Behavioral/Systems/Cognitive

Striatal Prediction Error Modulates Cortical Coupling

Hanneke E. M. den Ouden,[1,3] **Jean Daunizeau,**[1,4] **Jonathan Roiser,**[2] **Karl J. Friston,**[1] **and Klaas E. Stephan**[1,4]

[1]Wellcome Trust Centre for Neuroimaging, Institute of Neurology and [2]Institute of Cognitive Neuroscience, University College London, London, WC1E 6BT, United Kingdom, [3]Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behaviour, Centre for Cognitive Neuroimaging, 6525 EN, Nijmegen, The Netherlands, and [4]Laboratory for Social and Neural Systems Research, Institute for Empirical Research in Economics, University of Zurich, CH-8006 Zurich, Switzerland

Both perceptual inference and motor responses are shaped by learned probabilities. For example, stimulus-induced responses in sensory cortices and preparatory activity in premotor cortex reflect how (un)expected a stimulus is. This is in accordance with predictive coding accounts of brain function, which posit a fundamental role of prediction errors for learning and adaptive behavior. We used functional magnetic resonance imaging and recent advances in computational modeling to investigate how (failures of) learned predictions about visual stimuli influence subsequent motor responses. Healthy volunteers discriminated visual stimuli that were differentially predicted by auditory cues. Critically, the predictive strengths of cues varied over time, requiring subjects to continuously update estimates of stimulus probabilities. This online inference, modeled using a hierarchical Bayesian learner, was reflected behaviorally: speed and accuracy of motor responses increased significantly with predictability of the stimuli.

We used nonlinear dynamic causal modeling to demonstrate that striatal prediction errors are used to tune functional coupling in cortical networks during learning. Specifically, the degree of striatal trial-by-trial prediction error activity controls the efficacy of visuomotor connections and thus the influence of surprising stimuli on premotor activity. This finding substantially advances our understanding of striatal function and provides direct empirical evidence for formal learning theories that posit a central role for prediction error-dependent plasticity.

Introduction

One of the major reasons for the remarkable flexibility and adaptive repertoire of human behavior is that we construct and update estimates of conditional probabilities that describe uncertain causal relationships in the world. For example, human subjects can infer changing conditional probabilities among sensory events (Behrens et al., 2007; Brodersen et al., 2008), even when these probabilities are not relevant for overt behavior (den Ouden et al., 2009). Such learning of stimulus probabilities has been shown to be reflected by activity changes in visual (Summerfield et al., 2008; Summerfield and Koechlin, 2008), auditory (Pincze et al., 2002), and somatosensory (Akatsuka et al., 2007; Iannetti et al., 2008) areas. The general principle that emerges from these studies is that sensory responses increase with the size of prediction error, i.e., the less expected (or more surprising) a stimulus, the greater the response. This is in accordance with current theoretical accounts of brain function, such as predictive coding and the free-energy principle (Rao and Ballard, 1999; Friston, 2005; Friston and Stephan, 2007; Friston and Kiebel, 2009), which posit a fundamental role of prediction errors for adaptive behavior and learning.

Efficient learning of probabilities can be used to form predictions that guide motor behavior. For example, once the predictive strength of a cue has been learned, the premotor cortex shows preparatory activity (Tanji and Evarts, 1976; Wise and Mauritz, 1985; Crammond and Kalaska, 2000) and reaction times (RT) decrease (Requin and Granjon, 1969; Strange et al., 2005; Bestmann et al., 2008).

However, the neurobiological mechanisms that underlie adaptive changes in motor behavior when predictions fail are not fully understood. According to the free-energy principle, any prediction error should induce learning through synaptic plasticity, reconfiguring connection strengths such that prediction error is minimized at both sensory and motor levels, thus optimizing perceptual inference and motor actions (Friston and Stephan, 2007). In line with this view, a recent electrophysiological study by Bestmann et al. (2008) suggested that information about prediction errors is ". . . continuously channelled into motor regions to control the excitability of expected motor outputs." In this study, we provide direct empirical evidence for this hypothesis, exploiting recent advances in computational models of learning (Behrens et al., 2007) and nonlinear dynamic causal models (DCMs) of functional magnetic resonance imaging (fMRI) data (Stephan et al., 2008). In particular, we link the physiological mechanisms proposed by the free-energy principle (and other formal theories of learning and decision making), i.e., prediction error-dependent changes in connectivity, to a large body of literature that has described responses in the striatum correlated with prediction error (McClure et al., 2003; O'Doherty et al., 2003, 2004; Corlett et al., 2004; Seymour et al., 2004; Jensen et al., 2007; Menon et al., 2007). Specifically, we show that the observed learning-

Received Sept. 4, 2009; revised Dec. 1, 2009; accepted Dec. 29, 2009.

This work was supported by the Wellcome Trust (H.E.M.d.O., K.J.F., K.E.S.), the University Research Priority Program Foundations of Human Social Behaviour and the NCCR "Neural Plasticity" at the University of Zurich (K.E.S.), and the Swiss Initiative in Systems Biology SystemsX.ch (J.D., K.E.S.). We thank Tim Behrens, Nathaniel Daw, and Quentin Huys for helpful discussions.

Correspondence should be addressed to Hanneke den Ouden, Donders Centre for Cognitive Neuroimaging, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands. E-mail: h.denouden@donders.ru.nl.

DOI:10.1523/JNEUROSCI.4458-09.2010

0270-6474/10/303210-10$15.00/0

dependent changes in blood oxygenation level-dependent (BOLD) activity are compatible with a mechanistic model in which the connection strengths between visual and motor regions are modulated by prediction error-related activity in the striatum.

Materials and Methods

Participants

Twenty healthy right-handed volunteers, 24.4 ± 2.1 years of age (mean ± SD, 10 female) took part in this study. The participants had no history of psychiatric or neurological disorders. Written informed consent was obtained from all volunteers before participation. The study was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee.

Experimental design

Associative learning task. To investigate the mechanisms underlying adaptive motor behavior and the role of striatal and sensory prediction errors, we used a simple audiovisual associative learning task in which the participants were presented with auditory cues (high or low beeps) that differentially predicted upcoming visual target stimuli (Fig. 1*A*). To ensure attention and assess learning, participants were instructed to respond as quickly as possible by button press (right middle and index finger, counterbalanced across subjects), reporting whether the target stimulus was a face (F) or a house (H). The associative contingencies of the auditory and visual stimuli changed over the course of the experiment (Fig. 1*B*). To ensure that participants' responses were not biased by learned expectations about the relative frequencies of the visual stimuli, we designed the sequence of changes in probabilities such that the marginal probabilities of faces and houses were identical, as described in more detail below. Subjects were instructed that the relation between auditory and visual stimuli was probabilistic, that these probabilistic relations would change unpredictably in time. They were not informed about the nature of the probabilistic relations or about the temporal intervals over which they changed.

Timing and stimuli

On each trial, one of two auditory cue conditioned stimuli (CS_1 and CS_2) was followed by a visual target stimulus (Fig. 1*A*). To prevent anticipatory responses or guesses, both the intertrial interval (2000 ± 650 ms) and visual stimulus onset latency (150 ± 50 ms) were jittered randomly (Fig. 1*A*). Auditory stimuli were presented binaurally for 300 ms. The auditory stimuli were matched for perceived loudness under scanning conditions as described previously (den Ouden et al., 2009). The frequencies of the auditory stimuli used in this experiment were 1125 and 500 Hz, and the adapted volume of the high tone was 98 ± 4.1% (mean ± SD) with respect to the low tone. The visual stimuli were presented centrally for 150 ms to prevent saccades, and subjects were required to fixate a central cross throughout the experiment. The visual stimuli consisted of eight pictures of neutral facial expressions drawn from the Ekman Series of Facial Affect (Ekman and Friesen, 1976) and eight pictures of houses, matched for overall luminance and presented on a gray background. Stimuli were presented using the software package Cogent (www.vislab.ucl.ac.uk/Cogent).

