMRI analysis," GLM2); fifth, inferior frontal gyrus response to changes of first-stage states across trials (**Supplementary Fig. 1b**, GLM4).

In addition to the multilevel models, in order to ensure that subjects' choice behavior was interpretable, we fit this logistic regression model (without covariates from fMRI data or multilevel structure) separately for each subject and compared model fit quality to an intercept-only (baseline) model, controlling for the different number of parameters in each model with Akaike's information criterion[49]. Two subjects were better fit by the baseline model, indicating their choices were not significantly related to any of the feedback in our task, and were excluded from subsequent analysis.

fMRI procedures. Gradient echo T2*-weighted echo-planar images (EPI) with blood oxygenation level dependent (BOLD) contrast were collected on a 3-T

Siemens Allegra MRI scanner. Forty axial slices (3 × 3 × 3 mm voxels) were acquired in oblique orientation of 30 degrees to the anterior commissure–posterior commissure line, with a repetition time (TR) of 1,750 ms, TE of 17 ms, 80° flip angle, 240 mm × 192 mm field of view. A high-resolution T1-weighted anatomical image (magnetization-prepared rapid-acquisition gradient echo sequence, 1 × 1 × 1 mm voxels) was also collected.

Preprocessing and analysis were conducted with SPM8 (Wellcome Trust Centre for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/). Region of interest (ROI) analyses were conducted with the MarsBaR toolbox (http://marsbar.sourceforge.net/). Functional images were realigned for head motion, coregistered across runs and to the structural image, resampled to 2 × 2 × 2 mm voxels and smoothed with an 8-mm FWHM Gaussian kernel.

We present analyses in both native subject brain space (for those involving ROIs based on image category specific activity estimated per subject from separate localizer runs), and normalized to the Montreal Neurological Institute (MNI) template (with SPM8 "segment and normalize"). In all general linear model (GLM) analyses of the fMRI time series, the first seven volumes of each run were discarded for equilibration, and data were high-pass-filtered with bandwidth 128 s. Additionally, the six motion parameters from realignment and any scanner spike artifacts were added as nuisance regressors to all GLMs. Regressors of interest were convolved with the canonical hemodynamic response function.

fMRI analysis. *Localizer model and ROI selection.* We first estimated a GLM of the time series in the four localizer runs (two pre- and two post-task runs). This model contained separate 14 s boxcar regressors for each of the five categories of stimuli (*faces*, *tools*, *body parts*, *scenes* and *scrambled tools*; only the first four were used in the construction of ROIs). Our analysis of prospective activation (GLM1) focused on ROIs that showed image category selective responses to the stimulus categories used in the second stage of the decision task (body parts and scenes). To restrict the body part– and scene-sensitive ROIs to regions most strongly associated with these stimulus categories and to eliminate any potential confound caused by categories recruiting overlapping brain regions, each ROI was obtained by taking the intersection of the three relevant contrasts in the temporal and occipital lobes (for example, scene ROI: intersection of *scenes* > *body parts*, *scenes* > *faces*, *scenes* > *tools*). These ROIs were obtained in each subject's native brain space, where each contrast was thresholded at $P < 0.001$ (**Fig. 3a**,**b**; see **Supplementary Fig. 2** for group level depiction of these ROIs). This approach did not identify ROIs for body parts in two subjects or for scenes for one other (non-overlapping) subject.

