Behavioral/Systems/Cognitive

Decoding the Formation of Reward Predictions across Learning

Thorsten Kahnt,[1,3] Jakob Heinzle,[1] Soyoung Q Park,[3,4] and John-Dylan Haynes[1,2,3]

[1]Bernstein Center for Computational Neuroscience and [2]Berlin Center for Advanced Neuroimaging, Charité–Universitätsmedizin Berlin, D-10115 Berlin, Germany, [3]Berlin School of Mind and Brain, Humboldt–Universität zu Berlin, D-10117 Berlin, Germany, and [4]Department of Education and Psychology, Freie Universität Berlin, D-14195 Berlin, Germany

The predicted reward of different behavioral options plays an important role in guiding decisions. Previous research has identified reward predictions in prefrontal and striatal brain regions. Moreover, it has been shown that the neural representation of a predicted reward is similar to the neural representation of the actual reward outcome. However, it has remained unknown how these representations emerge over the course of learning and how they relate to decision making. Here, we sought to investigate learning of predicted reward representations using functional magnetic resonance imaging and multivariate pattern classification. Using a pavlovian conditioning procedure, human subjects learned multiple novel cue–outcome associations in each scanning run. We demonstrate that across learning activity patterns in the orbitofrontal cortex, the dorsolateral prefrontal cortex (DLPFC), and the dorsal striatum, coding the value of predicted rewards become similar to the patterns coding the value of actual reward outcomes. Furthermore, we provide evidence that predicted reward representations in the striatum precede those in prefrontal regions and that representations in the DLPFC are linked to subsequent value-based choices. Our results show that different brain regions represent outcome predictions by eliciting the neural representation of the actual outcome. Furthermore, they suggest that reward predictions in the DLPFC are directly related to value-based choices.

Introduction

A fundamental prerequisite for adaptive and goal-directed action is a reliable representation of the predicted outcomes of different behavioral options (Montague et al., 2006). Sensory cues signal potential rewards, but the associations between cues and rewards must be learned so that they can be used to guide choices. The acquisition and representation of outcome predictions involves the striatum (Knutson et al., 2001; O'Doherty et al., 2004; Samejima et al., 2005; Kahnt et al., 2009; Tobler et al., 2009) and prefrontal areas including the orbitofrontal cortex (OFC) (Tremblay and Schultz, 1999; Schoenbaum et al., 1998; O'Doherty et al., 2002; Gottfried et al., 2003; Plassmann et al., 2007) and the dorsolateral prefrontal cortex (DLPFC) (Watanabe, 1996; Barraclough et al., 2004; Roesch and Olson, 2004). Reinforcement learning (RL) provides a theoretical framework for how sensory cues acquire reward value (Sutton and Barto, 1998). Specifically, during repeated pairings between sensory cues and reward outcomes, the value of the outcome is gradually taken over by the sensory cue by minimizing the difference between the expected and the actual outcome (Schultz et al., 1997).

Neuronal recording studies have shown that in the primate prefrontal cortex, reward value is coded by different neuronal populations that increase and decrease activity with increasing reward value, respectively (Kennerley et al., 2009; Morrison and Salzman, 2009; Kobayashi et al., 2010). In line with this distributed coding scheme, patterns of functional magnetic resonance imaging (fMRI) activity in the OFC have been shown to carry distributed information about the value of reward-predicting cues (Kahnt et al., 2010, 2011b). Moreover, the idiosyncratic fMRI patterns coding predicted rewards are similar to the fMRI patterns coding the actual reward outcome (Kahnt et al., 2010). This suggests that the distributed neural signature coding the value of the outcome is taken over by the sensory cue. Thus, the similarity between the fMRI patterns coding reward value during the sensory cue and the outcome (hereafter, "pattern similarity") can serve as an fMRI measure for the acquisition of reward predictions.

Here, we study cue–outcome pattern similarity to investigate the acquisition of outcome predictions in the human brain. In each of seven scanning runs, subjects learned six novel cue–outcome associations using a pavlovian conditioning task. After each run, these predicted reward representations were probed using a free-choice task. We hypothesized that fMRI patterns coding predicted and actual outcomes are unrelated before learning, when no reward expectancies have been established yet. However, once outcome predictions have been acquired, similar fMRI patterns should code the reward value during the cue and the actual out-

Received July 5, 2011; revised July 22, 2011; accepted Aug. 17, 2011.

Author contributions: T.K., J.H., S.Q.P., and J.-D.H. designed research; T.K. and J.H. performed research; T.K. analyzed data; T.K., J.H., S.Q.P., and J.-D.H. wrote the paper.

This work was supported by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung Grant 01GQ0411), the Excellence Initiative of the German Federal Ministry of Education [Deutsche Forschungsgemeinschaft (DFG) Grant GSC86/1-2009], and the Sonderforschungsbereich (SFB 779 A3) of the German Research Foundation (DFG).

Correspondence should be addressed to Thorsten Kahnt, Bernstein Center for Computational Neuroscience, Philippstrasse 13, House 6, D-10115 Berlin, Germany. E-mail: thorsten.kahnt@bccn-berlin.de.

