Mapping value based planning and extensively trained choice in the human brain

Klaus Wunderlich[1], Peter Dayan[2] & Raymond J Dolan[1]

Investigations of the underlying mechanisms of choice in humans have focused on learning from prediction errors, leaving the computational structure of value based planning comparatively underexplored. Using behavioral and neuroimaging analyses of a minimax decision task, we found that the computational processes underlying forward planning are expressed in the anterior caudate nucleus as values of individual branching steps in a decision tree. In contrast, values represented in the putamen pertain solely to values learned during extensive training. During actual choice, both striatal areas showed a functional coupling to ventromedial prefrontal cortex, consistent with this region acting as a value comparator. Our findings point toward an architecture of choice in which segregated value systems operate in parallel in the striatum for planning and extensively trained choices, with medial prefrontal cortex integrating their outputs.

An overarching view of adaptive behavior is that humans and animals seek to maximize reward and minimize punishment in their choices. Solutions to such value-based decision problems fall along a crude spectrum. At one end, on-the-fly planning, based on a model of the relevant domain, can determine which of the available actions lead to a desired outcome. Finding optimal actions in this type of choice context, for instance, by searching the branches of a decision tree for the best outcome, poses severe demands on computation and memory and rapidly becomes intractable with growing complexity. This end of the planning spectrum is of particular importance when we have relatively little experience in an environment or where its aspects change quickly.

In contrast, when subjects have extensive practice in a relatively stable domain, they can directly learn from experience about the affective consequences of different actions. Decision-making at this end of the spectrum can become highly automated and need no longer be based on a complex representational model of the world. One of the main results in the field of reinforcement learning[1], and indeed one of the earliest insights in artificial intelligence[2], is that it is possible to learn optimal actions in complex, but stable, domains by making and measuring errors in predictions over the course of extended experience.

A rich body of work on value-based decision-making in humans has focused on learning on the basis of prediction errors[3,4]. Although there has been extensive investigation of various tasks involving planning, such as the Tower of London[5], these tasks have typically not focused on value and have not been designed to compare the two ends of the spectrum referred to above. More recent investigations targeting this spectrum[6–8] have been revealing, but have not directly addressed the computational mechanisms or neural encoding of value-based planning, or the integration of extensive training and planning–based evaluation and choice.

We designed a value-based choice task for human subjects and used functional magnetic resonance imaging (fMRI) to examine the neural mechanisms underlying forward planning and choices based on learned values after extensive behavioral training. Our task allowed us to separately index planning and extensively trained contexts, which allowed us to specifically investigate value representations in the brain associated with the computational processes of each type of choice. We found that medial striatum was more strongly engaged during planning and lateral striatum was more strongly engaged during choices in extensively trained contexts. Notably, the blood oxygen level–dependent signals in caudate pertained to individual computational components of planned choice values, whereas signals in posterior putamen selectively fluctuated with the values during extensively trained responses. Our results provide direct evidence in humans for multiple decision systems that operate independently and in parallel and recruit neural structures along a medio-lateral axis in basal ganglia. Furthermore, prefrontal cortex, specifically ventromedial prefrontal cortex (vmPFC), represented the value of the chosen option across systems, highlighting its possible role as a value comparator across both decision systems.

RESULTS

We asked 21 subjects to participate in a decision task in which decision values that were either derived from forward planning or learned through extensive training could be distinguished on a trial-by-trial basis. One component of the task required subjects to navigate a tree-shaped branching maze to reach one of several available terminal states. Each state was associated with distinct probabilities of obtaining reward, thus rendering the value components of individual branches in the decision tree computationally transparent. In pure planning trials (**Fig. 1a**), probabilities of reward were visually displayed, but could change on each trial. Three consecutive choices led

