Published in final edited form as:
J Neurophysiol. 2016 June 1; 115(6): 3195–3203. doi:10.1152/jn.00046.2016.

The involvement of model-based but not model-free learning signals during observational reward learning in the absence of choice

Simon Dunne[1,2], **Arun D'Souza**[1,3], and **John P. O'Doherty**[1,2,4]

[1]Trinity College Institute of Neuroscience, Trinity College Dublin, Ireland [2]Computation and Neural Systems Program, California Institute of Technology [3]Department of Psychology, University of Freiburg, Germany [4]Division of Humanities and Social Sciences, California Institute of Technology

Abstract

A major open question is whether computational strategies thought to be used during experiential learning, specifically model-based and model-free reinforcement-learning, also support observational learning. Furthermore, the question of how observational learning occurs when observers must learn about the value of options from observing outcomes in the absence of choice, has not been addressed. In the present study we used a multi-armed bandit task that encouraged human participants to employ both experiential and observational learning while they underwent functional magnetic resonance imaging (fMRI). We found evidence for the presence of model-based learning signals during both observational and experiential learning in the intraparietal sulcus. However, unlike in experiential learning, model-free learning signals in the ventral striatum were not detectable during this form of observational learning. These results provide insight into the flexibilty of the model-based learning system, implicating this system in learning during observation as well as from direct experience, and further suggest that the model-free reinforcement-learning system may be less flexible with regard to its involvement in observational learning.

Keywords

Observational learning; reward prediction error; reinforcement learning; ventral striatum; fMRI

Introduction

The ability to modify a behavior after experiencing its consequences is vital to the survival of any animal (Rescorla and Wagner 1972; Sutton and Barto 1998). However, learning solely by experiential trial and error can be time-consuming and dangerous and many

Address for reprint requests and other correspondence: S. Dunne, California Institute of Technology, MC 228-77, 1200 East California Boulevard, Pasadena, CA 91125 (sdunne@caltech.edu).

Author contributions

J.P.O. conceived and designed research; S.D. and A.D. collected data; S.D. analyzed data, prepared figures and drafted manuscript; S.D. and J.P.O. edited and revised manuscript; S.D, A.D. and J.P.O. approved final version of manuscript.

Europe PMC Funders Author Manuscripts

species have developed the ability to learn from other sources of information (Tomasello et al. 1987; Galef and Laland 2005; Whiten et al. 2005; Grüter and Farina 2009). Humans are particularly adept at such non-experiential learning, in particular that achieved by observing others take actions and receive their consequences, or *observational learning* (Berger 1962; Bandura 1977).

While a handful of studies have been conducted to examine the neural computations underlying observational learning (Burke et al. 2010; Cooper et al. 2012; Suzuki et al. 2012), this form of learning remains relatively unexplored compared to its experiential counterpart. In particular, studies investigating observational learning have focused on the situation where observers learn from the consequences of actions freely chosen by an observed person (Burke et al. 2010; Cooper et al. 2012; Suzuki et al. 2012). In that situation, there are multiple learning strategies that an agent could use to guide their choice. One strategy, akin to belief learning in economic games (Camerer and Ho 1999; Hampton et al. 2008), is to learn predictions for which action the observee is likely to choose based on the observee's past choices, and to use that information to simply mimic the observee's behavior. Another strategy is to learn the value of the outcome associated with each option as the observer experiences them, and to use that value information to guide one's own choices. While there is some evidence to suggest that individuals are capable of learning predictions about others' actions in either observational learning or strategic interactions (Hampton et al. 2008; Burke et al. 2010; Suzuki et al. 2012), much less is known about the brain's ability to learn the value of different decision options through observation in the absence of free-choice in the actions being observed.

The goal of the present study was to address whether the human brain can learn about the value of stimuli through observation, in the absence of choice behavior in the observee that could be used to drive action-based learning. In the experiential domain, it has been proposed that learning about the value of decision options can occur via two distinct computational mechanisms: a model-free reinforcement learning mechanism, in which options are evaluated according to a "cached" history of their reinforcement and a model-based learning mechanism, which acquires a model of the decision problem that it uses to compute the values of different decision options (Daw et al. 2005; Balleine et al. 2008). Although computationally simple, a model-free mechanism has limited flexibility. For example, a model-free agent will persist in choosing a option, even if its contingent outcome suddenly becomes no longer valuable. This is because, lacking a model of the environment, it has no representation of the current value of the outcome and relies solely on how rewarding the option has been in the past. In contrast, a model-based learning mechanism can adapt immediately to such a change, but it is more computationally complex for it to evaluate different available options because because, rather than relying a cached value for an action, it needs to consider all possible future implications of choosing that option.

Recent evidence suggests that both of these mechanisms may be present during experiential learning in humans. Model-free learning algorithms iteratively update the value of an action using reward prediction errors (RPE), which represent whether the outcome of taking that action was more or less rewarding than expected. Such RPE signals have been extensively reported within striatum during experiential learning in human neuroimaging studies

(McClure et al. 2003; O'Doherty et al. 2003; Gläscher et al. 2010). This literature suggests that the ventral aspects of striatum may be involved in encoding RPEs when learning about the value of stimuli as opposed to actions (O'Doherty et al. 2004; Cooper et al. 2012; Chase et al. 2015). In addition, there is growing evidence for the encoding of model-based state-prediction errors (SPEs) by a network of frontoparietal regions (Gläscher et al. 2010; Liljeholm et al. 2013; Lee et al. 2014). These update signals reflect how surprising the outcome of a given action is, irrespective of its reward value, and can be used to update the probabilistic model of contingencies linking actions and the identity of their outcomes that is maintained by a model-based learning algorithm.

On the basis of evidence for the existence of both model-based and model-free learning signals in the experiential domain, a key objective of the present study was to establish whether learning about the value of different options through observation would involve a model-free mechanism, a model-based mechanism or both.

To address these questions, we recruited human participants to play a multi-armed bandit task (see Figure 1) while they underwent functional magnetic resonance imaging (fMRI). In the task, participants watched an observee play different colored slot machines. Importantly, because the observee only had a single slot machine to choose from on each trial, participants could not learn about the value of the slot machines from the observee's actions; they could only learn by observing the payouts experienced by the observee. In order to assess whether the participants had learned from observation, they occasionally made choices between the slot machines they had watched the observee play. The chosen slot machine's payout on these trials was added to the participants earnings but was hidden from them, which allowed us to incentivise learning from observation while preventing participants from learning about the machines experientially. In order to enable a direct comparison of the neural mechanisms underlying observational and experiential learning, we also included an experiential learning condition, which used a different set of slot machines. This condition was identical to the observational one, except that the participants themselves played and experienced the payouts. We then fitted model-free and model-based learning algorithms to participants' choices in this task, and derived from these fitted models both reward prediction error and state prediction error regressors for use in our analysis of the participants' fMRI data. While in this task the estimated values of the slot machines generated by the model-free and model-based algorithms coincide, their respective update signals do not, allowing us to distinguish their neural representations. Although the observational condition differs from the experiential condition in that the observational choice trials are more valuable than the observational learning trials, our effects of interest occur only on the learning trials in both conditions. Therefore, these differences would not be a confounding factor in our analysis.