Two-stage decision task models. We estimated four GLMs of the two-stage decision task. With GLM1, we sought evidence of prospective activation. To assess prospective activation, GLM1 contained delta (stick) regressors at first-stage (face and tool states) onset on catch trials indicating which of the two second-stage states subjects had chosen to visit (regressors: *choose scenes* and *choose body parts*). The model also contained separate regressors for each category of on-screen task event (*faces*, *tools*, *body parts*, *scenes*, *aborted trials*, *wins*, *losses* and *catch trial gray boxes*; each modeled with boxcars spanning the duration of each on-screen event). This model was estimated in native brain space for each subject for the average response across voxels in each of the relevant ROIs (scenes and body parts) derived from functional localizer data. We assessed the contrasts *choose scenes* > *choose body parts* in the scenes ROI and *choose body parts* > *choose scenes* in the body part ROI. The averaged results in the two ROIs for each subject provided an index of prospective activation, which was taken to the group level via entry as a covariate into the logistic regression analyses (**Fig. 3c**) and via correlation with the weighting parameter w estimated from the computational model fit (**Supplementary Table 1**). Three subjects had ROIs for only one of the categories (see above). In these cases, the estimated contrast for that single ROI served as the prospective index.

With GLM2 we assessed (in brains normalized to MNI coordinate space) BOLD correlations with the choice values of the two strategies at the first task stage (**Fig. 4**), and also with the model-free reward prediction errors at the second task stage (**Fig. 5**). These parametric time series were derived from the full reinforcement learning model (see "Computational model" below), using each subject's choice data and a single set of free parameters fit across all subjects (excepting the strategy weighting parameter w, which was set to 0 and 1), as in previous studies[12].

In the first task stage, we modeled reinforcement learning model–estimated choice probabilities (see equation (7) for the two approaches). Specifically, delta regressors and parametric covariates were entered for the model-derived probabilities of selecting the chosen stimulus under the model-free ($w = 0$) account, and also for the difference that arises when subtracting those probabilities from the probabilities given by the model-based ($w = 1$) strategy[12]. Note that by using regressors for both $w = 1$ and $w = 0$, GLM2 interpolates between these two extremes by fitting the neural activity with the weighted sum of these two regressors, thereby approximating the response for any w (ref. 12). In particular, the estimated weighting on the regressor encoding the difference in probabilities captures the extent to which the net neural activity more closely resembles the model-based rather than the model-free probabilities. In the second task stage, only model-free ($w = 0$) reward prediction errors were modeled (these errors are always zero under a model-based account; see equation (4)). GLM2 additionally contained nuisance regressors for the prediction errors at reward time (which are identical for model-based and model-free approaches; see equation (2)).

GLM2 effects were first taken to a second-level analysis by way of a one-sample t-test at each voxel (**Figs. 4a**,**b** and **5a**). Effects were related to behavioral and neural quantities across subjects in two ways: first, by entry of per-subject covariates (neural prospection and model-based choice) into the whole-brain second-level analysis (**Fig. 5b**,**c**, left), and second, by reestimating, in each subject, the effects averaged across voxels in functional ROIs (putamen, dmPFC, FPC) defined by the group-level effects (**Fig. 4b**,**5a**). These resulting neural summary scores for each subject were then correlated with neural (**Fig. 5c**, right) measures and with behavior (**Figs. 4c**,**d** and **5b**, right) via entry as a covariate into logistic regression analyses.

The putamen ROI (used in **Fig. 5b**,**c**, left) was defined as the clusters correlating with prediction error (**Fig. 5a**) that survived small-volume correction in an anatomical mask of striatum ($P < 0.05$). The dmPFC and PFC ROIs (used in **Fig. 4c**,**d**) were defined as the clusters correlating with the difference in model-based and model-free choice probability (**Fig. 4b**) surviving correction for whole-brain family-wise error due to multiple comparisons at $P < 0.05$. Note that the multiple comparisons involved in selecting the ROI on the basis of their being an overall main effect, on average across subjects, do not bias the subsequent, independent test of whether this contrast covaries, from subject to subject, with either the degree to which their choices were model-based versus model-free or the degree to which neural prospection was observed.