Figure 1 Task and behavioral results. (**a**) Task flow in planning trials: subjects navigated a three-layer maze before reaching probabilistic rewards. Eight numbers (randomly changing from trial to trial) displayed reward probabilities of each terminal room. Second layer choice was determined by a deterministic value minimizing computer agent that implemented the lowest value option. (**b**) Exemplary planning maze: nodes represent rooms and lines represent transitions between rooms. Subjects moved forward by freely choosing at the first and third level (cyan circles); the computer determined choice at the second level (gray circles). The optimal path (arrows) was determined by backward induction of state values using a minimax strategy. State (black) and action (red) values are shown along the choice path. (**c**) Prior training over 3 d in four single-level mazes with invariant contingencies and distinct reward probabilities ($P = 0.15, 0.40, 0.65$ and 0.90). Wall colors provided distinguishable contexts that allowed subjects to distinctively identify each trained maze. No explicit information about reward probabilities or contingencies was given. (**d**) Combination of planned and trained options in the same trial; colored doors transitioned into trained maze of same color, the other door followed reduced planning branch with four outcome states. No reward probabilities were shown above colored doors. (**e**) Average fraction of correct choices according to a tree search planning strategy (plan) and two alternative heuristics: highest value (max) and higher average value (avg). Subjects' choice behavior pertained to tree search planning strategy and could not be explained by any of the heuristics. Vertical lines represent s.e.m. See **Supplementary Table 2** for individual subject behavior.
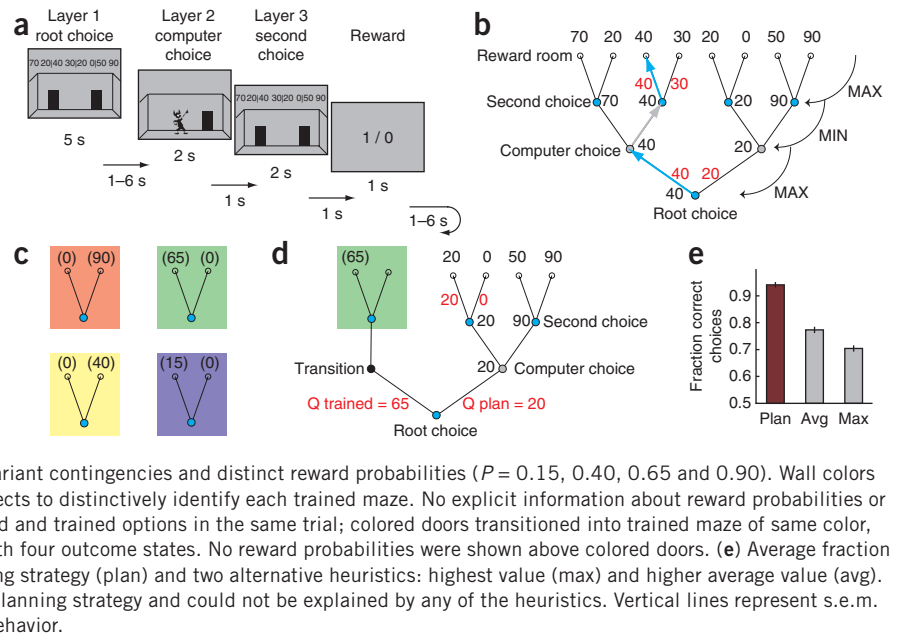
from the start state to the terminal state. Subjects planned the first and last choices; the middle choice was made by a predictable computer agent acting according to a fully disclosed rule (minimax, selecting the tree branch having the lower maximum value). This latter step induced a tree search strategy for calculating planned values, whereas, for instance, a mere requirement to compare displayed values might fail to invoke sufficient forward planning. Thus, the best possible choice in the task required a form of dynamic programming, involving the estimation of values at distinct stages in the tree (**Fig. 1b**).
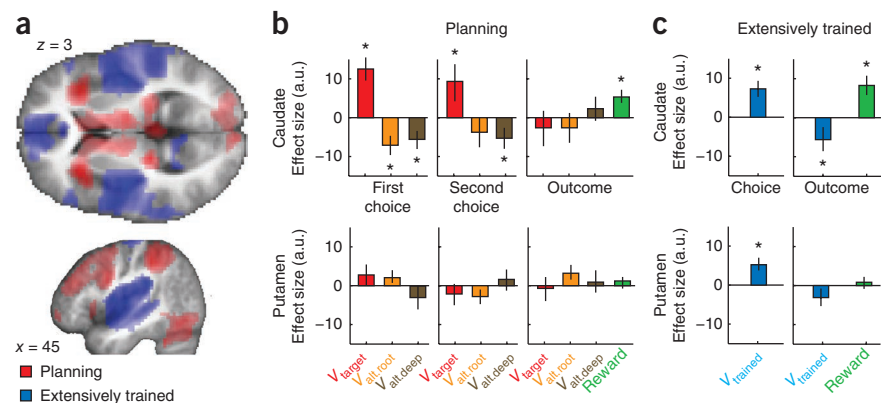
A second component of the overall task design involved trials that did not require forward planning and were instead extensively exercised during 3 d of behavioral training. In these, subjects had to make single choices between two available actions after having learned values from samples of a probabilistic reward delivery process (**Fig. 1c**). The inclusion of separate planning and extensively trained trials allowed us to investigate neural computations unique to each decision system. Subsequent to this, we examined mixed trials involving choices between either a planning or

a trained branch (**Fig. 1d**) and choices between two trained branches. From a normative perspective, the combination of both components in the same trial entails a direct comparison of planned values from one branch with values derived from the extensively trained task on the other branch.

Behavioral results

Subjects' choices were largely consistent with choice values over all trial types (**Supplementary Table 1**), confirming that planning was cognitively tractable, and consistent with subjects having learned values and action mappings in the trained trials. On average, subjects chose optimally in 94% of planning trials and chose the rewarding door in 98% of extensively trained trials. Our task was designed such that only a tree search strategy would yield good performance in planning trials. To test whether participants indeed used a tree search strategy, we compared individual subjects' choices to the optimal minimax strategy and to other simpler heuristics, such as picking the path with the largest maximum value later in the tree or picking the path with the highest average value

Figure 2 Neural correlates of planning versus extensively trained choices. (**a**) Significant categorical effects for planning > trained trials (red) and trained > planning trials (blue). Medial sectors of basal ganglia, including medial caudate, thalamus, bilateral anterior insula, dorsomedial prefrontal cortex, bilateral medial frontal gyrus and precuneus, showed enhanced BOLD responses on planning compared with extensively trained trials. Lateral posterior putamen, posterior insula extending into the medial temporal gyrus and somatosensory cortex, including postcentral gyrus, were more activated when subjects made a response in the extensively trained context. (**b**) Effect size plots in a regression of planned values, convolved with a canonical HRF, against BOLD data at three

time points: first choice, subjects' second choice and outcome. Signals in caudate pertained to the value difference between actual target and alternative values in the choices along the traversed path, as indicated by both significant positive effects for target values and significant negative effects for the alternative values. Asterisks mark significant effects ($P < 0.05$; see **Supplementary Table 5** for individual effect sizes and P values), a.u. = arbitrary units; vertical lines represent s.e.m. Posterior putamen did not significantly correlate with planned values. (**c**) Caudate activity related to classic reward prediction errors during trained trials. Posterior putamen showed significant value representations in extensively trained mazes at time of choice.

Figure 3 Comparing values from planning and values from extensively trained mazes. Value-based effect sizes at choice time in mixed planning/trained trials and trials comparing two trained branches. Mixed trials are separately plotted conditional on subjects' choices. V_c = value of the chosen option, V_{nc} = value of the not chosen option. (**a**) Caudate represented planned values of the planning branch regardless of choice. (**b**) Putamen fluctuated with values of the colored trained branch regardless of choice. (**c**) vmPFC encoded the value of the chosen option, representing the output of a comparison process. *$P < 0.05$; a.u., arbitrary units; vertical lines represent s.e.m.

in the leaf nodes. Subjects' behavior was better explained by the minimax planning strategy than any of the alternative heuristics ($P < 10^{-8}$, Wilcoxon rank sum test; **Fig. 1e**). Moreover, choices in every individual subject matched choices predicted by the planning strategy more closely than those predicted by the heuristics (**Supplementary Table 2**).

Before undergoing fMRI, subjects were trained for 3 d on extensively trained trials to ensure that associated values had stabilized. Over the course of that training, subjects' responses converged to the optimal action in each context; there was no difference between the rate of correct responses in higher and lower valued contexts from day 2 onwards (**Supplementary Fig. 1**).