DOI:10.1523/JNEUROSCI.3412-11.2011

 0270-6474/11/3114624-07$15.00/0

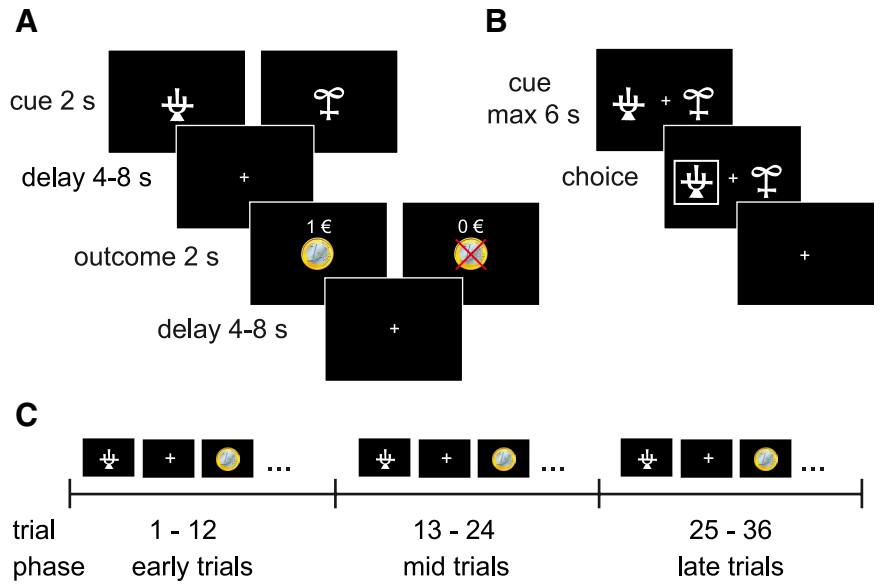

Figure 1. Experimental design. ***A***, In each of the seven scanning runs, six novel cues were associated with reward and no-reward outcomes (3 reward and 3 no reward). Subjects had to respond to the presentation of the cue as quickly as possible using a button press. ***B***, After each scanning run, subjects performed a value-based choice task. ***C***, Trials were sorted into an early, mid, and late learning phase for all analyses.

come (Kahnt et al., 2010). First, we made no assumptions about the formation of reward predictions and investigated pattern similarity during early, mid, and late learning trials separately. We find that cue–outcome pattern similarity emerges across learning in the striatum and the prefrontal cortex. Second, we fit an RL model to the behavioral data and show that the predicted reward representations in the dorsal striatum follow an RL process and precede those in prefrontal regions.

Materials and Methods

Subjects. Twenty-three subjects (mean age ± SD, 25.43 ± 3.23 years; 10 female) participated in the experiment. All subjects had normal or corrected-to-normal vision and were free of neurological and psychiatric history. The experimental procedure was approved by the local ethics review board of the Charité–Universitätsmedizin Berlin, and subjects gave informed written consent to participate in the study.

Pavlovian conditioning task. In each of the seven scanning runs, subjects performed a pavlovian conditioning task where they learned the associations between six novel cues and monetary outcomes or no outcomes (three rewards, three no rewards). In each trial, one cue was shown for 2 s followed by a variable delay of 4–8 s (mean, 6 s), after which the outcome was presented for 2 s (Fig. 1*A*). To obtain a behavioral measure of conditioning during scanning, subjects were instructed to respond to the presentation of the cue by pressing a button with the right index finger. If the button was not pressed within 2 s, no outcome was presented and the trial was aborted. Trials were separated by a variable interval of 4–8 s. Each cue–outcome pairing was repeated six times per scanning run (36 trials per run in total), and the trial order was pseudo-randomized such that each cue was presented twice during each one-third of the scanning run. In each scanning run, six new cues were introduced, and each cue predicted the outcome with 100% contingency. This combination of deterministic reinforcement and randomized presentation of multiple cues was used to maximize the number of trials (cues and outcomes) that can be used for data analysis while keeping learning speed relatively slow.

Postscanning value-based choice task. To test whether subjects acquired cue–outcome associations by the end of each scanning run, subjects performed a value-based choice task within the scanner (without fMRI data acquisition) after each run (Fig. 1*B*). In each trial of the choice task, a reward and a no-reward cue were presented to the left and right side of fixation, respectively (randomized). All combinations of reward and no-reward cues were presented once, resulting in nine trials per session. Subjects were asked to pick their preferred cue by pressing the left or the right button on a response box using the index or middle finger of their right hand, respectively. Importantly, to exclude further (instrumental) learning, no feedback was provided during the choice task. The monetary reward of the chosen cues was paid to the subjects after the experiment.

Temporal partition of trials into learning phases. The trials of each run were divided into three learning phases (Fig. 1*C*): early (trials 1–12; first and second presentation of each cue), mid (trials 13–24; third and fourth presentation of each cue), and late (trials 25–36; fifth and sixth presentation of each cue) learning. All analyses of the behavioral and fMRI data were based on this temporal partition. The partition was used to have a sufficient number of data points for the decoding analysis within each learning phase. For early learning trials (the first and second presentation of each cue), we can assume that no reward predictions have been formed yet, whereas the results of the postscanning preference task show that associations were established at least after the last trial (Fig. 2*A*).