We predicted that experiential and observational learning about the value of the different slot machines would share similar neural substrates, Specifically, we hypothesized that a frontoparietal network would encode state prediction errors during both experiential and observational learning, while the ventral aspect of striatum would be involved in encoding reward prediction errors during both experiential and observational learning.

Materials and Methods

Procedures

Seventeen healthy young adults (mean age 23.3 years, SD 3.62 years, 8 males) participated in our neuroimaging study. All participants provided written informed consent. The study was approved by the Research Ethics Committee of the School of Psychology at Trinity College Dublin.

Each participant attended Trinity College Institute of Neuroscience where they received instruction in the task (See Figure 1 and below). The participant was then introduced to a confederate (the *observee*), who they would subsequently watch playing a subset of the trials of the bandit task. Immediately before the task began, participants saw the experimenter supposedly test the video connection to the observee. The participant then completed the bandit task while undergoing MR imaging. In a post-scan debriefing, all participants reported having believed the video of the observee shown during the bandit task to have been a live feed. In reality, this video was recorded before the experiment.

Bandit Task

Participants faced slot machines that, when played, delivered a positive monetary payoff (€0.20) or nothing, with differing reward probabilities that changed independently and continuously over the course of the task. Each machine's reward probability time-course was a sine curve that drifted between 0 and 100% plus a small amount of Gaussian noise on each trial (M = 0, SD = 6), with a random starting point and half-period randomly set between 0.87 and 1.67 times the number of trials per condition. The reward probabilities were constrained to be correlated with each other at no greater than r = 0.02. The reward probability time-courses assigned to each condition were counterbalanced across participants. These slot machines were uniquely identifiable by their color. Three slot machines were assigned to an *experiential learning* condition and the remaining three to an *observational learning* condition. This separation of slot machines by condition allowed us to be confident that any neural effects of learning were solely attributable to experiential or observational learning. In each condition, one 'neutral' slot machine always paid €0.00 with 100% probability. These neutral slot machines were intended to control for visuomotor effects, but were not ultimately utilized in the analysis, because our use of parametric regressors implicitly controls for such effects. In both conditions, participants faced a mixture of *forced-choice* trials, on which they could learn about the probability of payoff associated with each slot machine, and *free-choice* trials, on which they could use this knowledge to maximize their earnings. The use of forced-choice trials allowed us to exclude the possibility that in the observational condition participants were mimicking the choices of the observer rather than learning the value of each slot machine. Free-choice trials were included to allow us to assess whether participants were learning from the forced-choice trials. The task was blocked by condition, with 28 trials presented in each block. Free-choice trials made up one quarter of all trials, and were randomly interleaved among the forced-choice trials. The task was presented in four runs of 3 blocks or approximately 17 minutes each.

On forced-choice trials, a single slot machine appeared on the left or right side of the screen. The player had a maximum of two seconds to play the machine by indicating the side of the screen it was on, using a keypad. The slot machine lever was pulled and its reel spun for four seconds. The slot machine then disappeared and was replaced by either an image of a coin, indicating to the participant that they had earned €0.20, or an image of a scrambled or crossed-out coin indicating to the player that they had earned nothing. After two seconds the payoff image disappeared and the trial was followed by an inter-trial interval (ITI) with a duration drawn from a uniform distribution (minimum = 1 seconds, maximum = 7 seconds), during which a white crosshairs was displayed on a black background. On forced-choice trials in the experiential condition, the participant played the slot machine and earned the payoff it delivered, while on forced-choice trials in the observational condition the participant watched video of the slot machine being played by the observee. The observee was shown seated on the left side of the screen with their back to the camera in front of a monitor, which displayed the task. On these trials, the participant observed and did not earn the payoffs delivered by the slot machine. However, it remained in the interest of participants to attend to these observed trials because they would make choices between the slot machines shown on these trials on subsequent free-choice trials.

On free-choice trials, the two slot machines with a non-zero probability of reward from the current condition appeared on screen. Participants had a maximum of two seconds to choose to play one of the machines. The lever of the selected slot machine was pulled and its reel spun for four seconds. The inter-trial interval began immediately after the reel had finished spinning. As in forced-choice trials, the slot machines paid out according to their associated reward probability. The payoff was not displayed to the participant on free-choice trials in order to confine their learning to the forced-choice trials. This also had the benefit of preventing potential indirect effects of receipt of reinforcement, such as increased attention, from influencing participants learning through observation. The payoff earned on a free-choice trial was however added to the participant's earnings thus incentivising them to choose the machine they believed most likely to pay out.

If participants failed to respond to a trial within two seconds or responded incorrectly to a forced-choice trial in the experiential condition, the slot machine cues disappeared and were replaced by text stating 'Invalid or late choice'. This remained onscreen for the remainder of the trial.

Imaging Procedures

Magnetic resonance imaging was carried out with a Philips Achieva 3T scanner with an eight-channel SENSE (sensitivity encoding) head coil. T2*-weighted echo-planar volumes with BOLD (blood oxygen level dependent) contrast were acquired at a 30 degree angle to the anterior commissure-posterior commissure line, to attenuate signal dropout at the orbitofrontal cortex (Deichmann et al. 2003). Thirty-nine ascending slices were acquired in each volume, with an in-plane resolution of 3×3mm, and slice thickness of 3.55mm [TR: 2000ms; TE: 30ms; FOV: 240×240×138.45mm; matrix 80×80]. Data was acquired in four sessions, each comprising 516 volumes. Whole-brain high-resolution T1-weighted structural scans (voxel size: 0.9×0.9×0.9mm) were also acquired for each participant.

Computational Modeling

Participants' choices were modeled using both model-free and model-based learning algorithms. The model-free algorithm used was a variation on the SARSA reinforcement learning algorithm (Sutton and Barto 1998) together with a softmax decision rule. This algorithm iteratively updates a 'cached' value for taking an action in a particular context. The values of slot machines played by the participant and those played by the observee were updated in the same manner. Specifically, all slot machines began with an initial value of 0. If a given slot machine did not display a payoff on a particular trial, its value remained unchanged. If a slot machine *i* displayed a payoff on trial *t*, its value *V* was updated according to the rule:

$$V_i^{t+1}=V_i^t+\alpha\delta_{i,RPE}^t$$

$$\delta_i^t=O_i^t-V_i^t$$

where α, δ_{RPE} and O refer to the learning rate, reward prediction error and payoff value respectively.