Categorical neural differences: planning versus trained choices

We first compared activity at the time of initial choice during planning trials with activity in trials involving extensively trained choices. Activity dissociated along an anteromedio-posterolateral axis in basal ganglia (**Fig. 2a**). Structures that were preferentially activated during planning included caudate and medial striatum, thalamus, bilateral anterior insula, dorsomedial prefrontal cortex, dorsolateral prefrontal cortex, bilateral medial frontal gyrus and parietal cortex (precuneus extending into intraparietal sulcus). In contrast, lateral posterior putamen, posterior insula extending into the medial temporal gyrus, vmPFC and somatosensory postcentral gyrus were more strongly activated in trained trials (all $P < 0.05$ familywise error corrected; **Supplementary Table 3**). Anatomically defined region of interest (ROI) analyses confirmed that BOLD responses in caudate increased significantly only during planning trials, whereas posterior putamen activity was selective to trained trials.

Neural correlates of choice relevant values

As choice crucially depends on value, we next investigated neural responses pertaining to valuations of available choices in two striatal regions that are strongly linked to decision-making, namely anterior caudate nucleus, implicated in goal directed choices[9,10], and posterior putamen, which has been associated with overtrained choices[8]. In addition, we examined responses in vmPFC, a region also widely implicated in value-based choice[11–13]. We delineated ROIs a priori (see **Supplementary Fig. 2** for location details) based on previous research and anatomical criteria and regressed various values against fMRI signals in these regions. One set of values concerned were those of the
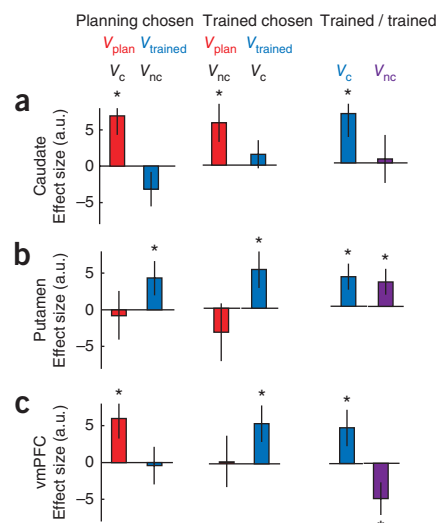
target (the choice leading to the best reachable outcome, taking account of the computer's minimax strategy) and of the alternative choices along the traversed maze path. This was motivated by the fact that these are the values that need to be compared during tree search. Indeed, consistent with this hypothesis, fMRI signals in the caudate covaried with the difference between target and alternative values, as shown by the significant positive effects for target and negative effects for the alternatives ($P < 0.05$; **Fig. 2b** and **Supplementary Table 5**). Notably, during the root choice, caudate activity related to several values relevant for a given choice, including those at the present ($V_{target} - V_{alt.root}$), and to the consecutive choice deeper in the tree ($V_{target} - V_{alt.deep}$). Note that successful forward search of the decision tree required a consideration of the latter values even while at the root state. During the subjects' second choice (layer 3), caudate activity was still associated with the values of both alternatives at the now current choice, but was no longer associated with the value of the previously rejected root branch. These value difference signals are likely to reflect the output of value comparisons, a predicted hallmark of a cognitive implementation of tree search. The effects seen in caudate for planning value components were not evident in posterior putamen. Instead, the putamen solely encoded values on extensively trained trials at the time of choice (**Fig. 2c**).

We next examined how the two networks interact in decisions that require a comparison of the respective values represented in these two distinct clusters. We presented subjects with a choice between a planning branch and a trained branch (**Fig. 1d**), that is, a task in which subjects need to access both planned and trained values. By design, the value of the trained branch was uncorrelated with the values of the planning branch, which allowed us to distinguish the influence of both value types. The caudate consistently represented the difference between the planned target value and the value of the alternative option on the planning branch, consistent with it performing the same value difference computations as in pure planning trials. Notably, the caudate represented these planned values and not the
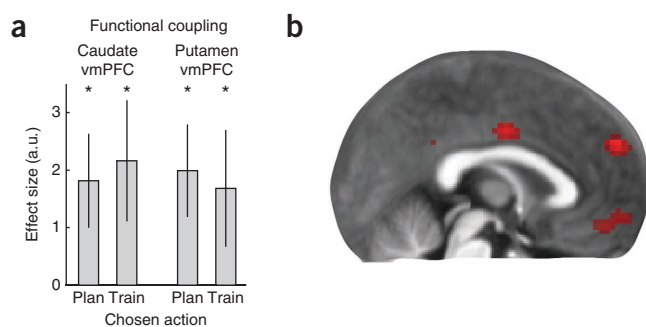

Figure 4 Functional coupling between caudate-vmPFC and putamen-vmPFC is significantly increased during choice in mixed trials. (**a**) We tested statistical significance of the PPI interaction contrast between our a priori defined ROIs, for which the effect size is shown as bar graphs (all $P < 0.05$). Vertical lines represent s.e.m. The increase in coupling was independent of actual choice, consistent with the hypothesis that vmPFC mediates in the decision process by accessing pre-choice values from both choice systems. (**b**) Shown are areas of increased coupling with both caudate and putamen during mixed choices (conjunction analysis).

values of the trained branch, regardless of which branch was later chosen (**Fig. 3a**). In contrast, activity in putamen pertained solely to values of the available trained branches, also irrespective of later choice (**Fig. 3b**). The putamen also represented the stimulus values of both available actions in trials comparing values from two trained branches.

The finding that activity in caudate and putamen covaried with planned values and values from the trained trials respectively, even for actions that were not chosen, provides direct evidence for a parallel and independent operation of two separate controllers. In turn, this parallel operation afforded us the opportunity to examine how these systems compete at the time of choice. Similar to action values[14,15], the striatal correlates of planned values and values of the trained branches fulfill the criteria for pre-choice values and are likely to serve as inputs to a final decision comparator. The region most commonly implicated in comparative valuation is the vmPFC[16-19]. We observed that vmPFC activity covaried with the value of the chosen branch, a post choice signal, irrespective of whether it was planned or trained (**Fig. 3c**). Notably, we found no evidence for the representation of mere stimulus values in the vmPFC cluster: if vmPFC activity had related to some form of representation of both value options (or their sum), then we would have expected to see a positive effect for both chosen and unchosen values in this contrast. Furthermore, we ruled out that the vmPFC signals represented the best option (maximum value) rather than the chosen option by re-estimating our general linear model (GLM) with a maximum choice value regressor and performing Bayesian model comparison[20] between both GLMs. This analysis provided strong evidence for a choice-related signal in both mixed and trained/trained trials (exceedance probability >0.99).