Cue–outcome contingencies

The two tones differentially predicted the identity of the visual target stimulus, and these contingencies changed over the course of the task (Fig. 1*B*). Because each CS was followed by one of two stimuli (F or H), the probability of one visual stimulus, given a particular auditory CS, was one minus the probability of the other visual stimulus:

$$p(\mathrm{F}|\mathrm{CS}_i) = 1 - p(\mathrm{H}|\mathrm{CS}_i) \quad [i \in \{1, 2\}]. \tag{1}$$

To ensure that participants' responses were not biased by learned expectations (e.g., about the relative frequencies of the visual stimuli), the marginal probabilities of faces and houses were identical at any point in time. This was achieved because, first, the probability of one visual outcome given CS_1 was the same as the probability of the other visual outcome given CS_2 (compare with Fig. 1*B*):

$$p(\mathrm{F}|\mathrm{CS}_1) = p(\mathrm{H}|\mathrm{CS}_2). \tag{2}$$

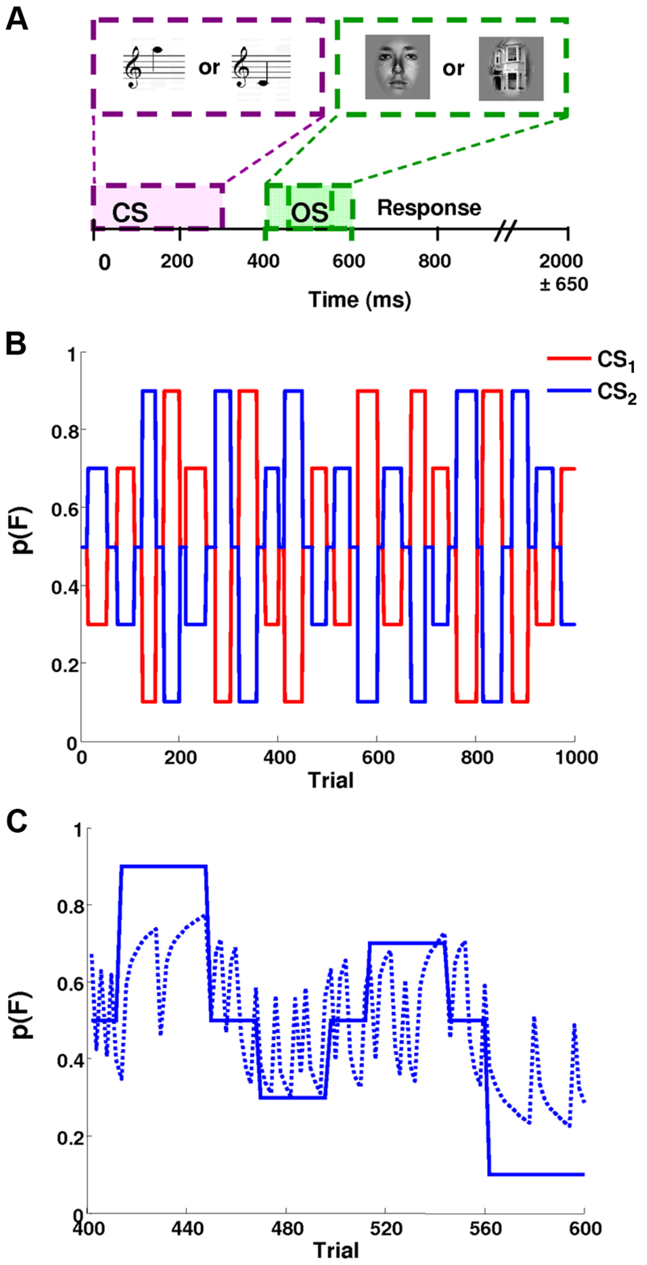

Figure 1. ***A***, Timeline for a single trial. At trial onset, the auditory cue stimulus (CS) was presented for 300 ms. The visual outcome stimulus (OS) lasted for 150 ms and was presented 150 ± 50 ms after the CS. The intertrial interval lasted for 2000 ± 650 ms on average. ***B***, Temporal evolution of the probability of a face occurring, *p*(F), given either CS. Note that the probability of a house being presented is simply the mirror image of this sequence. ***C***, The posterior mean of *p*(F|CS) as estimated by the Bayesian learner (dashed line) tracks the underlying blocked probabilities (solid line). Note that the blocked probabilities are CS_2 zoomed in from ***B*** (trials 400 – 600, session 3). Because blocks of stable probabilities are short, however, the estimated probabilities never quite reach their true values during a given block. Note that the estimates change rapidly at block transitions. When an unexpected stimulus occurs, the estimates briefly move toward $p = 0.5$. Note that, for clarity, we only show a single session (session 3) here.

Second, each block contained equal numbers of randomly intermixed CS_1 and CS_2 trials. With these two manipulations, we ensured that, on any given trial, before the CS was presented, the a priori probability of a face (or house) occurring was always 50%. Thus, any expectations about the visual stimulus could depend only on the auditory stimulus.

Each subject completed five sessions of 200 trials each. In each session, the predictive strengths of the two CS types changed pseudorandomly over time, taking one of five different discrete levels of predictive association. Specifically, the cue could be (1) strongly predictive ($p = 0.9$), (2) predictive ($p = 0.7$), (3) nonpredictive ($p = 0.5$), (4) antipredictive ($p = 0.3$), and (5) strongly antipredictive ($p = 0.1$) of the visual stimulus. Each predictive level was presented as a block of stimuli once per scanning session. Predictive block lengths varied between 14 and 20 trials per CS type, so that participants could not predict exactly when a change in contingencies would occur. Furthermore, to avoid instantaneous and complete reversals of the contingencies, blocks containing predictive cues alternated with short blocks (6–10 trials) containing nonpredictive cues (i.e., $p = 0.5$) in order.

fMRI data acquisition

A 3 T head scanner (Allegra Magnetom; Siemens) was used to acquire a T1-weighted fast-field echo structural image and multislice T2*-weighted echo-planar volumes with BOLD contrast (repetition time, 2.73 s; echo time, 30 ms). Before the functional scans, a B_0 field map was acquired using a gradient echo field map sequence. Functional data were acquired in five scanning sessions lasting ~8 min each. One hundred eighty-nine volumes were acquired per session (945 scans in total per subject). The first six volumes of each session were discarded to allow for T1 equilibrium effects. Each functional brain volume comprised 42 axial slices with 2 mm thickness and a 1 mm interslice gap and an in-plane resolution of 3 × 3 mm. The field of view was chosen to cover the whole brain, except for the brainstem.

Data analysis

Behavioral data analysis

First, the data were screened for outliers in reaction times, and responses faster than 150 ms were excluded. We then tested whether the distributions of RT and response speeds (RS) (i.e., inverse reaction times) showed significant deviations from normality using a Kolmogorov–Smirnov test. Because RS, but not RT, were well described by a Gaussian distribution, the former were entered into a repeated-measures ANOVA with outcome probability (five levels: 0.9, 0.7, 0.5, 0.3, and 0.1), CS type (CS_1, CS_2), and outcome type (F, H) as within-subjects factors. The Greenhouse–Geisser correction was used when significant nonsphericity was detected. The analogous analysis was performed on error rates.

Bayesian learning model

The linear model based on the true probabilities described above (henceforth referred to as the "categorical model") explained a significant proportion of variance in reaction times and errors and as such provided a reasonable model of the behavioral data. However, it makes the unrealistic assumption that the participants have instantaneous and precise knowledge of the true probabilities that generated the stimulus sequence. In reality, the participants had to estimate these unknown probabilities trial by trial from the observed stimulus sequence. Thus, we aimed to find a more realistic model reflecting how subjects' continuously updated their estimates of the associations, based on past observations. Clearly, there are numerous potential models of such a process (Dayan and Daw, 2008; Dayan and Niv, 2008). Here we focused on a generic Bayesian learner model that accounts for both trial-by-trial updates of probability estimates and learning about the volatility of the environment (Behrens et al., 2007).

Previous studies have shown that many human perceptual and motor processes approximate the behavior of an ideal Bayesian learner (Kersten et al., 2004; Körding and Wolpert, 2006). Bayesian learners continually update their estimates of hidden contingencies by combining previous information from past experience with current observations in the present. In standard Bayesian learner models, the learning rate, and thus the relative influence of past versus current observations on the estimates, is constant. This, however, is not optimal when the underlying probabilistic associations are changing in an unknown and irregular manner, as in the task used in this study. In such an environment, an optimal learner would not only estimate the probabilities but also their instability in time (i.e., volatility) and would increase the weight of current observations, relative to past experience, with increasing volatility. To model the ongoing estimation of associative cue–outcome relationships based on observed outcomes, we used a hierarchical Bayesian learning model developed by Behrens et al. (2007). Given a series of observed events, this model estimates, at any given point in time, the posterior probability density function (PDF) of both the probabilistic associations and the volatility of the environment (see Fig. 2). Here, we adopted this model (for details of our implementation, see supplemental data, available at www.jneurosci.org as supplemental material) and used the posterior mean of the PDFs as estimates of the probability and volatility. To verify that the probability estimates of this Bayesian model were better linear predictors of the behavioral RS than the true probabilities from which the stimulus sequence was generated, we used Bayesian model selection (BMS) (see below). Given the clear superiority of probability estimates from the Bayesian model in explaining the behavioral data (see Results), they were used in the subsequent analyses of the fMRI data.

Based on suggestions by our reviewers, we tested three additional models that could potentially explain the behavioral data. First, we tested a standard Rescorla–Wagner model; in the reinforcement learning literature, this is referred to as a "model-free" approach because it has no knowledge about the environment or task structure. Furthermore, we evaluated two variants of a hidden Markov model that explicitly reflected task structure by representing transitions between the five different association levels (a "model-based" approach). Bayesian model comparison showed that the Bayesian learner model proved to be superior to any of these models. The main text of this manuscript therefore focuses on comparing the Bayesian learner to the categorical model based on the true probabilities. Details of the additional models and their comparison are included in the supplemental data (available at www.jneurosci.org as supplemental material). In particular, supplemental Figure S2 (available at www.jneurosci.org as supplemental material) juxtaposes the predictions from the different computational models about trial-by-trial estimates of cue strength, and supplemental Table S1 (available at www.jneurosci.org as supplemental material) lists the results of model comparison.