In contrast, the model-based algorithm estimates the transition probabilities linking each slot machine to the two possible outcome states (reward and no reward). If a slot machine *i* from condition *j* led to an outcome state *s* on trial *t*, the transition probability $T(i, s)$ was updated according to the rule:

$$T(i,s)^{t+1}=T(i,s)^t+\eta\delta_{\mathrm{i,SPE}}^t$$

$$\delta_{i,SPE}^t=1-T(i,s)^t$$

where η and δ_{SPE} refer to the learning rate and state prediction error respectively. The estimated transition probability for the outcome state not arrived in was reduced according to $T(i,s')^{t+1}=T(i,s')^t(1-\eta)$ to ensure $\sum_s T(i,s)^{t+1}=1$. The reward function over outcome states was defined as $r(s)=1$ if *s* is the reward outcome state, otherwise $r(s)=0$. The estimated transition probabilities were integrated with the reward function to compute the expected reward from playing a slot machine, i.e.

$$V_i^t=\sum_s T(i,s)^t r(s)$$

For both model-based and model-free learning algorithms, the predicted choice probabilities were obtained by passing the values of the slot machines available for choice to a softmax choice rule with temperature parameter θ. Due to the simple structure of the bandit task, the

model-free and model-based algorithms make identical behavioral predictions and the state prediction errors of the model-based algorithm were equivalent to the absolute value of the reward prediction errors from the model-free algorithm. Importantly, such unsigned prediction errors cannot be used to update the cached values maintained by a model-free learning algorithm because they do not reflect the reinforcing value of the outcome of an action. Separate learning rate and temperature parameters were fit to the pooled participants' choices by maximizing the product of the model-predicted probabilities of the participants' choices (maximum likelihood estimation) using fminsearch in MATLAB. The same parameter estimates were used for the model-free and model-based learning algorithms. We used the Bayesian Information Criterion (BIC) to evaluate the learning algorithms according to their goodness of fit and complexity. Using this procedure, we selected a model with a single learning rate (BIC = 1428.6, learning rate = 0.14, $\theta = 6.18$) that outperformed alternative models which allowed for different learning rates to be associated with the observed and experienced payoffs (BIC = 1434.7), or allowed different softmax temperatures to be associated with free-choice trials in the observed and experienced conditions (BIC = 1463.9). Prediction error values were derived from this fitted model and used in our analysis of the BOLD data. In order to assess the sensitivity of any neural effects to the fitted parameter values we also carried out analyses of the BOLD data with prediction error regressors generated using learning rates that deviated from the fitted values (0.05, 0.25).

fMRI Preprocessing

All image preprocessing and analysis was performed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK; available at http://www.fil.ion.ucl.ac.uk/spm). All functional volumes were corrected for differences in acquisition time between slices (to the middle slice), realigned to the first volume, and coregistered with the high-resolution structural image. The coregistered high-resolution structural image was segmented and normalised to Montreal Neurological Institute (MNI) space using Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL). The resulting transformation was applied to the functional volumes. The functional volumes were spatially smoothed with a Gaussian kernel (full-width at half-maximum = 8 mm) and high-pass temporally filtered (60s).

fMRI Statistical Analysis

We analyzed the BOLD data using a GLM with participant-specific design matrices. Forced-choice trials were modeled with a regressor indicating the onset times of the slot machine cue and another indicating the onset times of the payout cue. Forced-choice trials containing slot machines that deterministically paid out €0.00 were modeled with separate cue and payout onset regressors. A parametric regressor representing the trial-by-trial prediction error estimates obtained from the behavioral model modulated the payout onset regressor. Another regressor modeled the cue onset on free-choice trials. The experiential and observational conditions were modeled with separate sets of onset and parametric regressors, allowing us to control for any average differences in the BOLD data between the conditions that may be due to visual, social or motor factors. Each regressor was convolved with a canonical haemodynamic response function after being entered into SPM8 to generate a

design matrix. In order to be confident in our estimates of the effects of our regressors on BOLD, we tested for the presence of multicollinearity by measuring the Variance Inflation Factor (VIF) associated with each convolved regressor for each participant. This was obtained by performing OLS regression of each regressor on all other regressors. The VIF is then defined as $1/(1 - R^2)$, where R^2 is the coefficient of determination from this regression. The VIF values for all regressors of all participants were less than 2.5, substantially less than conventional cutoff values of 10, 20 or 40 (O'Brien 2007), indicating that multicollinearity was not present. We do not include regressors representing the times of decision in each condition because a VIF analysis indicated that these would introduce a high degree of multicollinearity into the design matrix (VIF_{OBS} = 10.13, VIF_{EXP} = 23.56; averaged across participants). This is attributable to the small temporal separation between the appearance of the cue and the time of decision leading to extremely high correlations between their regressors in both the observational and experiential learning conditions (R_{OBS} = 0.94, R_{EXP} = 0.97; averaged across participants). The presence of such multicollinearity would prevent us from accurately distinguishing the effects of cue onset and decision onset. Motion parameters estimated during the realignment procedure were also included as regressors of no interest.

Maps of the voxel-wise parameter estimates for the regressors of interest were entered in between-subjects random effects analyses testing the effect of the regressors across the group. Unless otherwise stated, we report statistics with an uncorrected threshold of $p_{UNC} < 0.005$, after correction for multiple comparisons across the whole brain (WBC) to $p_{FWE} < 0.05$ using 3dFWHMx to estimate the smoothness of the residual images of each contrast and AlphaSim to calculate extent thresholds (AFNI, Cox 1996). A substantial number of studies have reported reward prediction error encoding in ventral striatum (Pagnoni et al. 2002; O'Doherty et al. 2004; Pessiglione et al. 2006). Therefore, we also carried out small-volume corrections (SVCs) on ventral striatum using an independent anatomical mask composed of two spherical ROIs of radius 5mm centered on the bilateral MNI coordinates of ventral striatum reported by a previous study of functional connectivity (Di Martino et al. 2008).

Results

Behavioral results

We began by confirming that all participants chose the slot machine with the higher reward probability significantly more often than an agent choosing at random (p=0.5), as determined by a one-tailed binomial test. We then tested for a performance difference between the experiential and observational learning condition using a paired *t*-test (see Figure 2A). Again, performance was defined as the proportion of choice trials on which the slot machine with the higher win probability was chosen. We found significantly higher performance in the experiential learning condition relative to the observational learning condition ($t(16)=2.26$, $p<0.04$, two-tailed). Participants' performance in the first half of the task (M = 0.84, SD = 0.11) in the experiential condition was significantly ($t(16) = 2.54$, $p = 0.02$, two-tailed) better in the second half (M = 0.72, SD = 0.15). Participants' performance in the observational condition did not differ between the first (M = 0.73, SD = 0.18) and

second half (M = 0.69, SD = 0.15) of the task. We found a significant effect of task difficulty, defined as the difference in the reward probabilities of the slot machines available for choice, on trial-by-trial performance in 9/17 participants in the experiential condition, and in 8/17 participants in the observational condition, after probit regressions. Levene's tests did not indicate inhomogeneity of variance between their performance on the experiential and observational condition (F=0.004, p=0.95). Participants' reaction times on observational free-choice trials (M = 827ms, SD = 115ms) were also significantly slower (t(16)=-2.59, p <0.02, two-tailed) than those on experiential free-choice trials (M = 776ms, SD = 123ms).

Neuroimaging results

Model-free learning signals—We tested for the presence of a relationship between BOLD activity and the model-derived RPEs, signals used by model-free learning algorithms to update the cached value of performing an action in a particular context (see Figure 3 and Table 1).