Reinforcement learning model. We used a standard RL model (Sutton and Barto, 1998) to generate predictions about learning. Specifically, a Rescorla-Wagner model (Rescorla and Wagner, 1972) learned the expected value of the cues via a simple delta rule. In each trial, the expected value EV was updated by a prediction error δ, the difference between the actually received R and the expected reward: $EV_{t+1} = EV_t + \alpha \times \delta_t$, where $\delta_t = R_t - EV_t$. These models have previously been shown to account for learning-related changes in deterministic pavlovian conditioning tasks (Seymour et al., 2004; O'Doherty et al., 2006). Because reaction time (RT) latencies to conditioned stimuli have been shown to correlate with the expected values derived from RL models (Critchley et al., 2002; Gottfried et al., 2003; Seymour et al., 2004; O'Doherty et al., 2006), the free parameter in the RL model (the learning rate α) was estimated for all subjects together using the RT data as an on-line measure of conditioning (Bray and O'Doherty, 2007). Specifically, we computed model-derived expected values for a range of learning rates (0.01–0.99, in steps of 0.01), and the expected values from each model were then regressed (simultaneously for all runs and subjects) against the log-transformed trial-by-trial RT latencies. To control for unspecific RT differences between runs as well as unspecific RT drifts within runs, the regression model included run-wise constants and within-run linear trends besides the model-derived expected values. We finally used the learning rate that produced expected values that explained the most variance in the RT data (R^2). This estimation procedure yielded a learning rate of $\alpha = 0.14$. Please note that this learning rate is comparable to that used in previous pavlovian conditioning tasks with deterministic reinforcement schedules involving six predictive cues simultaneously (O'Doherty et al., 2006). The expected values derived from our model indicate that reward predictions are low in early learning trials, whereas learning is almost completed in late learning trials (Fig. 2*C*).

fMRI data acquisition and preprocessing. Functional imaging was conducted on a 3 tesla Trio (Siemens) scanner equipped with a 12-channel head coil. In each of the seven scanning runs, 285 T2*-weighted gradient-echo echo-planar images containing 33 slices (3 mm thick) separated by a gap of 0.75 mm were acquired. Imaging parameters, resulting in a voxel size of 3 × 3 × 3.75 mm, were as follows: repetition time (TR), 2000 ms; echo time (TE), 30 ms; flip angle, 90°; matrix size, 64 × 64; field of view (FOV), 192 mm. A T1-weighted structural data set was collected for the

purpose of anatomical localization. The parameters were as follows: TR, 1900 ms; TE, 2.52 ms; matrix size, 256 × 256; FOV, 256 mm; 192 slices (1 mm thick); flip angle, 9°.

Preprocessing, parameter estimation, and group statistics of the functional data were performed using SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). For preprocessing, images were slice time corrected, realigned, and spatially normalized to the MNI template. No smoothing was applied at this point of the analysis.

Searchlight decoding and pattern similarity. We have previously shown that the reward value during anticipation and receipt of reward is coded by similar fMRI patterns in the human OFC (Kahnt et al., 2010). Here, we used the same searchlight decoding approach to track the similarity of the fMRI patterns during learning. Specifically, as input to the decoding analysis, for each scanning run we set up a GLM with four regressors [(1) reward cues, (2) no-reward cues, (3) reward outcomes, and (4) no-reward outcomes] for each of the three learning phases. This resulted in 12 regressors ([2 cues + 2 outcomes] × 3 learning phases). Note that each regressor modeled six trials (3 cues × 2 presentations per learning phase). All regressors were convolved with a canonical HRF and simultaneously regressed against the BOLD signal in each voxel. The resulting parameter estimates represent the response amplitudes for each voxel and each condition. Importantly, estimating parameters simultaneously for all conditions in the context of a GLM assures that cue- and outcome-related BOLD signals are separated.

The parameter estimates were then used to search for brain regions in which similar fMRI patterns code reward value during the sensory cue and the actual outcome in each learning phase separately. For this, we used a searchlight decoding approach (Kriegeskorte et al., 2006; Haynes et al., 2007) that examines the information in local fMRI patterns surrounding each voxel (radius, 10 mm; Fig. 3*A*). Importantly, because we were only interested in information encoded in the distributed fMRI patterns, the mean of each pattern was subtracted. This assures that the average signal of each pattern is zero and thus cannot contain any information at all. We then used six of the seven scanning runs to train a linear support vector classifier (SVC) to separate fMRI patterns corresponding to reward versus no-reward outcomes. For this, we used the LIBSVM (Library for Support Vector Machines) implementation (www.csie.ntu.edu.tw/~cjlin/libsvm/). We tested the SVC by classifying fMRI patterns corresponding to reward versus no-reward cues in an independent test data set (run 7) (Fig. 3*B*). We also performed the classification in the reverse direction, by training on fMRI patterns during the cue and testing on fMRI patterns during the outcome (reported results are averaged across both directions). Importantly, because (1) we used different sensory cues in each scanning run and (2) we trained on data during the outcome and tested on data during the cue, significant above-chance decoding is only possible if (1) fMRI patterns code the predicted reward independent of the sensory features of the cue and (2) similar patterns code the value of predicted and actual outcomes. This classification was repeated seven times using a leave-one-run-out cross-validation. Note that cross-validation using independent training and test data sets excludes biased results because of overfitting or "double dipping." The procedure was done for each searchlight (i.e., each center voxel) and each learning phase resulting in three voxel-wise maps of decoding accuracy, one for each learning phase. If the similarity between reward coding response patterns during cue and outcome emerges during learning, significant classification should be possible in late but not in early learning trials.