Bayesian model selection

When comparing different models for observed data, it is critical that the decision is not only based on the relative fit but also on the relative complexity of the competing models (Pitt and Myung, 2002). BMS provides a principled foundation for comparing competing models of different complexity (Penny et al., 2004). We used a novel hierarchical method for BMS that allows for group-level random-effects inference about the relative goodness of multiple competing models (Stephan et al., 2009). First, as described in the supplemental data (available at www.jneurosci.org as supplemental material), for all models considered, we computed the evidence $p(y|m)$, i.e., the probability of the data y being generated by model m, for each subject. The model evidence balances fit and complexity, enabling one to compare non-nested models with different levels of complexity. For the linear models applied to the behavioral data, there is an analytic expression for the model evidence (for details, see supplemental data, available at www.jneurosci.org as supplemental material). For the nonlinear dynamic causal models of the fMRI data described below, we used the negative free-energy approximation to the log model evidence (Friston et al., 2007; Stephan et al., 2007b).

Subsequently, the models were compared at the group level, using random-effects BMS (Stephan et al., 2009). This method has been shown to be considerably more robust than either the conventional fixed-effects analysis using the group Bayes factor (Stephan et al., 2007b) or frequentist tests applied to model evidences, especially in the presence of outliers (Stephan et al., 2009). It uses variational Bayes to infer the posterior density of the models per se. One can then derive the exceedance probability, φ_k, i.e., the probability that a particular model k is more likely than any other model considered, given the group data. Exceedance probabilities are particularly intuitive when comparing two models, as in our analysis of the behavioral data (see Fig. 2*D*).

Functional neuroimaging analysis

fMRI data were analyzed using the SPM5 software package (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). The 915 echo planar images from each subject were corrected for geometric distortions caused by susceptibility-induced field inhomogeneities. A combined approach was used that corrects for both static distortions and changes in these distortions attributable to head motion (Andersson et al., 2001; Hutton et al., 2002). The static distortions were calculated for each subject by acquiring a B_0 field map and processing it using the FieldMap toolbox implemented in SPM5 (Hutton et al., 2004). The images were then realigned unwarped, slice-time corrected, and coregistered to the subject's own structural scan. The structural image was processed using a unified segmentation procedure combining segmentation, bias correction, and spatial normalization (Ashburner and Friston, 2005); the same normalization parameters were then used to normalize the echo planar images. Finally, the echo planar images were smoothed with a Gaussian kernel of 8 mm full-width half-maximum and resampled to 3 × 3 × 3 mm voxels.

The data were then modeled voxelwise, using the general linear model (GLM) for each of the 20 participants. In the GLM, correct and error trials were modeled as separate events. For correct trials, face and house trials were modeled as the two main conditions of interest. These were collapsed across the two different CS types (high and low tones), because the predictive strengths of the two CSs were counterbalanced over time and thus no differential effects were to be expected (this was confirmed by analysis of the behavioral data). Condition-specific effects were modeled in an event-related manner, convolving a sequence of delta functions with a canonical hemodynamic response function. The probability estimates from the Bayesian learner as well as the subject-specific response speeds were included as first-order parametric modulators of face and house trials such that the delta functions representing the presence of a face were modulated by the trial-specific probability estimate that a face should have occurred on this trial (equivalently for house trials). We also included the volatility estimates from the Bayesian learner as parametric modulators (orthogonalized with respect to the probability estimates). Finally, the six motion parameter vectors from the realignment procedure were included as regressors of no interest to account for variance caused by head motion.

After computing subject-specific contrast images of interest, random-effects group analyses across all 20 subjects were performed (Friston et al., 2005), using one-sided one-sample *t* tests and testing for both positive and negative BOLD responses. We report any responses that survived whole-brain correction at the cluster level ($p < 0.05$), with a voxel-level threshold of $p < 0.001$. For anatomically constrained a priori hypotheses concerning fusiform face area (FFA) and parahippocampal place area (PPA) (for stimulus-specific effects), putamen (for prediction errors), and anterior cingulate cortex (ACC) (for volatility; see supplemental material, available at www.jneurosci.org), we report activations that survived correction at the cluster level ($p < 0.05$; with $p < 0.001$ voxel-level threshold) within the region of interest. For the putamen and ACC, search volumes were generated using the PickAtlas toolbox using the AAL atlas (Maldjian et al., 2003); for stimulus-specific visual areas, we used in-built localizer contrasts that were orthogonal to the other contrasts of interest (see below).

First, we assessed the main effect of probability, that is, in which brain regions the hemodynamic response reflected the probability of the stimulus occurring, independently of which stimulus it was. We tested for both BOLD responses that increased with the likelihood of the outcome and responses that increased the less likely (or more surprising) the outcome was. In other words, these contrasts tested for stimulus-independent responses that reflected predicted or surprising outcomes, respectively. Given the results from our previous study (den Ouden et al., 2009), our a priori hypothesis was that the response in the putamen would correlate positively with prediction error, i.e., negatively with the probability of the observed outcome.

Second, we tested for stimulus × probability interactions, that is, prediction error-dependent responses that differed between faces and houses. Our a priori hypothesis was that responses in stimulus-specific areas would scale inversely with the probability of the presented stimulus (cf. Friston, 2005; Summerfield et al., 2008). In other words, responses of the FFA to face stimuli should decrease when they were more probable, and responses of the PPA to houses should decrease with the probability of a house being presented. This can be regarded equivalently as testing for prediction error-dependent increases in the response of category-specific areas. To accommodate intersubject variability in the exact location of FFA and PPA, we identified these regions for each subject separately, using an embedded orthogonal localizer contrast (i.e., the main effect of faces versus houses and vice versa) (Friston et al., 2006). We verified, using the whitened design matrix, that localizing and test contrasts were indeed orthogonal (cf. Kriegeskorte et al., 2009). In each subject, the individual maximum within 8 mm of the group maximum of face- and house-specific responses (Table 1) was determined. Subsequently, given these voxels with individually maximal stimulus specificity, we tested for (orthogonal) stimulus × probability interactions by entering the parameter estimates of regressors encoding trial-by-trial stimulus probability estimates into two-tailed paired-sample *t* tests. In other words, this procedure tested whether face- and house-specific responses in FFA and PPA, respectively, were modulated by the trial-by-trial probability estimate of a face or a house occurring.

Table 1. Montreal Neurological Institute coordinates and *Z* scores for significantly activated regions

Region	*x*	*y*	*z*	*Z* score
Prediction error effects: negative correlation with *p*(F) and *p*(H)				
Motor areas				
Left precentral gyrus (dorsal premotor cortex)*	−18	−18	60	4.13
Right intraparietal sulcus*	42	−33	39	4.02
Right superior parietal gyrus*	15	−60	63	4.16
Striatum				
Right putamen**	27	3	6	3.42
Left putamen**	−24	15	3	3.39
Probability effects: positive correlation with *p*(F) and *p*(H)				
No significant activations				

*$p < 0.05$, cluster-level corrected across the whole brain; **$p < 0.05$, cluster-level corrected for a priori region of interest.

Nonlinear dynamic causal models

Numerous studies have demonstrated previously that hemodynamic responses in the striatum reflect prediction errors (McClure et al., 2003; O'Doherty et al., 2004; Pessiglione et al., 2006; Jensen et al., 2007; den Ouden et al., 2009). According to theoretical models of learning, the size of prediction errors should control the strength of synaptic connections encoding stimulus–stimulus and stimulus–response links (McLaren et al., 1989; Schultz and Dickinson, 2000; Friston, 2005). In this study, we tested this notion directly by modeling how activity in the putamen gated the influence of visual areas onto the dorsal premotor cortex (PMd) (see Fig. 5). We expected that increased BOLD responses in the putamen, induced by a surprising face, should increase the strength of the FFA → PMd connection, thus enhancing the influence of face information on PMd activity and facilitating an update of the motor plan. This type of analysis, which requires the assessment of modulatory (second-order) effects on connectivity, has become possible with the recent introduction of nonlinear DCMs (Stephan et al., 2008).

In nonlinear DCMs, the hidden neural dynamics (i.e., not directly observed by fMRI) are modeled by the following equation:

$$\frac{dx}{dt} = \left(A + \sum_{i=1}^{m} u_i B^{(i)} + \sum_{j=1}^{n} x_j D^{(j)} \right) x + Cu. \tag{3}$$

Here, u are the experimentally controlled inputs, the A matrix represents the endogenous (context-independent or fixed) connection strengths between the modeled regions, the matrices $B^{(i)}$ represent the modulation of these connections induced by the *i*th input u_i as an additive change, and the C matrix represents the influence of direct (exogenous) inputs to the system. Finally, the $D^{(j)}$ matrices encode how connection strengths are modulated or gated by activity in area *j* (for details, see Stephan et al., 2008).

DCM specification. Based on our SPM results, we constructed a nonlinear DCM including the right putamen, PPA and FFA, and the left

PMd. As shown in Figure 3 and Table 1, several other areas showed a prediction error-dependent response and may also be involved in the visuomotor information transfer. Therefore, the model we chose, including the above four regions, should be regarded as the most parsimonious model that enabled us to test whether prediction error activity in the putamen gated visuomotor connections. Although putamen, FFA, and PPA showed peak activations in the right hemisphere, we included the left premotor cortex because participants responded with their right hand.