We found a significant effect of RPE in the experiential condition (RPE_{EXP}) on BOLD in ventral striatum [$p_{FWE} < 0.05$, SVC; x,y,z = 10,10,-6], as well as posterior cingulate [$p_{FWE} < 0.05$, WBC; x,y,z = 0,-38,32] and dorsomedial prefrontal cortex [$p_{FWE} < 0.05$, WBC; x,y,z = -4,62,6]. With the exception of the activation in dorsomedial prefrontal cortex, these effects were robust to deviations in the value of the learning rate parameter from the fitted value (see Table 1), consistent with a previous report (Wilson and Niv 2015). The RPE_{EXP} effects in the first and second halves of the task did not differ significantly nor did their difference covary with performance differences in the experiential condition between the first and second halves. In contrast to our findings for RPE_{EXP}, we found no positive or negative effect on BOLD of the RPE associated with the observed outcomes in the observational condition (RPE_{OBS}) in our ventral striatum ROI after small-volume correction, or elsewhere at our whole-brain threshold. We do not believe this finding is attributable to imprecise fitting of the behavioral model parameters because this absence of RPE_{OBS} encoding was robust to deviations in the learning rate parameter from the fitted value.

We then performed a formal two-tailed *t*-test on the differences between the effect of the RPE signal in the experiential and observational conditions ($RPE_{OBS} > RPE_{EXP}$, $RPE_{EXP} > RPE_{OBS}$). Of the areas we had found to be sensitive to RPE_{EXP}, only ventral striatum [$p_{FWE} < 0.05$, two-tailed; SVC; x,y,z = -6,10,-4] responded more strongly to RPE_{EXP} than to RPE_{OBS}. BOLD in a region of right middle occipital gyrus [$p_{FWE} < 0.05$, two-tailed, WBC; x,y,z = 22,-84,10] was not significantly related to either RPE_{EXP} or RPE_{OBS} but showed a significantly more positive effect of RPE_{OBS} than RPE_{EXP}.

In order to determine whether these differences between experiential and observational learning in the neural encoding of RPE were associated with differences in task performance between the conditions, we repeated this two-tailed *t*-test, including a covariate representing the difference between the conditions in the proportion of choices for the slot machine with the greater probability of reward for each participant. Following this test, RPE_{EXP} signaling in ventral striatum remained significantly greater than that of RPE_{OBS} [$p_{FWE} < 0.05$, two-tailed; SVC; x,y,z = -6,10,-4], indicating that this difference is not attributable to

performance differences between experiential and observational learning. In contrast, the cluster in right middle occipital gyrus that showed a significantly more positive effect of RPE_{OBS} than RPE_{EXP} did not survive the inclusion of this covariate.

Model-based learning signals—Next we tested for the neural representation of SPEs, error signals used by model-based learning algorithms to update probabilistic representations of environmental contingencies linking actions to subsequent states

We found (see Figure 5 and Table 1) a significant effect of the SPE associated with the earned outcomes in the experiential condition (SPE_{EXP}) in left intraparietal sulcus [$p_{FWE} < 0.05$, WBC; x,y,z = -48,-68,28] and right dorsomedial prefrontal cortex [$p_{FWE} < 0.05$, WBC; x,y,z = 4,34,42]. We found significant effects of BOLD to SPE_{OBS} in left intraparietal sulcus/inferior parietal lobule [$p_{FWE} < 0.05$, WBC; x,y,z = -44,-62,42] as well as left precuneus [$p_{FWE} < 0.05$, WBC; x,y,z = 4, -62, 40]. In contrast to the model-free RPE activations, these model-based SPE effects were highly sensitive to deviations in the value of the learning rate parameter from the fitted value, with none remaining significant at both alternative values (see Table 1). The effects of SPE_{EXP} on BOLD in the first and second halves of the task also did not differ significantly nor did their difference covary with performance differences in the experiential condition between the first and second halves. After performing a two-tailed *t*-test on the difference between the effects of SPE_{EXP} and SPE_{OBS}, we found that BOLD in none of the areas that responded positively to either SPE_{EXP} or SPE_{OBS} exhibited an effects of SPE_{EXP} that was significantly different from the effect of SPE_{OBS}. A single region of right middle temporal gyrus showed a significantly more positive effect of SPE_{OBS} than of SPE_{EXP} [$p_{FWE} < 0.05$, WBC; x,y,z = 40,-72,12], but because this area did not respond significantly to SPE_{OBS} in its own right, we do not consider it further. Given the similarity of the neural effects of SPE_{EXP} and SPE_{OBS}, we merged the regressors into a single SPE regressor in order to provide a test with greater statistical power. However, this analysis did not reveal any additional regions that were uniquely sensitive to this merged SPE regressor (see Table 1).

No effects of task performance on BOLD—We did not find any effects of task performance on BOLD response to the cue on free- or forced-choice trials, to the outcome on forced-choice trials, or to the prediction error regressors in either the observational or experiential condition.

Discussion

In this study, we examined the neural correlates of computations underlying learning from the outcomes of actions we observe others take in the absence of choice.

We found that this form of observational learning differs significantly from its experiential analog in the neural representation of the RPE, a signal that can be used to update ‘cached’ model-free action values (Daw et al. 2005). BOLD activity correlated with model-free RPE during experiential learning in ventral striatum, replicating previous findings (Pagnoni et al. 2002; O’Doherty et al. 2003; Pessiglione et al. 2006) as well as dmPFC, which has also previously been associated with both social and non-social prediction error signals (Behrens

et al. 2009, 2009; Yau and McNally 2015). However, activity in ventral striatum differed significantly during observational learning, with no evidence of an analogous observational RPE signal in ventral striatum at our testing threshold.

It is unlikely that this difference is attributable to gross differences in the visual, social, or motor properties of the conditions, which were controlled for by the use of condition-specific event onset regressors. The absence of model-free signaling in the observational learning condition is also unlikely to reflect reduced salience of this condition because we successfully detect other, model-based, signals during observational learning. This selective encoding of RPE during experiential learning provides new insight into the computational role of ventral striatum, indicating that it may not be recruited for the acquisition of model-free associations when we learn from the consequences of actions we observe others take. Our experiential and observational learning conditions necessarily differ in multiple respects that may have influenced what type of learning mechanism is engaged; most prominently, the presence of the observee and the receipt of reinforcement by the participant. An important goal for future research should be to identify what elements of observational learning give rise to this absence of RPE signaling in ventral striatum. It may be that the lack of experienced reward during observational learning prevents engagement of a model-free learning mechanism that relies on the receipt of reinforcement. Alternatively, the presence of the observee may suppress the use of model-free learning in favor of model-based updating strategies.