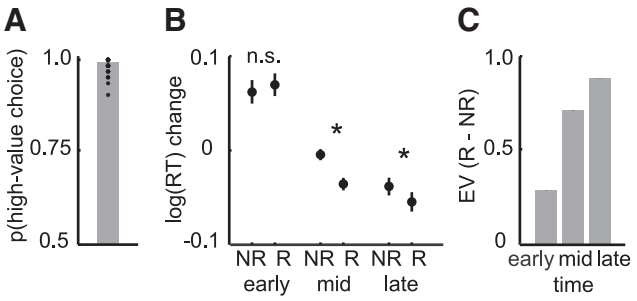

Figure 2. Behavioral results. ***A***, Bar plot shows average probability of a high-value choice in the postscanning choice task across subjects; asterisks indicate individual data points. ***B***, Log-transformed reaction time data as a function of reward value and learning phase (early, mid, late). Error bars indicate SEM for $n = 23$, and asterisks indicate significant RT (log-transformed) differences at $p < 0.05$ (one-tailed). ***C***, The EVs of reward–no-reward cues generated by the RL model are plotted as a function of learning phase. R, Reward; NR, no reward.

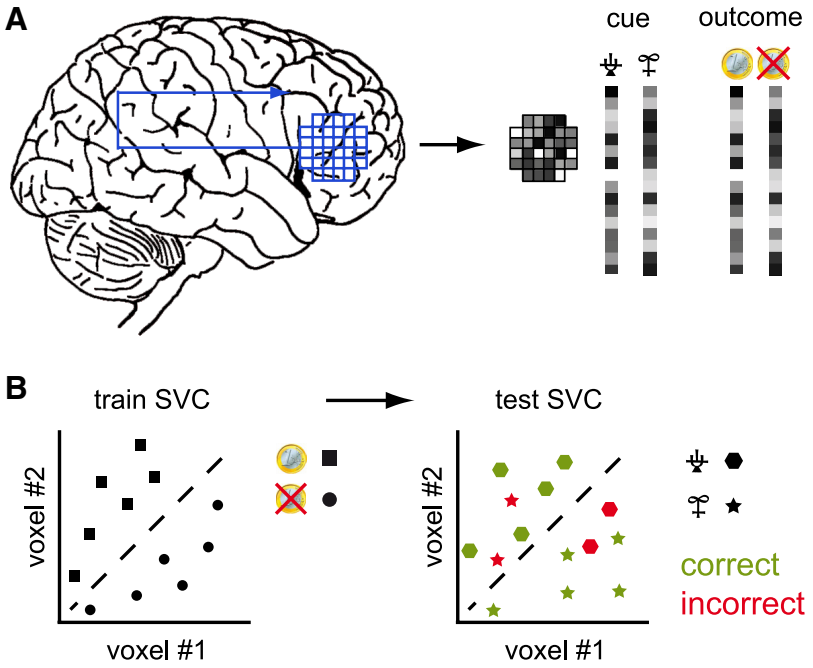

Figure 3. Multivariate searchlight decoding. ***A***, The local fMRI patterns in each searchlight were extracted for reward and no-reward cues as well as for reward and no-reward outcomes for each run and each learning phase. ***B***, Within each learning phase, we trained a linear SVC on a training data set to separate reward versus no-reward outcomes. The performance of the SVC was then assessed by testing how well the SVC classified reward versus no-reward cues in an independent test data set. Note that these cross-classifications are only successful if similar patterns code reward value during cue and outcome.

As a measure of decoding accuracy, we used the base-rate-independent area under the receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982; Swets, 1986). However, using percentage correct provides very similar results. For each searchlight, this ROC was created using the actual labels of the test data set and the predictions made by the SVC model ("decision values") when applied to the test data set. ROC curves relate the number of correct positive classifications to the number of false positive classifications across the entire range of possible decision boundaries. The area under the ROC curve (AUC; range, 0–1; chance level, 0.50) quantifies the accuracy of a classifier and has a straightforward interpretation: If one instance of each category (e.g., a reward and a no-reward cue) is chosen randomly from an independent test data set, the AUC represents the probability that the reward cue will have a numerically higher decision value than the no-reward cue. Because we trained and tested the SVC model on data during cue and outcome, respectively, AUC values, in our case, reflect pattern similarity. Specifically, AUC values are related to the similarity (i.e., the correlation)

between the fMRI patterns (coding reward versus no reward) during the sensory cue and the fMRI patterns (coding reward versus no reward) during the outcome.