Functional coupling between basal ganglia and vmPFC

In mixed trials, signals in caudate and putamen consistently pertained to the value of the same system, independent of choice, whereas vmPFC pertained to the value that was modulated by choice. In other words, activity in vmPFC depended on choice, whereas activity in putamen and caudate did not. This suggests the caudate and putamen are at an input stage to a value comparison process, whereas vmPFC is at an output stage.

To discriminate between alternative mechanisms for how choice values from both systems are compared, we used a connectivity analysis, derived from a psychophysiological interaction (PPI), and examined the functional relationship between vmPFC, caudate and putamen during mixed choices. One possibility is that the competition between the planning and extensive trained system is resolved in the basal ganglia and the outcome is transferred to vmPFC. This predicts that the PPI will show increased coupling of only the winning area (caudate or putamen) with vmPFC. The alternative hypothesis is that values from both areas are transferred to vmPFC and that competition is resolved in vmPFC. This predicts increased coupling between both precursor areas and vmPFC, regardless of choice.

The results of the PPI analysis support the latter pattern hypothesis, revealing a significant increase in the strength of coupling of both putamen and caudate with vmPFC during the time of choice, independent of the action that was finally chosen ($P < 0.05$; **Fig. 4**). In contrast, we did not find areas that showed a differential increased coupling with caudate on trials in which subjects chose the planning branch, but not on trials in which they chose the trained branch, or a differential increased coupling with posterior putamen on trained branch choices, but not on planning branch choices.

DISCUSSION

We found that behavior on trials invoking forward planning and trials with extensively trained options evoke activity in distinct neural systems during computations associated with choice. BOLD signals in the caudate pertained to values of the individual branches in a decision tree, whereas BOLD signals in the posterior putamen fluctuated with values associated with responses in an extensively trained context. Notably, during choices requiring a simultaneous comparison of values from both choice types, the individual striatal subsystems consistently represented their respective values regardless of final choice. These findings suggest that two independent systems represent the two choice types in our task. In contrast, activity in prefrontal cortex pertained to a value signal that depended on the actual decision that was made.

Converging evidence from animal and human studies has long suggested that two different learning processes govern behavior: one controlling the acquisition of goal-directed actions and one controlling the acquisition of habits[21,22]. According to this dissociation, an association between actions and outcomes governs action selection in the goal-directed case, whereas it is controlled through learned stimulus-response associations without any direct assessment of the outcome of those actions in the habitual case. As such, goal-directed control is performed with regard to the consequence of actions, whereas habits are determined by the predicting stimuli rather than the outcomes. Accounts suggesting a plurality of control are also supported by theoretical considerations of the computational mechanisms underlying different forms of reinforcement learning[23,24]. The defining criterion in the more computationally centered literature has been a functional one, focused on the differences in the computational mechanisms underlying different types of learning. The dissociation that we used is between model-free temporal difference learning of cached values and model-based choice that predict, on the fly, the immediate consequences of each action in a sequence. Our planning task, which was designed to be only solvable by searching the decision tree, typifies model-based control. The absence of a devaluation or contingency degradation test means that we cannot definitively prove that our extensive training created a true habit. Similarly, we cannot exclude the notion that subjects derived values in the extensively trained mazes by solving a decision tree during training and then memorizing it so that it could be retrieved at the time of choice. However, similar tasks in previous studies have shown that learning through numerous repetitions in stable contexts is solved by a prediction error–based mechanism[3,25-28].

In the caudate, we observed value differences, which are likely correlates of the choice values during the planning process. The existence of planning value representations in anterior caudate is consistent with evidence for goal-directed impairment after caudate lesions in rodents[29]. In addition, a human imaging study found elevated activity in anterior caudate when subjects were performing on a high-contingency schedule compared with when they were performing on a low-contingency schedule[10].

Although most of our results are consistent with previous findings implicating the caudate in explicit planning, it is notable that the BOLD signal in this structure also correlated with the values of the relevant options in extensively trained trials. There are a number of possible explanations for this. The simplest is that this activity is epiphenomenal for choice. That is, the main claim of dual systems accounts is not that redundant systems do not calculate (if they have the information to do so), but rather that their calculations do not influence behavior. Thus, in extensively trained trials, a planning system might estimate values, but with no effect or only a modest effect on behavior, or perhaps at most improving the prediction errors available to the other system[6]. When the planning system is engaged in its own unique computations, these calculations are no longer possible. We consider the mixed trials as showing this, although it would be interesting to design a more explicit test,

for instance, engaging the planning system with a distractor task while subjects make extensively trained choices. In such a scenario, we would expect value-associated signals in caudate to vanish, or rather to pertain to the concurrent planning task, while leaving choice performance on the extensively trained task essentially unimpaired. Diametrically opposed to this interpretation is the possibility that the caudate actually controls choice, even in trials that we consider to be non-planning. We believe this is unlikely, as the value of the trained branch was conspicuously absent from caudate in mixed trials, whereas if subjects based all choices on planning then we should have seen a value difference in the caudate, similar to the pattern of activity observed in pure planning trials. A third, and more radical, possibility is that the values from the trained branches are used to ground evaluations in the planning system. This interpretation would be most appropriate for trials involving two trained branches, as these values could then be compared by the planning system. Such an integration of values across systems has been widely predicted from the very earliest days of planning[2,30], but has not previously been observed. Our task is not ideal for testing this possibility, but it suggests an important area for future work. On a similar note, although we found that prediction error–based learning of action values did not affect choice in planning trials, we cannot exclude some form of concomitant model-free learning even in planning trials. However, in the absence of an overt expression of behavior from the model-free system, this would not be trivial to dissect using our methods. These questions are nevertheless important issues for future research.