While in the current study participants could only observe the outcomes received by the observee and could not be influenced by their choice of action, in previous studies participants watched the observee make explicit choices between actions and receive the resulting outcomes (Burke et al. 2010; Suzuki et al. 2011; Cooper et al. 2012). From these studies, there is evidence to suggest that the differential sensitivity of ventral striatum to RPE encoding we find may not be limited to learning from the outcomes of actions taken in the absence of choice, although these studies did not explicitly test for differences between observational and experiential learning signals. Our finding that experiential but not observational learning is associated with RPE signaling in ventral striatum is consistent with that of Cooper et al. (2012), who reported sensitivity of BOLD in ventral striatum to RPE when participants learned from the outcomes of their choices in a multi-armed bandit task but find no such RPE effect when participants learned by observing another player perform the same task. Suzuki et al. (2012) also find RPE encoding in ventral striatum when participants chose between stimuli associated with probabilistic monetary reinforcement, but this BOLD effect in striatum is absent when participants learn to predict the choices of others by observing them perform the task. Interestingly, Burke et al. (2010) report a negative effect of RPE in ventral striatum associated with the outcomes of observed choices as well as positive RPE encoding during experiential learning. They also suggested however that this result could potentially be attributable to a social comparison effect, by which observing rewards being denied to another person may itself be rewarding (Delgado et al. 2005; Fliessbach et al. 2007). Taken together, these results suggest that ventral striatum may not possess the computational flexibility required to allow the outcomes of actions we observe other social agents take to update stored values in the manner that experienced outcomes do. Alternatively, we cannot exclude the possibility that ventral striatum may only

engage in this form of updating in particularly evocative social contexts that have not yet been explored experimentally. This would be consistent with reports of modulation of ventral striatal BOLD responsiveness to reward received by others by interpersonal factors such as their perceived similarity to the participant (Mobbs et al. 2009) or whether they are cooperating or in competition with the participant (de Bruijn et al. 2009).

Previous studies have reported RPE effects accompanying observed outcomes in BOLD in dorsal striatum (Cooper et al. 2012) and vmPFC (Burke et al. 2010; Suzuki et al. 2012). When these results are taken together with the absence of RPE effects during observational learning in the current study, one possibility is that these neural circuits are more engaged during the processing of outcomes of observed actions that are freely chosen. This would echo accumulated evidence from studies of experiential learning indicating that dorsal striatum in particular is engaged selectively for model-free updating when actions can be freely chosen (O'Doherty et al. 2004; Gläscher et al. 2009; Cooper et al. 2012).

Despite the absence of model-free RPE signaling, participants' choices clearly indicated that they used the observed outcomes to learn values for the bandits, potentially by relying instead on model-based learning. We found neural representations of SPE signals used by such an algorithm to update the probabilistic contingencies linking environmental states in response to both experienced and observed outcomes in our task in left intraparietal sulcus – a region that has been associated with SPE signals in previous studies (Gläscher et al. 2010; Liljeholm et al. 2013; Lee et al. 2014). Our findings represent the first time such signals have been tested for and observed during observational learning.

Interestingly, although model-based learning is frequently associated with greater computational flexibility, our observational learning condition demands no more computational flexibility than the experiential learning condition; the conditions do not differ in their task structure, or the reward information available to the participant. Thus, the fact that we find only model-based update signals during observational learning is unexpected, and demonstrates that model-based learning signals occur across a broader range of domains than model-free learning, even when the greater computational flexibility of model-based learning is not required.

It should be noted that the observational SPE signal we discuss here differs, both computationally and in terms of its neural substrates, from other error signals reported during observational learning (Behrens et al. 2008; Burke et al. 2010; Suzuki et al. 2012) that reflected violations of predictions about the actions an observee will choose to take. These were used to improve those predictions (Suzuki et al. 2012), to learn about the observee's intentions (Behrens et al. 2008), or to bias ones own choices towards advantageous actions (Burke et al. 2010). In contrast, in our task the observee did not make choices between actions and the SPE signal instead reflected violations of the predicted consequences of the observee's actions. In addition, while we found BOLD correlates of the observational SPE in left intraparietal sulcus and left precuneus, action prediction errors have been reported in dmPFC (Behrens et al. 2008; Suzuki et al. 2012) and dlPFC (Burke et al. 2010; Suzuki et al. 2012), suggesting that learning to predict the outcomes of others

actions by observation is implemented by neural circuits that are distinct from those used to learn to predict the actions themselves (Dunne and O'Doherty 2013).

In addition to learning by observing the consequences of others actions, humans can exploit other forms of non-experiential learning. For example, we can also learn from the consequences of actions we could have but did not take, or by *fictive* learning. This can be seen in the investor who, having bought shares in a public company, can learn about the shrewdness of his choice from the changing share price of not only the company he invested in, but other companies he chose not to invest in. This phenomenon has been explored from a model-free standpoint (Daw et al. 2005; Lohrenz et al. 2007; Li and Daw 2011) but it remains to be seen whether fictive learning may also be supported by the strengthening of model-based associations.

In summary, we demonstrate that when learning by observing the experiences of others in the absence of free-choice, state prediction errors signals associated with model-based learning are present in the frontoparietal network, while reward prediction error signals associated with model-free reinforcement learning are not evident. These results illustrate the adaptability of the model-based learning system, and suggest that its apparent ability to incorporate information gleaned from both the observation of the consequences of others' actions and the experience of the outcomes of our own actions may be fundamental to our ability to efficiently assimilate diverse forms of information.

Acknowledgements

We thank Sojo Joseph and Bridget Finnegan for their assistance during data collection.

Grants

This work was funded by a grant from the Wellcome Trust to JPO.

References

1. Balleine BW, Daw ND, O'Doherty JP. Multiple forms of value learning and the function of dopamine. Neuroeconomics: decision making and the brain. 2008:367–385.
2. Bandura, A. Social Learning Theory. Englewood Cliffs, NJ: Prentice Hall; 1977.
3. Behrens TEJ, Hunt LT, Rushworth MFS. The computation of social behavior. Science. 2009; 324:1160–1164. [PubMed: 19478175]
4. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature. 2008; 456:245–249. [PubMed: 19005555]
5. Berger SM. Conditioning through vicarious instigation. Psychol Rev. 1962; 69:450. [PubMed: 13867454]
6. de Bruijn ERA, de Lange FP, von Cramon DY, Ullsperger M. When errors are rewarding. J Neurosci. 2009; 29:12183–12186. [PubMed: 19793976]
7. Burke CJ, Tobler PN, Baddeley M, Schultz W. Neural nechanisms of observational learning. Proc Natl Acad Sci. 2010; 107:14431–14436. [PubMed: 20660717]
8. Camerer C, Ho TH. Experience-weighted attraction learning in normal form games. Econometrica. 1999; 67:827–874.
9. Chase HW, Kumar P, Eickhoff SB, Dombrovski AY. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cogn Affect Behav Neurosci. 2015; 15:435–459. [PubMed: 25665667]