Group statistics were performed on a voxel-by-voxel basis. For this purpose, the individual AUC maps were smoothed with a 6 mm FWHM Gaussian kernel. Regions in which similar response patterns code reward value during prediction and receipt of reward were identified using voxel-wise *t* tests on AUC maps from late learning trials. We identified significant clusters using a statistical threshold of $p < 0.0001$ (uncorrected) together with a cluster-extend threshold of $k = 5$ continuous voxels. Furthermore, only voxels that also survived a whole-brain correction for multiple comparisons (false discovery rate, FDR) at $p < 0.01$ are reported in Results.

Results

Behavioral measures of predicted reward representations

Behavioral performance on the postscanning choice task revealed that subjects had acquired reward predictions during conditioning and were able to use these representations to guide their choices (mean, 98% high-value choices; Fig. 2*A*). Learning of reward predictions was also evident in the RT data during conditioning. A (3 × 2) time-by-reward ANOVA with repeated measures on log-transformed RT revealed significant main effects for reward ($F_{(1,22)} = 4.38$; $p < 0.05$) and time ($F_{(2,44)} = 32.51$; $p < 0.05$) as well as a significant time-by-reward interaction effect ($F_{(2,44)} = 4.23$; $p < 0.05$). *Post hoc t* tests showed significantly shorter log-transformed RT latencies to reward than no-reward cues in mid and late learning ($p < 0.05$, one-tailed), but not in early learning trials ($p = 0.56$; Fig. 2*B*). A standard RL model (Sutton and Barto, 1998; Kahnt et al., 2009; Park et al., 2010) was fit to the RT data to predict learning-related changes in expected value. The RL model suggests that reward predictions (difference between the expected values of reward and no-reward cues) are well established in late learning trials (Fig. 2*C*). Together, the behavioral results clearly show that subjects learned the associations between the cues and the outcomes and were subsequently able to use this information to guide their choices.

fMRI measures of predicted reward representations

In the following, we focus on the fMRI data to search for neural changes associated with the acquisition of reward predictions. First, we identified brain regions that reveal significant cue–outcome pattern similarity (as a neural measure for learning; see above) after reward predictions have been acquired (i.e., in late learning trials). To identify such regions, we used a searchlight decoding approach that examined the information in locally distributed fMRI patterns (Fig. 3*A*). Specifically, for the fMRI patterns in each searchlight (radius, 10 mm), we tested how well the reward value (reward vs no reward) of a sensory cue was classified (in an independent test data set), if the classifier was trained on fMRI patterns from reward versus no-reward outcomes (Fig. 3*B*). Pattern similarity thus reflects the information about reward versus no reward that is common to both prediction and receipt of reward. In late learning trials, we found significant ($p < 0.0001$; $k = 5$) AUC values in the medial and central OFC (medial BA 11, MNI [*x*, *y*, *z*] = [−3, 57, −9], $t = 4.32$; central BA 11, [−18, 57, −9], $t = 4.59$; Fig. 4*A*), in the bilateral DLPFC (left BA 9, [−18, 45, 42], $t = 5.32$; right BA 9, [27, 48, 39], $t = 4.97$; Fig. 4*B*), in the left dorsal striatum ([−21, −12, 15], $t = 4.41$; Fig. 4*C*), in the superior temporal gyrus (BA 22, [−54, −48, 12], $t = 4.03$), and in the posterior cingulate cortex (BA 23, [−6, −27, 24], $t = 4.31$). Thus, these brain regions code reward value (reward vs no reward) during prediction and receipt of reward by similar activity patterns after the cue–outcome associations have been acquired.

In a second step, we examined pattern similarity in these regions during the other time points of learning. From (1) the medial and central OFC, (2) the bilateral DLPFC, and (3) the dorsal striatum, exhibiting significant pattern similarity during late learning trials (see above), we extracted the averaged (across voxels) AUC values from all three time points (Fig. 4, right). We then applied these AUC values to a (3 × 3) time-by-region ANOVA with repeated measures. We found a significant main effect of time ($F_{(2,44)} = 13.23$; $p < 0.001$) that was driven by significant ($p < 0.01$) differences between AUC values during early and late as well as mid and late learning but not between early and mid learning ($p > 0.99$). We observed no significant main effect of region ($F_{(2,44)} = 1.80$; $p = 0.18$) but a significant time-by-region interaction ($F_{(4,88)} = 2.58$; $p < 0.05$). *Post hoc t* tests revealed that this interaction was driven by significantly higher AUC values in the dorsal striatum compared with the DLPFC ($t = 2.79$; $p < 0.05$) and the OFC ($t = 3.05$; $p < 0.05$) during mid learning trials. These results suggest that reward predictions in the dorsal striatum arise at an earlier time point than in the prefrontal cortex.