We constructed and compared several alternative models. The basic architecture, shown in Figure 4*A*, included connections from FFA and PPA to the PMd and modulations of these connections by activity in the putamen, which was driven by the trial-by-trial probability estimates provided by the Bayesian learning model. The connectivity layout of our basic model was guided by anatomical and physiological data. For example, in the Macaque monkey, both ventral and dorsal parts of the lateral premotor cortex contain substantial numbers of neurons that respond to visual stimuli (Fogassi et al., 1999), including symbolic action-selection cues (Yamagata et al., 2009). These responses may be mediated by direct anatomical connections from inferotemporal to ventral premotor cortex (Webster et al., 1994; Gerbella et al., 2010) or by polysynaptic connections from various visual areas to dorsal premotor cortex, which are relayed via parietal and temporal areas (Matelli et al., 1998; Luppino et al., 2001; Tanné-Gariépy et al., 2002; Gamberini et al., 2009). The connectivity structure of this DCM was subsequently optimized systematically by BMS; the optimal model was found to include reciprocal connections between FFA, PPA, and PMd (for a graphical representation of all models tested, see supplemental Fig. S1, available at www.jneurosci.org as supplemental material).

After the endogenous connections had been optimized, we conducted a final and critical model comparison. Because the putamen and the PMd showed similar prediction error-related responses (see Fig. 3), we wanted to establish the specificity of our model and demonstrate that putamen activity gated visuomotor connections, instead of PMd gating connections between visual areas and putamen. We therefore compared our model with one in which the roles of the PMd and the putamen were reversed (see Fig. 4*C*).

Time series extraction. Because the exact locations of activation maxima varied across participants, we used subject-specific anatomical and functional constraints in selecting regional time series (cf. Stephan et al., 2007a). In brief, a regional time series was extracted if (1) it passed a threshold of $p < 0.05$ (uncorrected) in the respective contrast of the GLM analysis and if (2) it was located within a certain distance from the group maximum. For FFA and PPA (identified by the contrast testing for main effect of faces vs houses, F > H and H > F, respectively), the individual maxima were required to be within 8 mm of the group maximum. For the putamen and PMd (identified by the contrast testing for a negative main effect of probability), the individual maxima were required to be within 16 mm of the group maximum (PMd) and within the putamen as defined by the participant's own anatomical scan. To summarize the regional time series, we computed the first eigenvector across all suprathreshold voxels within 4 mm of the selected maximum. After this procedure, we were able to extract time series for all four areas in 15 of 20 participants. We could not obtain a putamen time series in three participants and a PMd time series in two participants because of failure to meet the anatomical and functional criteria above. Because we could not specify the complete model in these participants, they were excluded from the DCM analysis.

Results

Behavioral data

Subjects responded correctly on 91 ± 0.8% (mean ± SE) of trials, on 5% of the trials they gave the wrong response or pressed multiple buttons, and on the remaining 4% of trials they did not respond before the end of the trial.

Subjects responded faster (Fig. 2*A*) and more accurately (Fig. 2*B*) to more likely stimuli, indicating that they successfully tracked the changing contingencies. The difference in average reaction time between unexpected ($p = 0.1$) and expected ($p = 0.9$) outcomes across subjects was 32 ms (Fig. 2*A*). For formal hypothesis testing, we used inverse reaction times, i.e., RS, because these were more normally distributed than reaction times (cf. Carpenter and Williams, 1995). Repeated-measures ANOVA showed a significant relationship between both RS ($F_{(2.4,45.4)} = 43.9$; $p < 0.001$) and error rates ($F_{(1.5,334.0)} = 12.52$; $p < 0.001$) and stimulus probability.

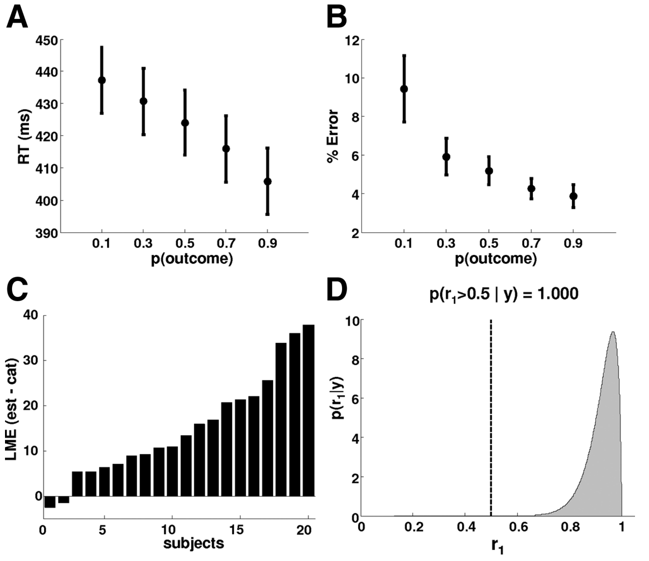

Figure 2. RTs (***A***) and percentage of errors as a function of outcome probability (***B***) (mean ± SE). Correct trials were averaged within each level of probability and collapsed across CS and visual outcome type (F/H). Subjects speed up and make fewer errors the higher the probability of the outcome. ***C***, Subject-specific differences in log model evidence (LME) for using the trial-by-trial probability estimates from the Bayesian model versus the true probabilities as linear predictors for behavioral measured response speeds. In all but two subjects, there is far greater evidence for the Bayesian model. ***D***, The Dirichlet density describing the probability of model m_1 (based on the probability estimates from the Bayesian learning model) relative to the alternative model m_2 (based on the true, blocked probabilities), given the measured response speeds across the group. The shaded area represents the exceedance probability of m_1 being a more likely model than m_2. This exceedance probability of $\varphi_1 = 100.0\%$ was strongly favoring m_1 as a more likely model than m_2.

As discussed in Materials and Methods, this ANOVA, although explaining a significant amount of variance, cannot be a realistic representation of the subjects' estimates of cue–outcome association strengths. We therefore used a hierarchical Bayesian learning model that estimates, from the observed cue–outcome combinations, the probabilistic associations given a series of observed events (Fig. 1*C*), and we tested whether the estimates of this Bayesian model were better linear predictors of the behavioral RS than the true probabilities from which the stimulus sequence was generated. The distribution of the log evidences across subjects (Fig. 2*C*) and random-effects BMS at the group level showed that the Bayesian learning model was unmistakably superior: the exceedance probability that the Bayesian learning model was the more likely model was almost 100% (Fig. 2*D*).

Following suggestions by our reviewers, we also compared the Bayesian learning model to a Rescorla–Wagner learning model and to two variants of a hidden Markov model reflecting the underlying task structure. The Bayesian model was found to be clearly superior to each of these additional models (for details, see supplemental data, available at www.jneurosci.org as supplemental material). Given this result, the trial-by-trial estimates from the Bayesian model were used in the subsequent analyses of the fMRI data.