The putamen encoded values associated with the extensively trained trials throughout our study. A recent imaging study[8] showed that cue-driven activation in dorsolateral posterior putamen increases with prolonged habitization and concluded that this region may contribute to the habitual control of behavior in humans. Although that study did not investigate the value-related parametric effects in this region, we found neural representations of values for extensively trained choices in the same area. However, it is less clear whether there is a process of consolidation by which values migrate in the striatum over the course of overtraining. Our ROI (which was based on the coordinates in ref. 8) was posterior to the location of many previous studies reporting prediction error signals in putamen during basic learning tasks[4,27,31]. When we tested for value signals in an ROI in anterior putamen, we did not find a reliable significant representation for values of the trained branches or planned values. This is consistent with evidence from studies on procedural sequence learning[32,33], suggesting a transfer of activity from rostral to more caudal parts of putamen with increasing learning.

Our data (including the behavioral effect that higher values in those trials reduce response times; **Supplementary Table 1**) also suggest that even extensively trained responses can still be influenced by learned values of the associated actions, rather than depending only on the sort of more arbitrary action propensities found in certain reinforcement learning models (notably the actor-critic[34]). It is interesting to note that in neither of the tasks did we observe a value difference for the extensively trained choices in posterior putamen, which might function as reinforcement learning cached memory. This is particularly clear in trials with two trained branches, where the values of both available options were simultaneously represented in putamen. This pattern of option values, but not a value difference, suggests that the putamen does not compare values, but needs the vmPFC or caudate (where we see such a difference between the chosen and unchosen option) to perform this task. It should also be noted that posterior putamen did not reflect a prediction error at the time of outcome (unlike caudate), which might underlie the persistence of extensively trained habits.

Our finding that vmPFC increased its coupling with both caudate and putamen during choice and encoded the winning outcome of a choice process (chosen value) is consistent with it having a putative role as a value comparator. These results challenge the view that prefrontal cortex is largely sensitive to model-derived values[10,11,35] and instead suggest that the vmPFC is engaged whenever values are compared, to prepare an action, regardless of whether this derives from a planning computation or from extensive training. The absence of vmPFC value representations during extensively trained choice trials, which do not require a comparison, implies that subjects immediately initiate the action in these trials. This interpretation would also explain why vmPFC does not represent choice values at the third stage of planning trials, as subjects might have already precomputed and stored deep choices at the root stage and then only executed the appropriate response at the deep stage. Our behavioral findings support this interpretation: subjects' response times increased with decision difficulty (measured as absolute value difference) at the first stage, but not at the second stage (**Supplementary Table 1**). In contrast, caudate represented values of the second stage choice both at the time when they were computed (root choice) and at the time when the associated action was activated (deep choice), consistent with its proposed role of organizing and representing the forward planning process. In summary, vmPFC might facilitate actual value comparisons, whereas caudate represents the planned actions (together with the planned values) as long as they are task relevant and until the required actions are initiated. Furthermore, it is of interest that vmPFC pertained to the value difference between chosen and unchosen options during choices requiring a comparison of values from only one system (such as pure planning trials and trials involving a comparison between two extensively trained branches), but only to the chosen value in mixed trials. We cannot rule out the possibility that our test is insensitive to the negative effect of the unchosen option during those trials. However, an alternative explanation is that the brain employs different mechanisms for the value comparison in both conditions. This hypothesis requires further investigation, as previous studies have reported vmPFC sensitivity to both chosen values[11,13,31,36,37] and to value differences between chosen and unchosen options[16,19], but whether behavior was guided by planning or non-planning computations was not explicitly controlled for. In addition to chosen values, a number of previous studies found evidence for goods or option values in medial PFC[12,38]. Our overall interpretation suggests a value comparison role for vmPFC after pre-choice values are transferred there from other structures such as the basal ganglia. It is possible that vmPFC also has a separate role in the valuation of economic goods[39,40], in which case it may also reflect stimulus values[41]. We used abstract monetary rewards and our task therefore did not require such an appraisal of real world items in common value space.

We note four caveats to our findings. First, nonsignificant results do not prove the absence of an effect. However, it should be mentioned that neural signals in caudate and putamen did not just correlate with a singular value signal, but instead pertained to a set of specific computationally meaningful patterns across several different tasks. Second, we concentrated on BOLD signals at the time of the choice rather than at the outcome. This was because we had no expectation for the computation at this time for either system. The outcome is irrelevant for planning, as values change on a trial-by-trial basis and the computer opponent's strategy is instructed. For the choices in the extensively trained context, substantial experience with fixed outcome probabilities (unlike the case in ref. 6) should render nugatory any prediction error. Third, although our results in relation to categorical differences between trial types (**Fig. 2a**) might be influenced by

variations in difficulty between conditions, this would not affect the parametric analysis of values, as those potential confounds are encompassed by the associated categorical regressor. Finally, all value signals are relative to the reference frame of the choosing agent[38,42] and any neural representation of values should ultimately reflect subjective values. We assumed that our subjects employed a linear transformation of reward probabilities to value, consistent both with subjects' choices and neural data showing a linear relationship between reward probability and BOLD in our ROIs (**Supplementary Fig. 3**).

Our findings add to recent investigations of value-based choices[6–8,11], suggesting that there are conserved processes in basal ganglia across species. Previous studies were limited with respect to the questions that we posed by either not dissociating value representations of multiple controllers or not involving actual planning[43]. Furthermore, we designed our task to minimize the possible indirect interactions between the two forms of control; for instance, even if the planning system were to calculate temporal difference prediction errors on planning trials (unnecessarily for it)[6], there would be little to do with them, as the values change on a trial-by-trial basis. Perhaps the most pressing possibility furnished by our results is to embed values derived from extensive training deeper in the tree. This would require those learned values to be assessed as part of a planning choice in a more thoroughgoing way than in our trained/trained trials. As mentioned above, that this actually happens is a critical prediction of theories and practice in planning in extended domains, but has never been experimentally tested.

METHODS
Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

AUTHOR CONTRIBUTIONS
K.W. and P.D. conceived the study. K.W. designed the task, performed the experiments and analyzed the data. K.W., P.D. and R.J.D. wrote the paper.

1. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: an Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
2. Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Develop.* **3**, 210–229 (1959).
3. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
4. Seymour, B. *et al.* Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667 (2004).
5. Shallice, T. Specific impairments of planning. *Phil. Trans. R. Soc. Lond. B* **298**, 199–209 (1982).
6. Daw, N.D., Gershman, S.J., Dayan, P., Seymour, B. & Dolan, R.J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
7. 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).
8. 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).
9. Tricomi, E.M., Delgado, M.R. & Fiez, J.A. Modulation of caudate activity by action contingency. *Neuron* **41**, 281–292 (2004).
10. Tanaka, S.C., Balleine, B.W. & O'Doherty, J.P. Calculating consequences: brain systems that encode the causal effects of actions. *J. Neurosci.* **28**, 6750–6755 (2008).
11. Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
12. Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W. & Rangel, A. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* **28**, 5623–5630 (2008).
13. 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).
14. Lau, B. & Glimcher, P.W. Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.* **27**, 14502–14514 (2007).
15. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
16. Boorman, E.D., Behrens, T.E., Woolrich, M.W. & Rushworth, M.F. 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).
17. Noonan, M.P. *et al.* Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl. Acad. Sci. USA* **107**, 20547–20552 (2010).
18. Basten, U., Biele, G., Heekeren, H.R. & Fiebach, C.J. How the brain integrates costs and benefits during decision making. *Proc. Natl. Acad. Sci. USA* **107**, 21767–21772 (2010).
19. FitzGerald, T.H., Seymour, B. & Dolan, R.J. The role of human orbitofrontal cortex in value comparison for incommensurable objects. *J. Neurosci.* **29**, 8388–8395 (2009).
20. Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J. & Friston, K.J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).
21. 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).
22. Redgrave, P. *et al.* Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat. Rev. Neurosci.* **11**, 760–772 (2010).
23. 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).
24. Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974 (1999).
25. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
26. Knutson, B. & Cooper, J.C. Functional magnetic resonance imaging of reward prediction. *Curr. Opin. Neurol.* **18**, 411–417 (2005).
27. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
28. Berns, G.S., McClure, S.M., Pagnoni, G. & Montague, P.R. Predictability modulates human brain response to reward. *J. Neurosci.* **21**, 2793–2798 (2001).
29. Yin, H.H., Ostlund, S.B., Knowlton, B.J. & Balleine, B.W. The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* **22**, 513–523 (2005).
30. Sutton, R.S. First results with Dyna, an interesting architecture for learning, planning, and reacting. in *Neural Networks for Control* (eds. Miller, T, Sutton, R.S. & Werbos, P.) 179–189 (MIT Press, Cambridge, Massachusetts, 1990).
31. Knutson, B., Taylor, J., Kaufman, M., Peterson, R. & Glover, G. Distributed neural representation of expected value. *J. Neurosci.* **25**, 4806–4812 (2005).
32. Jueptner, M., Frith, C.D., Brooks, D.J., Frackowiak, R.S. & Passingham, R.E. Anatomy of motor learning. II. Subcortical structures and learning by trial and error. *J. Neurophysiol.* **77**, 1325–1337 (1997).
33. Lehéricy, S. *et al.* Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. *Proc. Natl. Acad. Sci. USA* **102**, 12566–12571 (2005).
34. Barto, A.G. Adaptive critic and the basal ganglia. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J.C., Davis, J.L. & Beiser, D.G.) 215–232 (MIT Press, Cambridge, Massachusetts, 1995).
35. Valentin, V.V., Dickinson, A. & O'Doherty, J.P. Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* **27**, 4019–4026 (2007).
36. Wunderlich, K., Rangel, A. & O'Doherty, J.P. Neural computations underlying action-based decision making in the human brain. *Proc. Natl. Acad. Sci. USA* **106**, 17199–17204 (2009).
37. Tanaka, S.C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* **7**, 887–893 (2004).
38. Padoa-Schioppa, C. & Assad, J.A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
39. Chib, V.S., Rangel, A., Shimojo, S. & O'Doherty, J.P. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *J. Neurosci.* **29**, 12315–12320 (2009).
40. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.* **27**, 9984–9988 (2007).
41. Wunderlich, K., Rangel, A. & O'Doherty, J.P. Economic choices can be made using only stimulus values. *Proc. Natl. Acad. Sci. USA* **107**, 15005–15010 (2010).
42. Kable, J.W. & Glimcher, P.W. The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* **10**, 1625–1633 (2007).
43. Fitzgerald, T.H., Seymour, B., Bach, D.R. & Dolan, R.J. Differentiable neural substrates for learned and described value and risk. *Curr. Biol.* **20**, 1823–1829 (2010).

ONLINE METHODS

Subjects. 21 healthy subjects (9 female, 18–35 years old) with no history of neurological or psychiatric illness participated in this study. None showed color vision deficiency in the Ishihara test. All subjects participated in 3-d learning of trained mazes and scan session 1 (see below). 20 of the 21 subjects participated in scan session 2. The Institute of Neurology (University College London) Research Ethics Committee approved the study.

Task. Our experiment consisted of four conditions: pure planning (P trials), extensively trained choices (E trials), choices requiring a comparison of planned and extensively trained values (PE trials), and choice between two trained branches (EE trials).

Planning. Subjects navigated through a tree-shaped maze in search of maximal reward. Each state in the decision tree corresponded to a unique room in the maze with state transitions implemented through left and right forward exit doors (backtracking was not possible). Depending on the chosen doors, subjects progressed along different branches in the tree maze until they reached a reward room at the end of each branch (**Fig. 1b**). All participants acquired correct mappings between room transitions and maze positions before the functional imaging experiment. Each reward room contained probabilistic reward, shown to subjects as a chest full with gold coins or an empty chest. The reward probabilities of all terminal rooms were clearly available to subjects throughout the trial as a display of eight numbers at the top of the screen. The reward probabilities fluctuated in discrete 0.1 wide steps between 0 and 1 and were shown to subjects as percentage integer number (in the range [0, 100]). Transitions from state to state within the maze (the spatial layout of the maze) were deterministic and constant throughout the entire experiment. However, the reward probabilities for the eight terminal states changed completely on every planning trial, thereby effectively preventing successful application of model-free learning strategies.