Moreover, we searched for brain regions in which reward predictions change at the same pace as the expected values of the RL model. We reasoned that the higher the model-derived expected values, the better the reward predictions should be represented specifically in brain regions where learning proceeds in an RL-like fashion. To test this idea, we correlated the whole-brain images of AUC values (early, mid, and late learning) with the expected values of the RL model (for reward vs no-reward cues, as shown in Fig. 2*C*). Only in the dorsal striatum ([−21, −15, 12], $t = 4.98$) did this analysis reveal significant ($p < 0.0001$; $k = 5$) correlations between changes in expected value of the RL model and changes in neural representations of predicted reward (pattern similarity). Importantly, these voxels overlapped substantially with the striatal voxels observed above. We did not find any significant correlations with the OFC and the DLPFC. This result suggests that reward predictions only in the dorsal striatum change at a pace that is in line with classical RL processes. In fact, this result could have been expected based on the temporal profile of the AUC values in the striatum and the expected values of the RL model that was fitted using RT data.

We further explored whether outcome predictions in striatal and prefrontal regions are related to subjects' ability to use these outcome predictions to guide value-based decisions. Specifically, we tested whether cue–outcome pattern similarity during late learning trials was correlated with the percentage of high-value choices in the postscanning task. Because the percentages of high-value choices are not normally distributed (Kolmogorov–Smirnov test, $p < 0.05$), the nonparametric Spearman's rank correlation coefficient was used to explore this relationship. This correlation between pattern similarity in late learning trials and the percentage of high-value choices in the postscanning task was significant in the DLPFC ($r = 0.51$; $p < 0.05$) but not in the OFC ($p = 0.85$) or the dorsal striatum ($p = 0.20$). In addition, we performed a permutation test (2000 permutations) to obtain the exact two-tailed *p* value for our empirically observed relationship. This procedure yielded a value of $p = 0.014$. These results suggest that only outcome predictions in the DLPFC are directly related to value-based action selection, i.e., the DLPFC activity seems to be important for the ability to use reward predictions to guide choices.