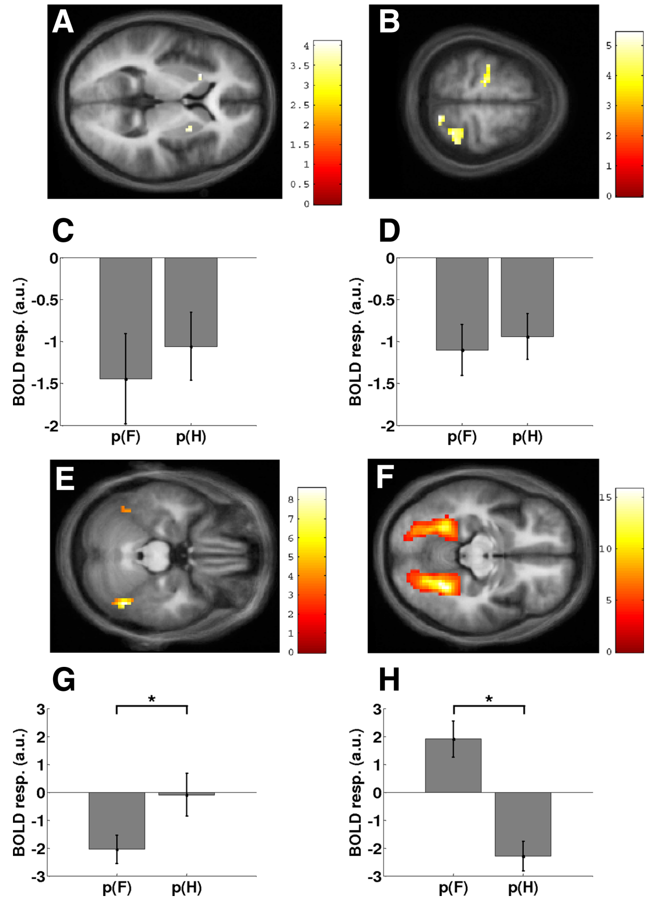

Figure 3. All parameter estimates show mean ± SE across all subjects, and all activations are displayed on the average anatomical scan. ***B*** and ***D–F*** show the results of a whole-brain analysis, and ***A***, ***C***, ***G***, and ***H*** show the results from region-of-interest analyses. ***A***, Effect of prediction error in the anterior putamen bilaterally. ***C***, Parameter estimates from the putamen showing the negative dependency on both p(F) and the p(H). ***B***, Effects of prediction error in PMd and the parietal cortex. a.u., Arbitrary units. ***D***, Parameter estimates for the left PMd, showing the same prediction error-dependent effect as the putamen. ***E***, Main effect of F > H in the right FFA, also showing the left FFA activation (see supplemental data, available at www.jneurosci.org as supplemental material). ***F***, Main effect of H > F in the bilateral PPA. ***G***, Parameter estimates of the modulatory effect of stimulus probabilities (from the individual maxima for the orthogonal F > H contrast in the FFA). There was a pronounced negative modulation of FFA responses to faces by the trial-by-trial probability estimates for faces ($\beta = -2.05 \pm 0.52$). In contrast, the modulation of FFA responses to houses by the trial-by-trial probability estimates for houses was marginal ($\beta = -0.09 \pm 0.78$). This difference was significant ($*p = 0.037$). ***H***, Parameter estimates of the modulatory effect of stimulus probabilities across subjects (from the individual maxima for the orthogonal H > F contrast in the PPA). PPA responses to houses showed a strongly negative modulation by the trial-by-trial probability estimates for houses ($\beta = -2.29 \pm 0.54$). In contrast, PPA responses to faces were positively modulated by the trial-by-trial probability estimates for faces ($\beta = 1.91 \pm 0.67$). This difference was significant ($*p = 0.00005$).

fMRI data

The main results of our SPM analysis are summarized graphically in Figure 3.

Regional responses reflecting stimulus-independent prediction error

Responses in the bilateral putamen were negatively correlated with the probability of the visual stimulus, regardless whether this stimulus was a face or a house (Table 1; Fig. 3*A*,*C*). In other words, the BOLD response in the putamen increased the more surprising the outcome was. Other areas that showed this type of response included the left PMd, right intraparietal sulcus, and right superior parietal gyrus (Table 1; Fig. 3*B*,*D*).

Prediction error-related responses in stimulus-specific areas

The factorial design provided an in-built localizer contrast for defining the FFA and PPA functionally in each participant (see Materials and Methods). The individual peak voxels in right FFA and right PPA that showed maximally selective face and house responses, respectively, also showed an (orthogonal) stimulus × probability interaction: in the FFA, there was a pronounced negative modulation of response to faces by the trial-by-trial probability estimates for faces ($\beta = -2.05 \pm 0.52$, mean ± SE). In other words, FFA responses to faces increased when the occurrence of a face was surprising (i.e., a large prediction error). In contrast, the modulation of FFA responses to houses by the trial-by-trial probability estimates for houses was negligible ($\beta = -0.09 \pm 0.78$) (Fig. 3*G*). The difference in regression slopes between the two conditions (i.e., the interaction) was significant ($t_{(19)} = 2.25; p = 0.037$).

The response in the PPA in response to houses showed a strongly negative modulation by the trial-by-trial probability estimates for houses ($\beta = -2.29 \pm 0.54$) (Fig. 3*H*), that is, in analogy to the FFA results, PPA responses to houses increased the higher the prediction error, i.e., the more surprising the presentation of a house was. However, unlike the FFA analysis, PPA responses to faces were positively modulated by the trial-by-trial probability estimates for faces ($\beta = 1.91 \pm 0.67$); this corresponds to a decrease in activity the more surprising the presentation of a face was. As for FFA, this interaction was significant ($t_{(19)} = 5.22; p = 0.00005$).

In summary, responses of PPA and FFA to their preferred stimuli were strongly modulated by prediction error, and this modulation was significantly higher than for their nonpreferred stimuli.

Nonlinear DCMs

In line with numerous previous studies (McClure et al., 2003; O'Doherty et al., 2004; Pessiglione et al., 2006; Jensen et al., 2007; den Ouden et al., 2009), the BOLD response in the putamen reflected prediction errors. According to theoretical models of learning, the size of prediction errors should control the strength of stimulus–stimulus and stimulus–response links and thus connection strength (McLaren et al., 1989; Montague et al., 1996, 2004; Schultz and Dickinson, 2000; Friston, 2005). We tested this notion directly by modeling how activity in the putamen gated the information flow from visual areas to the PMd (Fig. 4). First, the fixed connections were optimized; the optimal model included full reciprocal connectivity between PPA, FFA, and PMd but no direct connections from either visual areas or PMd to the putamen (Fig. 4*B*). The exceedance probability for this model was $\varphi_4 = 0.44$, surpassing the exceedance probabilities of the other tested models (which ranged from 0.01 to 0.28) (for details, see supplemental data, available at www.jneurosci.org as supplemental material).

Once the most likely pattern of connections among the areas was established, we constructed a model to verify the specificity of the modulatory influence exerted by the putamen, given that the putamen and the PMd showed similar prediction error-related responses (Fig. 3*C*,*D*). BMS showed that the model in which the roles of the putamen and premotor cortex were reversed (m_{pm}, with PMd as source of modulatory effects) was clearly inferior to the original model (m_{pt}, with the putamen as source of modulatory effects), with an exceedance probability of 99% in favor of the latter (Fig. 4*D*). In the optimal model m_{pt} (Fig. 4*B*), the parameter estimates reflecting gating effects of putamen activity on visuomotor connections were consistently positive and significant across subjects [PPA → PMd: $d = 0.01 \pm 0.003$ (mean ± SE), $t_{(14)} = 2.97, p = 0.010$; FFA → PMd: $d = 0.011 \pm 0.004$, $t_{(14)} = 2.71$, $p = 0.017$]. Therefore, in accordance with our initial hypothesis, prediction error-related activity in the putamen significantly modulated the strength of visuomotor connections.

Discussion

In this fMRI study, we used an audiovisual associative learning paradigm, in which we ensured that any expectations about the upcoming visual stimuli were entirely conditional on the auditory cues. Critically, the predictive strengths of cues were unknown and varied over time, requiring subjects to continuously update their estimates of cue–stimulus associations and thereby maximizing demands on adaptive changes in network connectivity. We modeled this online inference process using a Bayesian learner model that generated trial-by-trial probability estimates. These were used as predictor variables in the analysis of behavioral and fMRI data. Behaviorally, speed and accuracy of motor responses increased significantly with trial-by-trial predictability of visual stimuli (Fig. 2). Analysis of the fMRI data showed that responses in FFA and PPA reflected trial-by-trial prediction errors that were specific for their preferred stimulus (Fig. 3*C*,*D*). In contrast, both the putamen and dorsal premotor cortex represented stimulus-independent prediction errors (Fig. 3*A*,*B*). Using nonlinear DCMs, we found that prediction error responses in the putamen modulated the strength of connections from FFA and PPA to premotor cortex.

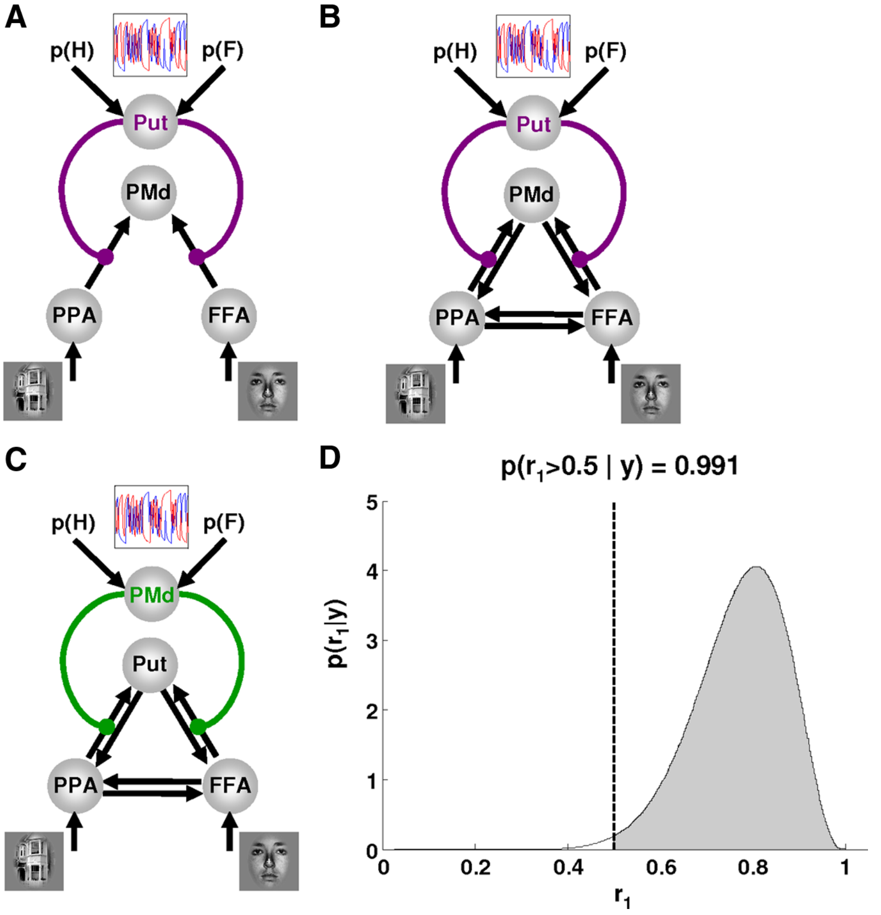

Figure 4. ***A***, A basic DCM for investigating modulation of visuomotor connections by prediction error-related activity in the putamen. ***B***, The optimal DCM (model m_{pt}), resulting from a systematic model search procedure, included full connectivity between the PMd, PPA, and FFA. Activity in the putamen significantly enhanced the connections from the PPA/FFA to the premotor cortex ($p = 0.010$ and $p = 0.017$, respectively). ***C***, Alternative DCM (model m_{pm}) in which the roles of the putamen and the PMd were swapped. ***D***, The Dirichlet density describing the probability of the "putamen" model m_{pt} relative to the alternative "premotor" model m_{pm}, given the measured fMRI data across the group. The shaded area represents the exceedance probability of m_{pt} being a more likely model than m_{pm}. This exceedance probability of $\varphi_1 = 99.1\%$ was strongly favoring m_{pt} as a more likely model than m_{pm}.