10. Cooper JC, Dunne S, Furey T, O'Doherty JP. Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. J Cogn Neurosci. 2012; 24:106–118. [PubMed: 21812568]
11. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res Int J. 1996; 29:162–173.
12. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8:1704–1711. [PubMed: 16286932]
13. Deichmann R, Gottfried JA, Hutton C, Turner R. Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage. 2003; 19:430–441. [PubMed: 12814592]
14. Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci. 2005; 8:1611. [PubMed: 16222226]
15. Di Martino A, Scheres A, Margulies DS, Kelly AMC, Uddin LQ, Shehzad Z, Biswal B, Walters JR, Castellanos FX, Milham MP. Functional connectivity of human striatum: A resting state fMRI study. Cereb Cortex. 2008; 18:2735–2747. [PubMed: 18400794]
16. Dunne S, O'Doherty JP. Insights from the application of computational neuroimaging to social neuroscience. Curr Opin Neurobiol.
17. Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A. Social comparison affects reward-related brain activity in the human ventral striatum. Science. 2007; 318:1305–1308. [PubMed: 18033886]
18. Galef BG, Laland KN. Social learning in animals: empirical studies and theoretical models. BioScience. 2005; 55:489–499.
19. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66:585–595. [PubMed: 20510862]
20. Gläscher J, Hampton AN, O'Doherty JP. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex. 2009; 19:483–495. [PubMed: 18550593]
21. Grüter C, Farina WM. The honeybee waggle dance: can we follow the steps? Trends Ecol Evol. 2009; 24:242–247. [PubMed: 19307042]
22. Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci. 2008; 105:6741–6746. [PubMed: 18427116]
23. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 2014; 81:687–699. [PubMed: 24507199]
24. Li J, Daw ND. Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci. 2011; 31:5504–5511. [PubMed: 21471387]
25. Liljeholm M, Wang S, Zhang J, O'Doherty JP. Neural correlates of the divergence of instrumental probability distributions. J Neurosci. 2013; 33:12519–12527. [PubMed: 23884955]
26. Lohrenz T, McCabe K, Camerer CF, Montague PR. Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci. 2007; 104:9493–9498. [PubMed: 17519340]
27. McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003; 38:339–46. [PubMed: 12718866]
28. Mobbs D, Yu R, Meyer M, Passamonti L, Seymour B, Calder AJ, Schweizer S, Frith CD, Dalgleish T. A key role for similarity in vicarious reward. Science. 2009; 324:900–900. [PubMed: 19443777]
29. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007; 41:673–690.
30. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003; 38:329. [PubMed: 12718865]
31. O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304:452–454. [PubMed: 15087550]

32. Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. Nat Neurosci. 2002; 5:97–98. [PubMed: 11802175]
33. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006; 442:1042–1045. [PubMed: 16929307]
34. Rescorla, R.; Wagner, A. A theory of classical conditioning: variations in the effectiveness of reinforcement and non-reinforcement. Classical conditioning II: Current research and theory. Black, A.; Prokasy, W., editors. New-York: Appleton Century Crofts; 1972. p. 64-99.
35. Sutton, RS.; Barto, AG. Introduction to reinforcement learning. Cambridge, MA: MIT Press; 1998.
36. Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H. Learning to simulate others' decisions. Neuron. 2012; 74:1125–1137. [PubMed: 22726841]
37. Suzuki S, Niki K, Fujisaki S, Akiyama E. Neural basis of conditional cooperation. Soc Cogn Affect Neurosci. 2011; 6:338–347. [PubMed: 20501484]
38. Tomasello M, Davis-Dasilva M, Camak L, Bard K. Observational learning of tool-use by young chimpanzees. Hum Evol. 1987; 2:175–183.
39. Whiten A, Horner V, de Waal FBM. Conformity to cultural norms of tool use in chimpanzees. Nature. 2005; 437:737–740. [PubMed: 16113685]
40. Wilson RC, Niv Y. Is model fitting necessary for model-based fMRI? PLoS Comput Biol. 2015; 11:e1004237. [PubMed: 26086934]
41. Yau JOY, McNally GP. Pharmacogenetic excitation of dorsomedial prefrontal cortex restores fear prediction error. J Neurosci. 2015; 35:74–83. [PubMed: 25568104]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

New and Noteworthy

Here we describe evidence for both common and distinct neural computations during reward learning through observation and through direct experience. Specifically, we report encoding of a state prediction error signal associated with model-based reinforcement-learning in the frontoparietal network during observational as well as experiential learning. In contrast, although we find encoding of model-free reward prediction errors in ventral striatum during experiential learning, this signal is absent during observational learning.

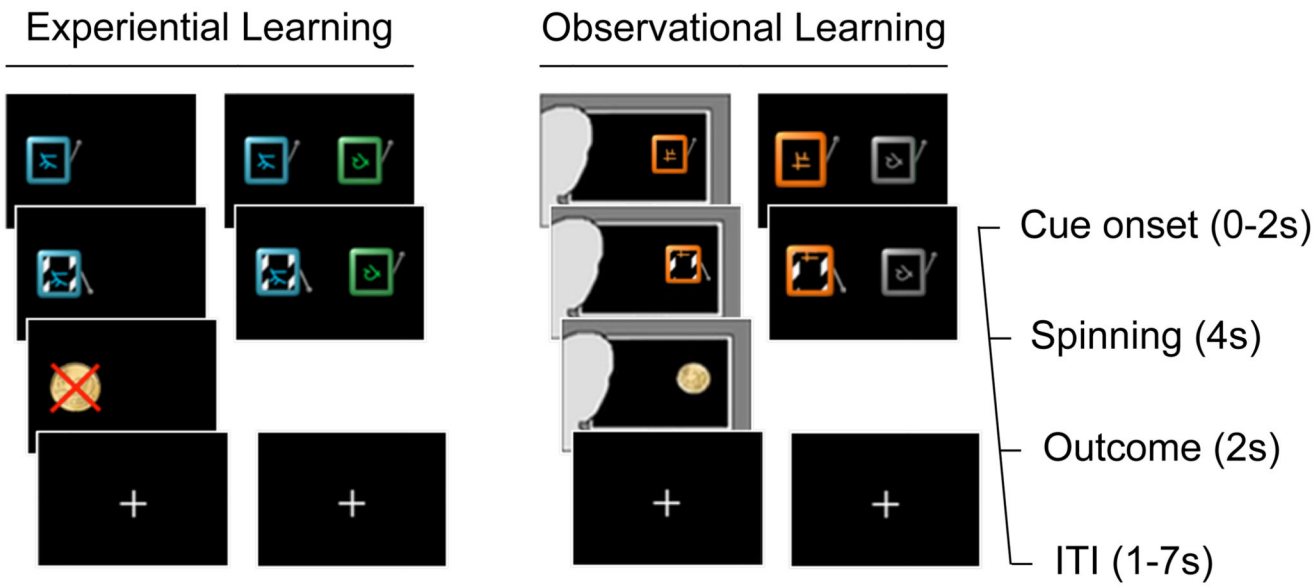

Figure 1. Task Schematic.

Participants completed a multi-armed bandit task, with an experiential and an observational learning condition. Individual slot machines were played on forced-choice trials and participants made choices between pairs of slot machines on free-choice trials. Each slot machine paid out a reward (€0.20) or nothing, with a reward probability that changed independently across machines and continuously throughout the task. On experiential forced-choice trials, the participant played the slot machine and earned the amount paid out, while on observational forced-choice trials they watched video of an observee playing the slot machine and earning the amount paid out. On all trials, a slot machine was selected for play within 2 seconds of the onset of a trial, after which the reels of that slot machine spun for 4 seconds. On forced-choice trials, the amount paid out was displayed for 2 seconds. On free-choice trials, the amount paid out was not displayed to the participant but was added to the participant's earnings. All trials were followed by an ITI whose duration was drawn randomly from a discrete uniform distribution (min=1s, max=7s).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