GLM3 was designed to assess the relationship of neural prospection to neural correlates of first-task-stage model-based valuation across trials. This model was computed separately for the average response across voxels in dmPFC and FPC. Functional ROIs in these regions were obtained from the negative main effect (**Fig. 4b**) of model-based valuation (see equation (7) and GLM2). To generate regressors for GLM3, we first estimated a trial-by-trial time series of neural prospection on catch trials in a model identical to GLM1 except for a unique prospection regressor on each catch trial. The resulting parametric prospection time series was then entered into GLM3. Also entered were the model-based and model-free first-stage relative valuations (on catch trials) for the unchosen stimulus (such that the effect was oriented positively) and the interaction of the valuations and prospection. The regressors of interest in GLM3 were the interaction of each of the valuation time series with the prospection series, which assess whether these quantities covary with one another across trials. This model included the same nuisance regressors as GLM2, and, because prospection was also entered, regressors for all on-screen visual events (as in GLM1) were also included.

With GLM4, we assessed activation that accompanied changes in first-stage state from trial to trial (**Supplementary Fig. 1**). This model was identical to GLM1, with two exceptions. First, it was conducted on images normalized to MNI coordinate space. Second, instead of delta regressors for expected outcomes at first-stage onset time on catch trials, there were instead delta regressors on each trial onset indicting whether the first-stage state was the same as or different from the first-stage state on the previous trial (irrespective of whether the state was faces or tools). This contrast (*different* > *same*) was estimated for each subject and taken to a second-level analysis by way of a one-sample t-test at each voxel (**Supplementary Fig. 1a**). The model was reestimated in the left inferior frontal gyrus cluster surviving correction for multiple comparisons. The resulting individual subject contrast estimates were taken to the group level by entry as a covariate into the multilevel logistic regression (**Supplementary Fig. 1b**) and by correlation with the weighting parameter w estimated from the computational model fit (**Supplementary Table 1**). Note that, as in GLM2, this ROI definition scheme does not bias the independent across subject correlations.

Computational model. In addition to the simplified linear regression analysis, which examines signatures of model-based and model-free learning in terms of how each trial's feedback affects the next choice, we also fit each subject's trial-by-trial choices using a full reinforcement learning model, in which the likelihood of choosing either action depends on its value, which is learned from the full sequence of rewards preceding it rather than just the single preceding trial (**Supplementary Table 1**). The model (a variant of a model used for many similar multi-step tasks[11,12,22]) explains choices as arising from a weighted combination of values learned according to both model-based and model-free approaches, and in fitting the model we estimate the relative weight of each sort of learning that best explains each subject's choices.

The two-stage decision task dissociates the predictions of the model-based and model-free approaches, though in learning from the immediate rewards that follow second-stage choices there are no further state transitions remaining and so the approaches coincide. Here, for each of the two second-stage task states $s2$ (body parts, scenes), state-action values $Q2$ are learned for each action $a2$ (each second-stage state having two such actions: $a2 \in \{a2X, a2Y\}$). These $Q2$ values are updated on each trial t when chosen

$$Q2(s2_t, a2_t) = Q2(s2_t, a2_t) + \alpha_2 \delta_{2,t} \tag{1}$$

where

$$\delta_{2,t} = r_t - Q2(s2_t, a2_t) \tag{2}$$

and α_2 is a free learning rate parameter, controlling the effect of reward prediction error δ_2 on action value update (received reward $r = 0$ or 1).

At the first task stage, the model-based and model-free strategies differ. The model-free approach, SARSA(λ) (state-action-reward-state-action temporal difference learning with eligibility trace parameter λ), learns a value, Q_{MF}, for each action $a1$ at each of the two first stage states $s1$ (faces, tools). The chosen action value in each state is updated on the basis of its outcome on each trial t as

$$Q_{\mathrm{MF}}(s1_t, a1_t) = Q_{\mathrm{MF}}(s1_t, a1_t) + \alpha_1 \delta_{1,t} + \lambda \alpha_1 \delta_{2,t} \tag{3}$$

where

$$\delta_{1,t} = Q2(s2_t, a2_t) - Q_{\mathrm{MF}}(s1_t, a1_t) \tag{4}$$

and α_1 is a free learning-rate parameter. Note that the second-stage prediction error δ_2 defined above differs from the model-free first-stage prediction error δ_1 defined here, as rewards are available following choice only in the second stage. The prediction error following reward in the second stage, δ_2, is used to update the value of the selected second-stage action value $Q2$ as written above, and also the value of the selected first-stage action by way of free eligibility trace parameter λ. The intermediate stage model-free prediction error defined here (which is nonzero for a model-free but not a model-based learner) is entered as a parametric time series in our neural analyses (**Fig. 5a**; GLM2).