Our fMRI data analysis showed a double dissociation of responses in the FFA and PPA. For both areas, responses to their preferred stimuli were strongly modulated by trial-by-trial prediction errors about these stimuli; moreover, this modulation was significantly higher than for their nonpreferred stimuli (Fig. 3*C*–*F*). This finding is consistent with predictive coding theories (Friston, 2005) and extends previous reports of enhanced FFA responses to surprising faces (Summerfield et al., 2008). In contrast, the bilateral putamen, left dorsal premotor cortex, right intraparietal sulcus, and superior parietal gyrus showed prediction error-related responses independent of whether the presented stimulus was a face or a house (Fig. 3*A*,*B*). In the present DCM analysis, we focused on the roles of the putamen and the premotor cortex. The prediction error-related responses of the latter likely reflects the updating of the motor plan that is necessary when the prediction evoked by the auditory cue is wrong (Mars et al., 2007b; Nakayama et al., 2008). In contrast, increased responses in the parietal areas more likely reflects an attentional updating process after unexpected visual stimuli. These areas have been studied extensively in attentional paradigms [e.g., using the Posner paradigm (Posner, 1980)] in which they show increased activation for unexpected stimuli, reflecting attentional changes in response to violations of previous expectations (Giessing et al., 2004; Thiel et al., 2004).

Previous neurophysiological and neuroimaging investigations of associative learning have primarily focused on region-specific prediction error responses, e.g., in the ventral tegmental area (Hollerman and Schultz, 1998; Yacubian et al., 2006; D'Ardenne et al., 2008) or the striatum (Schultz and Dickinson, 2000; McClure et al., 2003; O'Doherty et al., 2003, 2004; Corlett et al., 2004; Seymour et al., 2004; Tobler et al., 2006; Jensen et al., 2007; Menon et al., 2007). In contrast, to date, there has been only one empirical study examining effects of prediction errors on connectivity (den Ouden et al., 2009). This previous study of audiovisual associative learning found prediction error-related responses in the visual cortex and putamen, as well as a modulation of effective connectivity from auditory to visual cortex by prediction error. However, the source of this modulation remained anatomically uninformed. Furthermore, our previous study lacked behavioral evidence for learning (as a result of using an incidental learning paradigm), used nonspecific visual stimuli, and restricted learning to stationary probabilities. All of these limitations were overcome in the present study.

Prediction error-related responses in both dorsal and ventral parts of the striatum have been reported by numerous studies on very different types of learning. These results suggest that the striatum is sensitive to violations of learned contingencies, regardless of whether these contingencies signal reinforcement (McClure et al., 2003; O'Doherty et al., 2003, 2004; Seymour et al., 2004; Jensen et al., 2007; Menon et al., 2007), guide decision making (Corlett et al., 2004), or predict target stimuli (as in the current study), and even when these contingencies are not behaviorally relevant (den Ouden et al., 2009). Other fMRI studies showed that the striatum responds to nonrewarding, unexpected stimuli proportional to the salience of the stimulus (Zink et al., 2006), as well as to novel stimuli (Wittmann et al., 2007, 2008). These results suggest that the striatum may have a general role in processing unexpected events per se. One of the proposed functions of this striatal response is to reallocate processing resources to unexpected stimuli in both reward and nonreward contexts (Zink et al., 2006). Our results, showing that prediction error responses in the putamen indeed modulates information transfer from visual to motor areas, are consistent with such a gating role of the striatum.

Together, our results suggest that the increase of premotor activity for surprising visual outcomes is partially driven by stimulus-specific visual inputs that are gated by the degree of prediction error encoded by the putamen. In other words, the strength of effective connections from FFA and PPA to premotor cortex, which provide information about the appropriateness of the planned action, might change from trial to trial, depending on the mismatch between predicted and observed visual outcome, signaled by the putamen. This gating mechanism is consistent with anatomical studies that have reported indirect connections from the putamen to the premotor cortex via the ventrolateral thalamus (Alexander and Crutcher, 1990; Schultz, 2000). One possibility is that this increased input from sensory areas attributable to striatal gating serves to overcome the lower corticospinal excitability in trials with unexpected outcomes and to facilitate the execution of unprepared actions. Motor preparation increases the excitability of corticospinal projections (Mars et al., 2007a; van Elswijk et al., 2007), and, during probabilistic learning, motor output is biased according to contextual probabilities (Bestmann et al., 2008). Furthermore, there is increasing evidence suggesting that the striatum is involved in task switching and selection of motor programs (Mink, 1996; Cools et al., 2004, 2006; O'Reilly and Frank, 2006). Our finding that the striatum gates sensory information transfer to the premotor cortex suggests one mechanism by which the striatum could influence motor selection.

In a wider context, our results provide additional evidence for current theoretical accounts of brain function, such as predictive coding and the free-energy principle (Rao and Ballard, 1999; Friston, 2005), which posit a fundamental role of prediction errors for adaptive behavior and learning. A central notion of these accounts is the key role of prediction error-dependent synaptic plasticity in driving learning; this concept is also critical for other formal learning theories (Montague et al., 1996; Schultz et al., 1997; Schultz and Dickinson, 2000). In other words, the necessity of reconfiguring neuronal circuits during learning should be inversely proportional to how well those neuronal circuits are capable of predicting sensory stimuli. Although there is a vast literature on neuronal prediction error responses, only sparse direct evidence exists so far that synaptic plasticity (i.e., changes in effective connectivity) scales with the degree of prediction error (cf. den Ouden et al., 2009). The present study provides additional empirical evidence for this mechanism. Moreover, by identifying the putamen as a source of this prediction error-dependent modulation of connectivity, this study links the physiological mechanisms proposed by predictive coding and free-energy accounts of brain function to a large body of literature on prediction error-related responses in the striatum (McClure et al., 2003; O'Doherty et al., 2003, 2004; Corlett et al., 2004; Seymour et al., 2004; Jensen et al., 2007; Menon et al., 2007).

In summary, we used a combination of fMRI, computational (Bayesian) learning models, and DCMs to demonstrate parametric modulation of visuomotor coupling according to prediction error responses in the putamen. To our knowledge, this study is the first to demonstrate that trial-by-trial prediction error responses in a specific region modulate the coupling among other regions. Several neurobiological mechanisms for this type of plasticity have been suggested, including NMDA receptor-dependent short-term plasticity (for review, see Stephan et al., 2008), and it will be an important goal for future studies to identify which of these mechanisms is at work. The combination of computational models of learning and neurophysiological models of connectivity presented in this study represents a novel approach to model-based inference about synaptic plasticity during learning. This general approach may become useful in clinical studies, given the pathophysiological importance of synaptic plasticity for many brain diseases and our lack of other non-invasive methods for investigating it. Ultimately, this approach may help to establish neurophysiologically grounded diagnostic classifications of spectrum diseases, such as schizophrenia (Stephan et al., 2006).

References

Akatsuka K, Wasaka T, Nakata H, Kida T, Kakigi R (2007) The effect of stimulus probability on the somatosensory mismatch field. Exp Brain Res 181:607–614.

Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13:266 –271.

Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. Neuroimage 13:903–919.

Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26:839 – 851.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214 –1221.

Bestmann S, Harrison LM, Blankenburg F, Mars RB, Haggard P, Friston KJ, Rothwell JC (2008) Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. Curr Biol 18:775–780.

Brodersen KH, Penny WD, Harrison LM, Daunizeau J, Ruff CC, Duzel E, Friston KJ, Stephan KE (2008) Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw 21:1247–1260.

Carpenter RH, Williams ML (1995) Neural computation of log likelihood in control of saccadic eye movements. Nature 377:59 – 62.

Cools R, Clark L, Robbins TW (2004) Differential responses in human striatum and prefrontal cortex to changes in object and rule relevance. J Neurosci 24:1129 –1135.