To engage subjects in forward planning over and above a mere comparison of instructed values, the choice at layer 2 in the tree was made by a deterministic value minimizing computer agent. Before the experiment subjects were explicitly instructed about the computer agent's choice rule and, to avoid that subjects considered the computer as social agent, we emphasized that its choice rule would remain deterministic and predictable throughout the experiment. The only rational strategy in this task was to plan the best possible transit through the maze using a mimimax strategy[44] to rollback state value. This involves, already at the root choice, considerations of the choice at the third layer and the computer's choice in each of the two possible rooms in layer 2.

Choices in extensively trained contexts. Each of the four mazes consisted of one choice room with two doors and a reward room behind each door. Only one door led to probabilistic reward and those contingencies never changed throughout the experiment. Different wall coloring (red, yellow, green and blue) provided distinct contexts in which subjects acquired separate value associations from the set of 15, 40, 65 or 90 percent[45]. We counterbalanced mappings between color, reward contingencies, and actions across subjects.

Choices between planning and trained branches. Half of the decision tree was a planning branch with the same rules as in the planning maze; the doorframe in the root node of the other branch was colored and its choice led into the trained maze of that color. This required subjects to directly compare a planned target value from one branch with a trained value. We matched transitions in mixed trials to equate effort and time for traversing either branch. Note that this trial type did not provide subjects with an option to choose whether they would prefer to engage in planning or a choice based on the previously trained mazes. Instead, rational choice always required performance of both a planning part to calculate action values for the planning branch and retrieval of a value for the colored branch, followed by a direct comparison between values from both systems.

Choice between two extensively trained branches. Finally, trials in a fourth condition involved a comparison between two learned values. The root room contained two colored doors and choice of any transitions into the respective maze.

Training of values in colored mazes. To induce stable values in the colored mazes, we informed subjects that each color corresponds to a different maze with its own stable reward probabilities and then trained them on three consecutive days (720 trials in interleaved ordering) before the fMRI scan (**Supplementary Fig. 1**). We did not perform functional imaging during this training phase, but it is well established in numerous animal and human studies[3,25–28] that such a task induces prediction error mediated learning.

fMRI experiment. To prevent a deterioration of responses in trained mazes (PE trials might stimulate a formation of a new explicit value representation for each colored maze, inducing a strategy change on subsequent E trials), we blocked our experiment into two parts and first presented E and P trials and in a subsequent block PE and EE trials.

In scan session 1 we presented subjects with 96 P and 96 E trials, randomly intermixed, to measure choice related brain activity unique to either planning or decisions in extensively trained contexts. After a 15-min break outside the scanner, subjects participated in scan session 2, which contained intermixed 100 PE and 50 EE trials. Subjects' payout related to the earned rewards (£0.20 during fMRI session and £0.05 during training). In total, subjects accumulated approximately £60 in rewards (range £55–64).

Model predicted choice values. We used constant values of the true reward probabilities throughout the study. Due to the large number of training trials, and because subjects universally chose the better option toward the end of training, we can assume that subjects acquired learned values for the colored mazes during the training period and that those values had converged to the true value at the time of the fMRI study (trial-by-trial fluctuations in value would then be minimal due to a very small learning rate, adapted to the stable environment[46]).

Forward planning. We assumed that subjects would unroll values from the reward rooms (instructed on the screen) to every prior state and then plan in the root state the optimal transit through the maze. We modeled this forward search[47] for rewards R and calculated planned values for action a in each state s of Layer $L(s)$ using a maximizing strategy over available choices in states under subjects' control (layer 1 and 3), and a minimizing strategy in states under the computer's control (layer 2).

$$V(s,a) \leftarrow R(s') + \sum_{s'}\left(\max_{a'} V(s',a')\big[L(s) \neq 2\big] + \min_{a'} V(s',a')\big[L(s) = 2\big]\right)$$

Behavioral analysis. To investigate potential motivational (caused by a high target value) and difficulty based influences (originating from small differences between target and alternative values) on choice time we regressed V_{target}, the negative absolute value difference $(-|V_{chosen} - V_{unchosen}|)$, and trial number on logarithmic RT separately for each trial type. Note that we neither instructed subjects to respond quickly nor was it the case that fast responses had any monetary benefit to subjects (except for finishing the experiment slightly sooner). We similarly analyzed the influence of these parameters on correct choice (**Supplementary Table 1**).

Stimuli. We programmed stimulus presentation in MATLAB using Cogent 2000 (http://www.vislab.ucl.ac.uk/cogent.php).

fMRI data acquisition. Data were acquired with a 3T scanner (Trio, Siemens, Erlangen, Germany) using a 12-channel phased array head coil. Functional images were taken with a gradient echo T2*-weighted echo-planar sequence (repetition time = 3.128 s, flip angle = 90°, echo time = 30 ms, 64 × 64 matrix). Whole brain coverage was achieved by taking 46 slices in ascending order (2-mm thickness, 1-mm gap, in-plane resolution of 3 × 3 mm), tilted in an oblique orientation at −30° to minimize signal dropout in ventrolateral and medial frontal cortex. We also acquired a B0-fieldmap (double-echo FLASH, TE1 = 10 ms, TE2 = 12.46 ms, 3 × 3 × 2 mm resolution) and high-resolution T1-weighted anatomical scan of the whole brain (MDEFT sequence, 1 × 1 × 1 mm resolution).

fMRI data analysis. We used SPM8 (rev. 4068; http://www.fil.ion.ucl.ac.uk/spm/) for image analysis and applied standard preprocessing procedures (EPI realignment and unwarping using field maps, segmenting T1 images into gray matter, white matter, and cerebrospinal fluid, and using segmentation parameters to warp T1 images to the SPM Montreal Neurological Institute (MNI) template,

and spatially smoothing of normalized functional data using an isotropic 8-mm full-width half-maximum Gaussian kernel).

We regressed fMRI time series onto a composite GLM containing individual regressors representing the presentation of the root, second choice, computer choice/transition, and outcome. We modeled choice trials in all four conditions separately and further divided choices in the PE condition into planning and trained chosen trials. Additional regressors captured button presses and motion correction regressors estimated from the realignment procedure. Regressors at the choice time and outcome were parametrically modulated by task relevant decision variables as described in the separate section below. We did not apply orthogonalization when we entered regressors and modulators into the design matrix, ensuring that the regressors of interest were not confounded by spurious correlations from signals pertaining to any of the other value signals[48]. We assessed statistical significance with a second-level random-effects analysis using a one-sample t test against zero on the effect sizes in individual subjects.