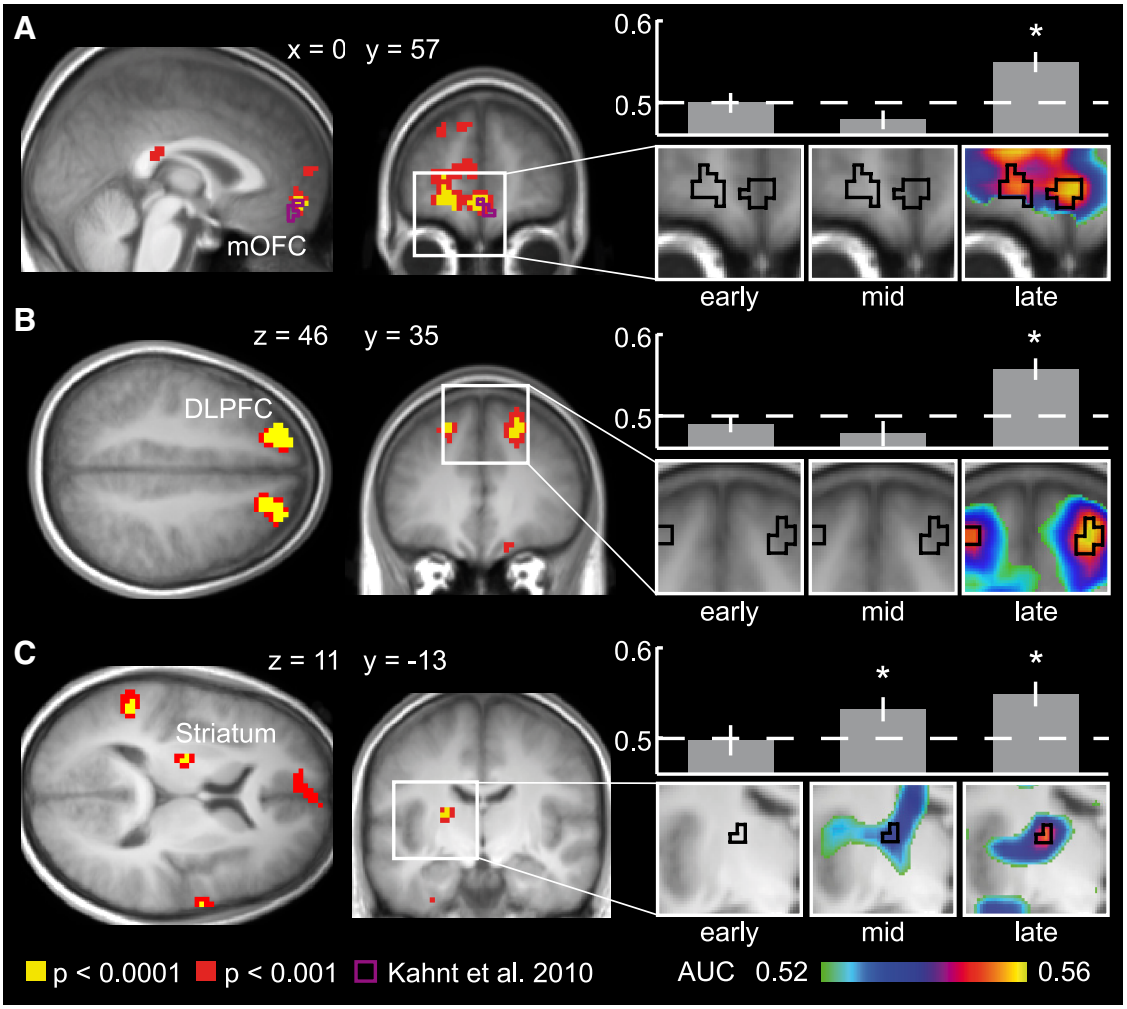

Figure 4. Similar reward coding response patterns for cue and outcome during late learning. Left, In the OFC (***A***), DLPFC (***B***), and dorsal striatum (***C***), similar fMRI patterns encode reward during prediction and receipt of reward. The purple outline in ***A*** represents a significant cluster for a similar analysis in a previous study (Kahnt et al., 2010). *T*-maps are thresholded at $p < 0.0001$ (yellow) and $p < 0.001$ (red) and are overlaid on a normalized anatomical image averaged across subjects. Right, Learning-related changes in pattern similarity in the OFC (***A***), DLPFC (***B***), and striatum (***C***). Coronal sections depict voxel-wise pattern similarity (AUC) during early, mid, and late learning. Black outlines indicate significant regions ($p < 0.0001$) corresponding to yellow areas on the left. Bar plots depict pattern similarity averaged across voxels in the black outline below. $^*p < 0.05$. White dashed lines indicate chance level (AUC of 0.5).

Discussion

In the current study, we have shown that spatially distributed fMRI patterns in the OFC, DLPFC, and striatum represent the reward value of predicted outcomes. Furthermore, we have shown how these representations are acquired across learning and that striatal representations precede those in prefrontal cortex. Finally, our results suggest that reward representations in the DLPFC specifically subserve value-based action selection.

The representation of outcome predictions in striatal and prefrontal regions is consistent with a large body of evidence from animal electrophysiology and human neuroimaging studies (Watanabe, 1996; Schoenbaum et al., 1998; Tremblay and Schultz, 1999; Breiter et al., 2001; Knutson et al., 2001; O'Doherty et al., 2002, 2004; Gottfried et al., 2003; Barraclough et al., 2004; Roesch and Olson, 2004; Samejima et al., 2005; Plassmann et al., 2007; Kahnt et al., 2009; Tobler et al., 2009; Park et al., 2011). However, the precise coding scheme as well as how these representations emerge during learning has remained primarily elusive. We have previously shown that similar fMRI patterns code the reward value of predicted and actual outcomes after outcome predictions have been established (Kahnt et al., 2010). Here, we extend this finding in a critical way by showing that this pattern similarity is present after, but not before, learning, suggesting that the distributed representations of predicted and actual outcomes become similar during learning. Importantly, this cue–outcome pattern similarity was extremely robust and generalized across the different sensory cues (predicting the same reward outcome) that were used in the seven independent scanning runs. Thus, the brain might represent reward predictions by eliciting the neural representation of the actual reward outcome independent of the cue that caused this prediction.

We found differences in learning between striatal and prefrontal brain regions. Specifically, representations of reward predictions in the striatum occurred earlier than in the DLPFC and the OFC, and only changes in the dorsal striatum were correlated with the expected values of an RL model. This suggests the presence of multiple reward signals in the brain that are not necessarily all related to RL processes. Similar dissociations of learning in striatal and prefrontal areas have been revealed in animal recording studies, where striatal processes have been shown to precede prefrontal representations (Pasupathy and Miller, 2005). In line with this, altered functional connectivity between the striatum

and the DLPFC has been shown to account for learning and decision-making impairments in addicted patients (Park et al., 2010). Furthermore, our results suggest that representations in the DLPFC are directly related to subsequent value-based decision making. This result is well in line with numerous findings linking DLPFC activity to information necessary for optimal decision making such as sensory evidence (Romo and Salinas, 2003; Heekeren et al., 2004; Philiastides et al., 2011), categories (Seger and Miller, 2010), rules (Wallis et al., 2001), and previous choices and rewards (Watanabe, 1996; Barraclough et al., 2004; Lee and Seo, 2007).

The fMRI patterns we observe to code specific rewards are likely to result from randomly distributed populations of neurons with different tuning properties for reward. Specifically, positive and negative value coding neurons have been shown to coexist in the primate and rat OFC (Schoenbaum et al., 1998; Tremblay and Schultz, 1999; Padoa-Schioppa and Assad, 2006; Kennerley et al., 2009; Morrison and Salzman, 2009; Kobayashi et al., 2010). Positive value coding neurons increase their firing rate with increasing value, whereas negative value coding neurons decrease their firing rate. Both populations are equally prevalent in the primate PFC (Kennerley et al., 2009) and could lead to slightly biased responses of individual MR voxels (Kahnt et al., 2010). Biased sampling of individual voxels, as well as similar mechanisms (Kriegeskorte et al., 2010; Swisher et al., 2010), has been suggested to account for orientation information in fMRI patterns obtained from the early visual cortex (Haynes and Rees, 2005; Kamitani and Tong, 2005) but also for higher cognitive and decision variables throughout the brain (Haynes and Rees, 2006; Norman et al., 2006; Kahnt et al., 2011a,b; Soon et al., 2008).