Cools R, Ivry RB, D'Esposito M (2006) The human striatum is necessary for responding to changes in stimulus relevance. J Cogn Neurosci 18: 1973–1983.

Corlett PR, Aitken MR, Dickinson A, Shanks DR, Honey GD, Honey RA, Robbins TW, Bullmore ET, Fletcher PC (2004) Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. Neuron 44:877– 888.

Crammond DJ, Kalaska JF (2000) Prior information in motor and premotor cortex: activity during the delay period and effect on pre-movement activity. J Neurophysiol 84:986 –1005.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267.

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8:429–453.

Dayan P, Niv Y (2008) Reinforcement learning: the good, the bad and the ugly. Curr Opin Neurobiol 18:185–196.

den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in associative learning. Cereb Cortex 19:1175–1185.

Ekman P, Friesen W (1976) Pictures of facial affect. Palo Alto, CA: Consulting Psychologists.

Fogassi L, Raos V, Franchi G, Gallese V, Luppino G, Matelli M (1999) Visual responses in the dorsal premotor area F2 of the macaque monkey. Exp Brain Res 128:194–199.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364:1211–1221.

Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2007) Variational free energy and the Laplace approximation. Neuroimage 34:220–234.

Friston KJ, Stephan KE (2007) Free-energy and the brain. Synthese 159:417–458.

Friston KJ, Stephan KE, Lund TE, Morcom A, Kiebel S (2005) Mixed-effects and fMRI studies. Neuroimage 24:244–252.

Friston KJ, Rotshtein P, Geng JJ, Sterzer P, Henson RN (2006) A critique of functional localisers. Neuroimage 30:1077–1087.

Gamberini M, Passarelli L, Fattori P, Zucchelli M, Bakola S, Luppino G, Galletti C (2009) Cortical connections of the visuomotor parietooccipital area V6Ad of the macaque monkey. J Comp Neurol 513:622–642.

Gerbella M, Belmalih A, Borra E, Rozzi S, Luppino G (2010) Cortical connections of the macaque caudal ventrolateral prefrontal areas 45A and 45B. Cereb Cortex 20:141–168.

Giessing C, Thiel CM, Stephan KE, Rösler F, Fink GR (2004) Visuospatial attention: how to measure effects of infrequent, unattended events in a blocked stimulus design. Neuroimage 23:1370–1381.

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309.

Hutton C, Bork A, Josephs O, Deichmann R, Ashburner J, Turner R (2002) Image distortion correction in fMRI: a quantitative evaluation. Neuroimage 16:217–240.

Hutton C, Deichmann R, Turner R, Anderson JM (2004) Combined correction for geometric distortion and its interaction with head motion in fMRI. Proceedings of the 12th Annual Meeting of the International Society for Magnetic Resonance in Imaging, Kyoto, May.

Iannetti GD, Hughes NP, Lee MC, Mouraux A (2008) Determinants of laser-evoked EEG responses: pain perception or stimulus saliency? J Neurophysiol 100:815–828.

Jensen J, Smith AJ, Willeit M, Crawley AP, Mikulis DJ, Vitcu I, Kapur S (2007) Separate brain regions code for salience vs. valence during reward prediction in humans. Human Brain Mapp 28:294–302.

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304.

Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends Cogn Sci 10:319–326.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.

Luppino G, Calzavara R, Rozzi S, Matelli M (2001) Projections from the superior temporal sulcus to the agranular frontal cortex in the macaque. Eur J Neurosci 14:1035–1040.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Mars RB, Bestmann S, Rothwell JC, Haggard P (2007a) Effects of motor preparation and spatial attention on corticospinal excitability in a delayed-response paradigm. Exp Brain Res 182:125–129.

Mars RB, Piekema C, Coles MG, Hulstijn W, Toni I (2007b) On the programming and reprogramming of actions. Cereb Cortex 17:2972–2979.

Matelli M, Govoni P, Galletti C, Kutz DF, Luppino G (1998) Superior area 6 afferents from the superior parietal lobule in the macaque monkey. J Comp Neurol 402:327–352.

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.

McLaren IP, Kaye H, Mackintosh NJ (1989) An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In: Parallel distributed processing: implications for psychology and neurobiology (Morris RGM, ed), pp 102–120. Oxford: Clarendon.

Menon M, Jensen J, Vitcu I, Graff-Guerrero A, Crawley A, Smith MA, Kapur S (2007) Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: effects of dopaminergic modulation. Biol Psychiatry 62:765–772.

Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381–425.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Montague PR, Hyman SE, Cohen JD (2004) Computational roles for dopamine in behavioural control. Nature 431:760–767.

Nakayama Y, Yamagata T, Tanji J, Hoshi E (2008) Transformation of a virtual action plan into a motor plan in the premotor cortex. J Neurosci 28:10287–10297.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

O'Doherty J, 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'Reilly RC, Frank MJ (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput 18:283–328.

Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. Neuroimage 22:1157–1172.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045.

Pincze Z, Lakatos P, Rajkai C, Ulbert I, Karmos G (2002) Effect of deviant probability and interstimulus/interdeviant interval on the auditory N1 and mismatch negativity in the cat auditory cortex. Brain Res Cogn Brain Res 13:249–253.

Pitt MA, Myung IJ (2002) When a good fit can be bad. Trends Cogn Sci 6:421–425.

Posner MI (1980) Orienting of attention. Q J Exp Psychol 32:3–25.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Requin J, Granjon M (1969) The effect of conditional probability of the response signal on the simple reaction time. Acta Psychol (Amst) 31:129–144.

Schultz W (2000) Multiple reward signals in the brain. Nat Rev Neurosci 1:199–207.

Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annu Rev Neurosci 23:473–500.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

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.

Stephan KE, Baldeweg T, Friston KJ (2006) Synaptic plasticity and dysconnection in schizophrenia. Biol Psychiatry 59:929–939.

Stephan KE, Marshall JC, Penny WD, Friston KJ, Fink GR (2007a) Interhemispheric integration of visual processing during task-driven lateralization. J Neurosci 27:3512–3522.

Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007b) Comparing hemodynamic models with DCM. Neuroimage 38:387–401.

Stephan KE, Kasper L, Harrison LM, Daunizeau J, den Ouden HE, Breakspear M, Friston KJ (2008) Nonlinear dynamic causal models for fMRI. Neuroimage 42:649–662.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017.

Strange BA, Duggins A, Penny W, Dolan RJ, Friston KJ (2005) Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw 18:225–230.

Summerfield C, Koechlin E (2008) A neural representation of prior information during perceptual inference. Neuron 59:336 –347.

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004 –1006.

Tanji J, Evarts EV (1976) Anticipatory activity of motor cortex neurons in relation to direction of an intended movement. J Neurophysiol 39:1062–1068.

Tanné-Gariépy J, Rouiller EM, Boussaoud D (2002) Parietal inputs to dorsal versus ventral premotor areas in the macaque monkey: evidence for largely segregated visuomotor pathways. Exp Brain Res 145:91–103.

Thiel CM, Zilles K, Fink GR (2004) Cerebral correlates of alerting, orienting and reorienting of visuospatial attention: an event-related fMRI study. Neuroimage 21:318 –328.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. J Neurophysiol 95:301–310.

van Elswijk G, Kleine BU, Overeem S, Stegeman DF (2007) Expectancy induces dynamic modulation of corticospinal excitability. J Cogn Neurosci 19:121–131.

Webster MJ, Bachevalier J, Ungerleider LG (1994) Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. Cereb Cortex 4:470 – 483.

Wise SP, Mauritz KH (1985) Set-related neuronal activity in the premotor cortex of rhesus monkeys: effects of changes in motor set. Proc R Soc Lond B Biol Sci 223:331–354.

Wittmann BC, Bunzeck N, Dolan RJ, Düzel E (2007) Anticipation of novelty recruits reward system and hippocampus while promoting recollection. Neuroimage 38:194 –202.

Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58:967–973.

Yacubian J, Gläscher J, Schroeder K, Sommer T, Braus DF, Büchel C (2006) Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J Neurosci 26:9530 –9537.

Yamagata T, Nakayama Y, Tanji J, Hoshi E (2009) Processing of visual signals for direct specification of motor targets and for conceptual representation of action targets in the premotor cortex. J Neurophysiol 102:3280 –3294.

Zink CF, Pagnoni G, Chappelow J, Martin-Skurski M, Berns GS (2006) Human striatal activation reflects degree of stimulus saliency. Neuroimage 29:977–983.