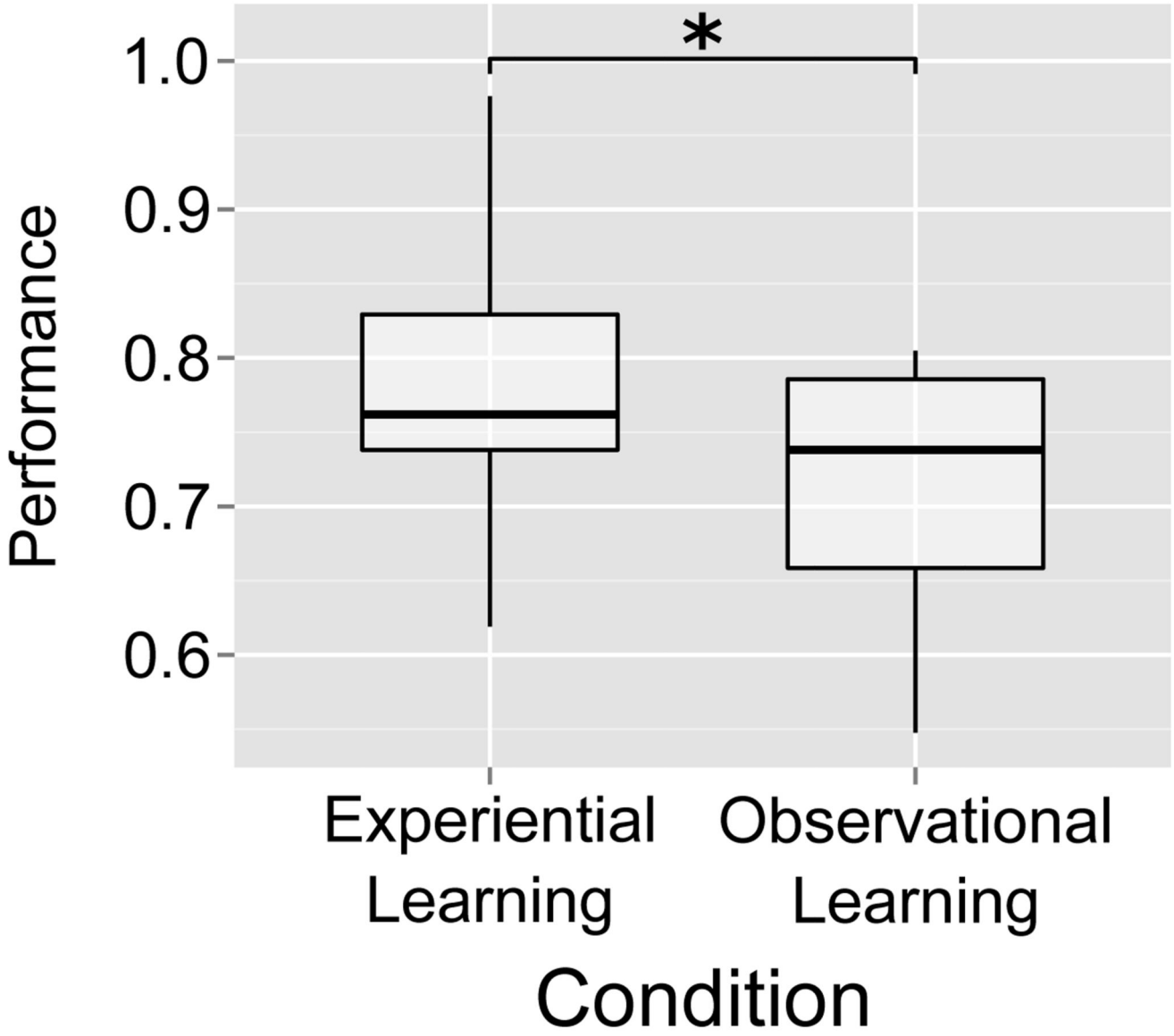

Figure 2. Behavioral performance.

Box plot of participants' performance, defined as the proportion of free-choice trials on which the slot machine with the highest probability of paying out was chosen, in the experiential and observational conditions. Horizontal bar represents median performance, box represents interquartile range, and whisker ends represent maximum and minimum performance values. Average performance was higher in the experiential learning condition (mean = 0.77) than in the observational learning condition (mean = 0.71).

Europe PMC Funders Author Manuscripts

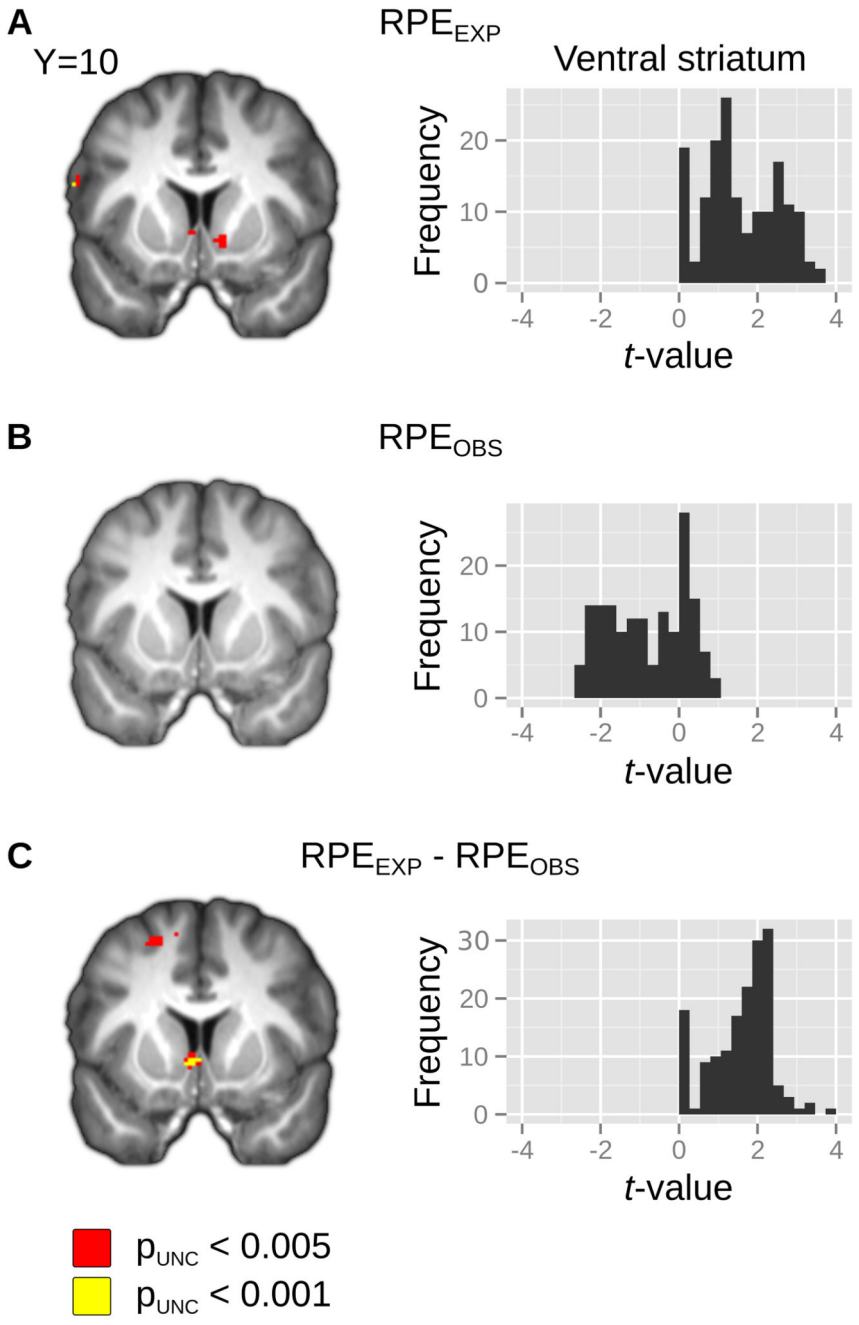

Figure 3. Effect of reward prediction error in ventral striatum.