Note that the model-free component updates the value of actions only when those actions are selected, and is thus unable to leverage the equivalency of the two first-stage states in making choices. The values of the equivalent actions in the tool and face states are each determined entirely and separately by their unique history of outcomes.

The model-based component uses a representation of the task's state-action transition contingencies to prospectively compute the value of actions on the basis of the states and rewards that are expected to follow each action. In the current task, this corresponds to, at the first stage, for each candidate action, computing the estimates of rewards available at the second stage that would arise given the choice of that action. Thus, for each first stage state $s1$ and action $a1$, the model-based action value Q_{MB} is

$$Q_{\mathrm{MB}}(s1_t, a1_t) = \max_{a' \in \{a2X, a2Y\}} Q2(S(s1_t, a1_t), a') \tag{5}$$

where $S(s1, a1)$ is the second-stage state that would be produced by choosing action $a1$ in the first-stage state $s1$, and the max measures the reward expected for whichever choice is believed to be better at that stage. Because the model-based component uses the task transition structure to evaluate first-stage actions in terms of the second stages to which they transition, this component effectively generalizes across the equivalent first-stage states in the task.

Finally, to model choices, the values of the two approaches in the first stage are combined according to free weighting parameter w (at $w = 1$, choice is exclusively model-based; at $w = 0$, exclusively model-free) as

$$Q_{\mathrm{net}}(s1_t, a1_t) = wQ_{\mathrm{MB}}(s1_t, a1_t) + (1 - w)Q_{\mathrm{MF}}(s1_t, a1_t) \tag{6}$$

At the second stage, the model-based and model-free solutions are identical, so $Q_{\mathrm{net}} = Q2$. The action values are used to compute the probability of each action via the softmax choice rule

$$P(a_t = a | s_t) = \frac{\exp(\beta Q_{\mathrm{net}}(s_t, a))}{\sum_{a'} \exp(\beta Q_{\mathrm{net}}(s_t, a'))} \tag{7}$$

As this rule applies equivalently to the first and second stages, we denote actions a and states s generalized across stages ($a \in \{a1, a2\}$, $s \in \{s1, s2\}$). The softmax inverse temperature β controlling the noise in choices is estimated separately at the two task stages (β_1, β_2). The softmax choice probabilities for the first-stage chosen stimulus under a model-free ($w = 0$) and the difference in a model-based ($w = 1$) and model-free account were entered as a parametric time series in our neural analysis (**Fig. 4a**,**b**; GLM2)

Computational model estimation. We estimated the six free parameters (β_1, β_2, α_1, α_2, λ, w; **Supplementary Table 1**) of the hybrid model for each subject by maximizing the log posterior of the choice data conditioned on the outcomes acquired. We took as uninformative priors over the parameters, gamma(1.2, 5) for softmax temperatures (β_1, β_2), and beta(1.1, 1.1) for the remaining parameters (α_1, α_2, λ, w), consistent with previously observed estimates and reports[12], and ensuring smooth parameter boundaries[50].

As in previous studies[12], we repeated this fitting procedure for fMRI analysis with the difference that a single set of parameters was fit across all subjects. These estimated parameters were used to generate time series for each subject that were entered into the general linear model of BOLD response.

A **Supplementary Methods Checklist** is available.

48. Otto, A.R., Gershman, S.J., Markman, A.B. & Daw, N.D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
49. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
50. Daw, N.D. in *Atten. Perform. XXIII* (Delgado, M.R., Phelps, E.A. & Robbins, T.W.) 1–26 (Oxford University Press, 2011).