Value modulated parametric analysis. For the first scan session of the P trials, we hypothesized that the most salient value signals would be the value of the optimal path (target choice) and the values of the two alternative decision branches that subjects follow along their way through the maze. We therefore expected to find neural value representations of the optimal target action, the alternative tree branch at the root node and the alternative value at the second choice, and in response to the outcome in the reward rooms. In the example shown in **Figure 1b**, $V_{target} = 40$, $V_{root\_alternative} = 20$, and $V_{deep\_alternative} = 30$, reward outcome = 100 on rewarded and 0 on non-rewarded trials. To investigate the temporal dynamics of value representations during planning over the entire trial, we modulated regressors at three time points: during the root choice, during the second choice in layer 3 and during presentation of the outcome. The regressor during outcome presentation was additionally modulated by actual reward. Although the time of third choice and outcome were fixed (to avoid confounding effects of any potential prediction errors), the effects of expected value during choice (on a continuous scale) and response to the actual outcome (either 1 or 0) are still dissociable through the principle of competing variances in unorthogonalized regressors.

There was a significant positive effect for the target value and negative effect for the alternative value in this analysis, indicating a value difference between the two components in the overall signal. Separate testing of minuend (a) and subtrahend (b) is a more thorough test for a difference representation than a direct regression of the difference value $a - b$: if a alone had a very strong effect, then the latter test might still be significant despite the fact that the signal was actually better explained by a than by $a - b$. Notably, if there is a significant positive effect for a and a significant negative effect for b, then a contrast testing for the difference between $a - b$ is necessarily also significant.

For the first scan session of E trials, we modulated regressors during presentation of the choice screen and at outcome with the true reward probability of the rewarding action ($V_{trained}$). The regressor at the time of the outcome was also modulated by the experienced reward.

For the second scan session of the PE trials, we split trials according to subjects' choices and modeled separately plan chosen and train chosen trials. Regressors during choice were parametrically modulated with the target value in the planning branch (V_{target}), the alternative value at the second choice of the planning branch ($V_{deep.alternative}$) and the value of the colored trained branch ($V_{trained}$). The regressor at the time of the outcome was modulated by the experienced reward.

For the second scan session of the EE trials, the regressor during presentation of the choice screen was modulated by the value of the chosen (V_{chosen}) and unchosen ($V_{unchosen}$) branch, at the time of the outcome by the experienced reward.

PPI analysis. We performed a PPI analysis[49] to examine the functional coupling between vmPFC and caudate and putamen BOLD during mixed choices.

The PPI term was Y × P, with Y being the BOLD time courses in either the caudate and putamen ROI, and P being an indicator variable for the times during which mixed choices were made. We entered the seed region BOLD Y, and the PPI interaction term along with all regressors from our model based parametric analysis (containing P and all value regressors) into a new GLM. Notably, this GLM also contained the parametric value signals for both branches, so any effect on the PPI interaction would reveal increased coupling that could not be explained from the mutual correlation of seed and target region with the choice values. We computed this PPI both for a seed in caudate and in putamen, thereby separately identifying areas that showed a significant increase in coupling with both areas. The conjunction highlights common regions that played a role in mediating between both choice systems. Alternatively we tested for choice-dependent changes in coupling, that is, areas that would differentially increase coupling with caudate on plan chosen trials, but not on train chosen trials, and vice-versa for putamen. This analysis did not reveal significant results anywhere in the brain, even at a lenient threshold of $P < 0.005$, uncorrected.

We also tested the possibility that vmPFC correlated with the choice dependent difference time course[18] between activity in caudate and putamen by estimating a GLM on the PPI = Y × P, where Y = $t_{caudate} - t_{putamen}$, and P = 1 on plan chosen trials and −1 on trained chosen trials. However, when we added the parametric choice values V_{plan} and $V_{trained}$ as covariates of no interest to this PPI GLM (to rule out the possibility that effects on this interaction were solely due to mutual correlations of seed and target areas with the choice values), we did not find significant remaining interactions ($P < 0.005$ uncorrected).

Whole brain analysis. A whole brain parametric analysis confirmed a selective representation of planned target values during P trials in anterior caudate (**Supplementary Fig. 4a**) and cached values during E trials within posterior putamen (**Supplementary Fig. 4b**). Besides precentral gyrus (putatively motivational motor preparatory) we did not observe any other significant correlation ($P < 0.05$ FWE corrected) with value signals outside of our a priori brain regions in any trial types (**Supplementary Table 4**).

ROI analysis. We analyzed value signals (results in **Supplementary Table 5**) within a priori anatomically defined ROIs (**Supplementary Fig. 2**). For each region we regressed our design matrix on a representative time course, calculated as first eigenvariate[50]. This provides us with a very sensitive analysis as only a single regression is performed per region (no multiple comparisons required).

Anterior caudate (xyz mm): right: 9, 15, 3; left: −9, 15, 3; size: 6 mm radius, 66 voxels. Sphere centered in the anterior caudate nucleus. Dorsolateral posterior putamen: right: 33, −24, 0; left: −33, −24, 0; 4 mm radius, 20 voxels. Location based on a previous habit learning study[8]. vmPFC: 0, 32, −13; 8 mm sphere, 65 voxels. Radii chosen to fit anatomical boundaries.

44. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton University Press, 1944).
45. Dickinson, A. & Balleine, B.W. The role of learning in the operation of motivational systems. in *Stevens' Handbook of Experimental Psychology* (eds. Pashler, H. & Gallistel, R.) 497–533 (John Wiley & Sons, New York, 2002).
46. Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
47. Bellman, R. On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA* **38**, 716–719 (1952).
48. Andrade, A., Paradis, A.L., Rouquette, S. & Poline, J.B. Ambiguous results in functional neuroimaging data analysis due to covariate correlation. *Neuroimage* **10**, 483–486 (1999).
49. Friston, K.J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).
50. Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P. & Henson, R.N. A critique of functional localisers. *Neuroimage* **30**, 1077–1087 (2006).