In the current study, we have shown that value-coding fMRI patterns during the reward-predicting cue and the rewarding outcome become similar during learning. Representations in the striatum appeared earlier than in the DLPFC and OFC, pointing toward different time courses of learning in these regions and the presence of different value signals. Furthermore, only reward representations in the DLPFC were related to the performance in subsequent value-based decisions, suggesting a critical role of DLPFC representations in reward-based action selection. These results shed light onto the basic functional architecture of learning and decision making and suggest that multiple reward representations in the human brain promote specific and dissociable functions.

References

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404–410.

Bray S, O'Doherty J (2007) Neural coding of reward-prediction error signals during classical conditioning with attractive faces. J Neurophysiol 97:3036–3045.

Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron 30:619–639.

Critchley HD, Mathias CJ, Dolan RJ (2002) Fear conditioning in humans: the influence of awareness and autonomic arousal on functional neuroanatomy. Neuron 33:653–663.

Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science 301:1104–1107.

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36.

Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–691.

Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534.

Haynes JD, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE (2007) Reading hidden intentions in the human brain. Curr Biol 17:323–328.

Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG (2004) A general mechanism for perceptual decision-making in the human brain. Nature 431:859–862.

Kahnt T, Park SQ, Cohen MX, Beck A, Heinz A, Wrase J (2009) Dorsal striatal-midbrain connectivity in humans predicts how reinforcements are used to guide decisions. J Cogn Neurosci 21:1332–1345.

Kahnt T, Heinzle J, Park SQ, Haynes JD (2010) The neural code of reward anticipation in human orbitofrontal cortex. Proc Natl Acad Sci U S A 107:6010–6015.

Kahnt T, Grueschow M, Speck O, Haynes JD (2011a) Perceptual learning and decision-making in human medial frontal cortex. Neuron 70:549–559.

Kahnt T, Heinzle J, Park SQ, Haynes JD (2011b) Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. Neuroimage 56:709–715.

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685.

Kennerley SW, Dahmubed AF, Lara AH, Wallis JD (2009) Neurons in the frontal lobe encode the value of multiple decision variables. J Cogn Neurosci 21:1162–1178.

Knutson B, Adams CM, Fong GW, Hommer D (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J Neurosci 21:RC159(1–5).

Kobayashi S, Pinto de CO, Schultz W (2010) Adaptation of reward sensitivity in orbitofrontal neurons. J Neurosci 30:534–544.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.

Kriegeskorte N, Cusack R, Bandettini P (2010) How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage 49:1965–1976.

Lee D, Seo H (2007) Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. Ann N Y Acad Sci 1104:108–122.

Montague PR, King-Casas B, Cohen JD (2006) Imaging valuation models in human choice. Annu Rev Neurosci 29:417–448.

Morrison SE, Salzman CD (2009) The convergence of information about rewarding and aversive stimuli in single neurons. J Neurosci 29:11471–11483.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430.

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826.

O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.

O'Doherty JP, Buchanan TW, Seymour B, Dolan RJ (2006) Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. Neuron 49:157–166.

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226.

Park SQ, Kahnt T, Beck A, Cohen MX, Dolan RJ, Wrase J, Heinz A (2010) Prefrontal cortex fails to learn from reward prediction errors in alcohol dependence. J Neurosci 30:7749–7753.

Park SQ, Kahnt T, Rieskamp J, Heekeren HR (2011) Neurobiology of value integration: when value impacts valuation. J Neurosci 31:9307–9314.

Pasupathy A, Miller EK (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433:873–876.

Philiastides MG, Auksztulewicz R, Heekeren HR, Blankenburg F (2011) Causal role of dorsolateral prefrontal cortex in human perceptual decision making. Curr Biol 21:980–983.

Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27:9984–9988.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton Century Crofts.

Roesch MR, Olson CR (2004) Neuronal activity related to reward value and motivation in primate frontal cortex. Science 304:307–310.

Romo R, Salinas E (2003) Flutter discrimination: neural codes, perception, memory and decision making. Nat Rev Neurosci 4:203–218.
Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340.
Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat Neurosci 1:155–159.
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.
Seger CA, Miller EK (2010) Category learning in the brain. Annu Rev Neurosci 33:203–219.
Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667.
Soon CS, Brass M, Heinze HJ, Haynes JD (2008) Unconscious determinants of free decisions in the human brain. Nat Neurosci 11:543–545.
Sutton R, Barto A (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.
Swets JA (1986) Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 99:100–117.
Swisher JD, Gatenby JC, Gore JC, Wolfe BA, Moon CH, Kim SG, Tong F (2010) Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. J Neurosci 30:325–330.
Tobler PN, Christopoulos GI, O'Doherty JP, Dolan RJ, Schultz W (2009) Risk-dependent reward value signal in human prefrontal cortex. Proc Natl Acad Sci U S A 106:7185–7190.
Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. Nature 398:704–708.
Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. Nature 411:953–956.
Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629–632.