We found that BOLD activity in ventral striatum was sensitive to RPEs associated with outcomes that were earned (A), but found no RPE representation in ventral striatum associated with outcomes that were observed (B). The differences between the effects of experiential and observational RPEs were significant in ventral striatum (C). SPMs of the effects (left) are overlaid on the group mean normalized anatomical image and thresholded at $p_{UNC} < 0.005$ and $p_{UNC} < 0.001$ for the purpose of illustration. Histograms (right) represent the frequency of voxel-wise *t*-statistics in our bilateral ventral striatum ROI. For

the contrast of negative effect of RPE_{OBS}, no voxels in our ROI survive an uncorrected threshold of $p_{UNC} < 0.005$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

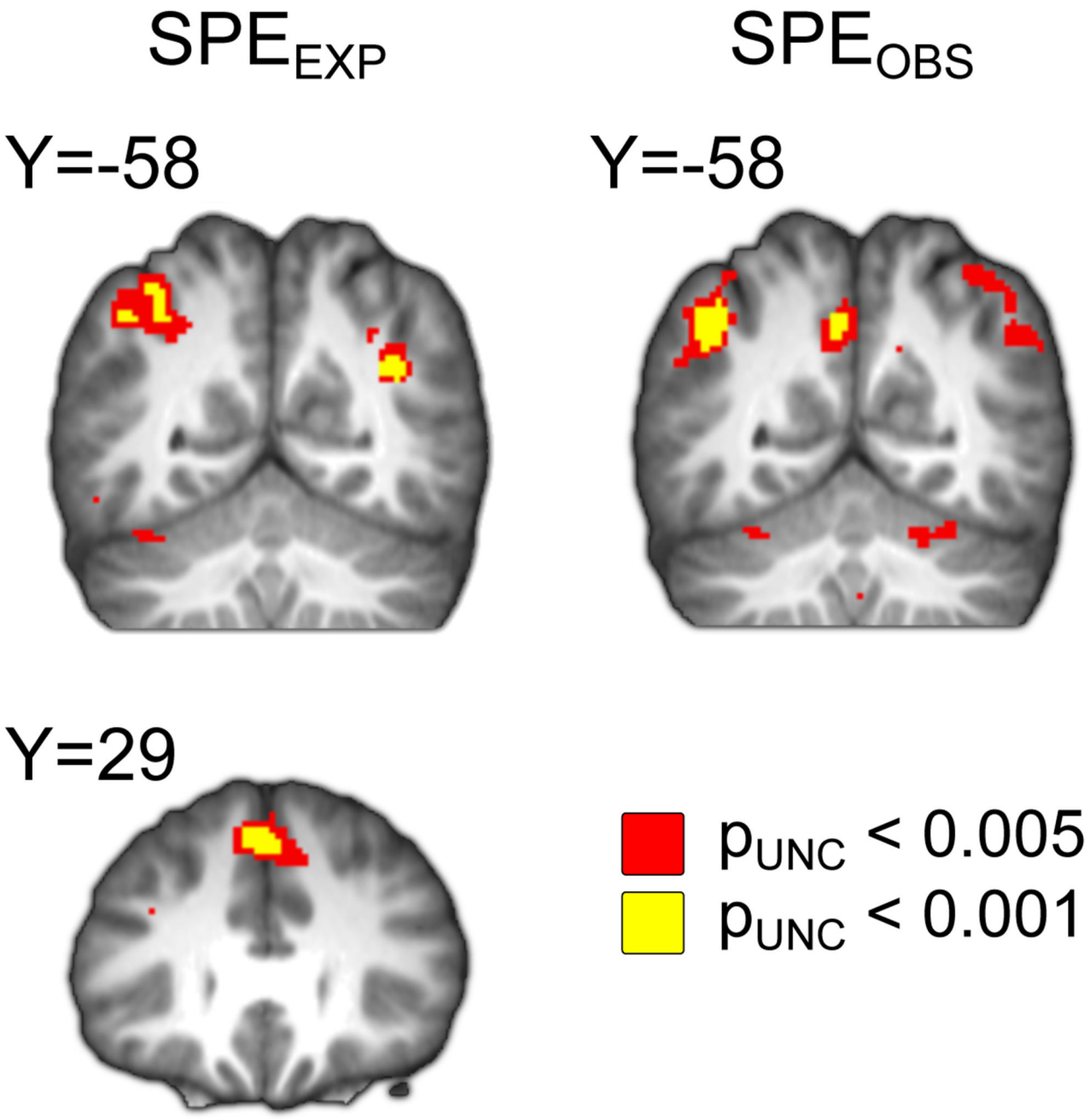

Figure 4. Effects of state prediction error.

We found effects of SPE_{EXP} on BOLD ($p_{FWE} < 0.05$, *WBC*) in left intraparietal sulcus and right dorsomedial prefrontal cortex, and effects of SPE_{OBS} in left intraparietal sulcus/inferior parietal lobule and left precuneus. A two-tailed *t*-test indicated that none of these clusters showed a significantly different effect of SPE_{EXP} than of SPE_{OBS}. SPMs are overlaid on the group mean normalized anatomical image and thresholded at $p_{UNC} < 0.005$ and $p_{UNC} < 0.001$ for the purpose of illustration.

Europe PMC Funders Author Manuscripts

Table 1

Peak co-ordinates of all significantly activated clusters, related to Figures 3,4, and 5; [1]$p_{FWE} < 0.05$ at cluster-level after whole-brain correction and a height threshold of $p_{UNC} < 0.005$; [2]$p_{FWE} < 0.05$ at peak-level after small volume correction and a height threshold of $p_{UNC} < 0.005$; [3]Survives correction for multiple comparisons with a learning rate of 0.05; [4]Survives correction for multiple comparisons with a learning rate of 0.25.

Contrast	Region	x	y	z
RPE_{EXP}	Posterior cingulate*	0	-38	32
	Dorsomedial prefrontal cortex*	-4	62	6
	Right ventral striatum**	10	10	-6
$RPE_{OBS} > RPE_{EXP}$	Right middle occipital gyrus*	22	-84	10
$RPE_{EXP} > RPE_{OBS}$	Left ventral striatum**	-6	10	-4
SPE_{OBS}	Left precuneus*	-4	-62	40
	Left intraparietal sulcus/inferior parietal lobule*	-44	-62	42
SPE_{EXP}	Left intraparietal sulcus*	-48	-68	28
	Right dorsomedial prefrontal cortex	4	34	42
$SPE_{OBS} > SPE_{EXP}$	Right middle temporal gyrus*	40	-72	12