Supplemental Materials

1. Log model evidence for behavioural data

This appendix describes how we evaluated the log-evidence for our linear observation model of reaction times, which has the form:

$$\begin{aligned} &Y = X\beta + \varepsilon \Rightarrow \\ &p\left(Y,\beta,\sigma^2 \mid m\right) = \left(2\pi\right)^{-d/2}\sigma^2 \exp\left(-\frac{\left(Y - X\beta\right)^T\left(Y - X\beta\right)}{2\sigma^2}\right), \end{aligned} \tag{A1}$$

Here, Y represents the data (response speeds), X is a design matrix and $\varepsilon \sim N(0,\sigma^2)$ are normally distributed errors. Using Jeffreys' (non-informative) priors for β and σ (i.e. $\beta \sim 1, \sigma \sim 1/\sigma$) , the evidence is given by

$$\begin{aligned} p(Y|m) &= \int\int p(Y,\beta,\sigma \mid m)d\beta d\sigma \\ &= (2\pi)^{(r-d)/2}\left|X^T X\right|^{-r/2}\Gamma(d-r-1)(\lambda/2)^{r+1-d} \end{aligned} \tag{A2}$$

where r is the number of parameters, d is the number of data-points and

$$\lambda = Y^T\left(I - X\left(X^T X\right)^{-1}X^T\right)Y \tag{A3}$$

is the sum of squared residuals. Therefore the log model evidence is

$$\log\left(p\left(Y|m\right)\right) = \frac{r-d}{2}\log\left(2\pi\right) - {}^{r}\!/\!_{2}\log\left(\left|X^T X\right|\right) + \log\left(\Gamma\left(d-r-1\right)\right) + \left(r+1-d\right)\log\left({}^{\lambda}\!/\!_{2}\right)$$

(A4)

This is an exact expression for the log evidence of this model. It can be generalized to include observation models whose design matrix is informed by the trial-b-trial estimates of an underlying learning model with parameters θ (e.g. the learning rate in the Rescorla-Wagner model):

$$y = X\left(\theta\right)\beta + \varepsilon \tag{A5}$$

Assuming that the residuals are Gaussian ($\varepsilon \sim N\left(0,\sigma^2 I_d\right)$) yields the likelihood function, i.e.:

$$p\left(y|\theta,\beta,\sigma\right) = N\left(X\left(\theta\right)\beta,\sigma^2 I_d\right) \tag{A6}$$

We now can integrate over both β and σ to yield the restricted data likelihood $p(y|\theta,m)$:

$$\begin{aligned} p(y|\theta,m) &= \int p(y|\theta,\beta,\sigma,m)\underbrace{p(\beta|m)}_{\propto 1}\underbrace{p(\sigma|m)}_{\propto 1/\sigma}d\beta d\sigma \\ &= (2\pi)^{(r-d)/2}\,\Gamma(d-r-1)\left|X(\theta)^T X(\theta)\right|^{-r/2}\left(\frac{\lambda(\theta)}{2}\right)^{r+1-d} \end{aligned} \tag{A7}$$

where we used Jeffreys' (non-informative) priors for β and σ, and the sum of squared estimated residual error $\lambda(\theta)=\hat{\varepsilon}^T\hat{\varepsilon}$ is given by:

$$\lambda(\theta)=y^T\left(I_d - X(\theta)\left(X(\theta)^T X(\theta)\right)^{-1} X(\theta)^T\right)y \tag{A8}$$

This (restricted) likelihood function is obviously not conjugate to, for example, simple Gaussian priors on θ. This means that there is no analytical expression for the model evidence $p(y|m)$:

$$p(y|m)=\int p(y|\theta,m)\,p(\theta|m)\,d\theta \tag{A9}$$

However, it is possible to use the so-called Laplace approximation (e.g. see (Friston et al., 2007)) to finesse this difficult integration problem. First, let us derive a second-order Taylor expansion to the log restricted likelihood (LReL) $t(\theta)$:

$$\begin{aligned} t(\theta) &= \ln p(y|\theta,m) \\ &\approx t(\hat{\theta})+\frac{1}{2}(\theta-\hat{\theta})^{\mathrm{T}}\underbrace{\left.\frac{\partial^2 t}{\partial\theta^2}\right|_{\hat{\theta}}}_{-H(\hat{\theta})}(\theta-\hat{\theta}) \end{aligned} \tag{A10}$$

where $H(\hat{\theta})=-\left.\frac{\partial^2 t}{\partial\theta^2}\right|_{\hat{\theta}}$ is the negative Hessian of the LReL. Then, under Jeffrey's non-informative priors for θ, the log model evidence can be approximated as:

$$\begin{aligned} \ln p(y|m) &= \ln\int \exp(t(\theta))\,d\theta \\ &\approx t(\hat{\theta})+\frac{r_\theta}{2}\ln 2\pi+\frac{1}{2}\ln\left|H(\hat{\theta})^{-1}\right| \quad . \\ &= \underbrace{\ln p(y|\hat{\theta},m)+\frac{r_\theta}{2}\ln 2\pi-\frac{1}{2}\ln\left|H(\hat{\theta})\right|}_{F} \end{aligned} \tag{A11}$$

where r_θ is the dimensionality of parameter vector θ (i.e. the number of free parameters in the underlying learning model).
In equation 7, $F \approx \ln p\left(y \middle| m\right)$ is referred to as the Laplace approximation to the model evidence. Note that the well-known Bayesian Information Criterion (BIC) is simply the asymptotic limit ($d \to \infty$) to F, as given above:

$$\begin{aligned} F &= \underbrace{\ln p\left(y \middle| \hat{\theta}, m\right)}_{O(d)} + \underbrace{\frac{r_\theta}{2}\ln 2\pi}_{O(1)} - \frac{1}{2}\underbrace{\ln\left|H\left(\hat{\theta}\right)\right|}_{O(r_\theta \ln d)} \\ &\xrightarrow[y\text{ iid}]{d\to\infty} \underbrace{\ln p\left(y \middle| \hat{\theta}, m\right) - \frac{r_\theta}{2}\ln d}_{\text{BIC}} \end{aligned} \qquad \text{(A12)}$$

Note that when the underlying learning model has no free parameters then $\frac{r_\theta}{2}\ln d = 0$ and Eq. A12 reduces to the exact expression for the log-evidence in Eq. A4. This means that Eq. A12 can be used to compare any observation models, even if they differ in the number of parameters of the underlying learning model.

2. Bayesian volatility-based associative learning model

We start with the premise that subjects represent or infer the causes of their sensory inputs and optimise their behaviour on the basis of this inference. From a Bayesian perspective, the brain is an *observer* of its own sensory signals. This means subjects invert some forward or generative model of sensory inputs to represent the unobserved (hidden) causes of that input. Any learning then relies strongly on the subject's model of the world (the perceptual model), which determines predictions, and hence, prediction error. This Bayesian perspective has already been used to model behavioural decisions (e.g. (Kording et al., 2007).

In what follows, we describe the volatility-based perceptual model used in this study to estimate the volatility and probabilities of the observed events (i.e. cue-outcome pairs). This model is based on the proposal by Behrens et al. (Behrens et al., 2007) and subsumes the set of probabilistic assumptions the brain encodes in order to represent the causes of paired audio-visual stimuli. The perceptual model generates sensory input u (*e.g.*, experimental stimuli) from hidden causes x (*e.g.*, experimental factors or environmental states) and can be expressed in terms of a likelihood model $p\left(u \mid x\right)$ and prior beliefs $p\left(x\right)$. The states of the world x are unknown to the subject but are under experimental control. In our example, u is a series of cue-outcome pairs, presented to the observer, and x encodes the probabilistic cue-outcome association that the subject has to learn in order to predict its environment adequately. The prior belief itself is decomposed into a hierarchy of conditional probability density functions, as will be described bellow.

Let u_t be the outcome at trial *t* be a multinomial random variate such that:

$$\begin{aligned} p\left(u_t \left| u_t^c, r_t \right.\right) &= Mult\left(u_t \left| r_t \right.\right) \\ &= \prod_{i=1}^{n} \left(r_t^i\right)^{u_t^i} \end{aligned},$$

where $\left(r_t^i\right)^{i=1,\ldots,n}$ is a $n\times1$ vector of probabilities describing completely the distribution of the n possible outcomes. This forms the likelihood of our generative model. Note that from there on, we will consider that each of the cues u_t^c is associated with its own likelihood, and consequently, its own generative model. This means that everything we state below is conditional on the given cue. As a consequence, the Bayesian inversion of such a set of generative models is cue-specific, and has to be replicated for all different cues.

This vector of cue-outcome association probabilities obeys *a priori* the following Dirichlet distribution:

$$\begin{aligned} p\left(r_t \left| r_{t-1}, v_t \right.\right) &= Dir\left(r_t \left| a_t \right.\right) \\ &= \frac{\Gamma\left(a_t^0\right)}{\prod_{i=1}^{n} \Gamma\left(a_t^i\right)} \prod_{i=1}^{n} \left(r_t^i\right)^{a_t^i - 1} \end{aligned}$$

This transition density is actually a martingale; i.e. it is a first order Markov process whose current first order moment is equal to its previous realization:

$$\left\langle r_t \right\rangle = r_{t-1}.$$

Furthermore, the precision (i.e. inverse variance) of the transition from r_{t-1} to r_t is parameterized by a scalar quantity v_t, which measures the volatility of the environment:

$$\sum_{i=1}^{n} a_t^i = \exp\left(-v_t\right) + 1$$

The volatility itself is assumed to vary over time as a martingale, and the above parameterization makes a simple AR(1) model possible:

$$\begin{aligned} p\left(v_t \left| v_{t-1}, K \right.\right) &= N\left(v_t \left| v_{t-1}, K \right.\right) \\ &= \frac{1}{\sqrt{2\pi K}} \exp\left(-\frac{1}{2K}\left(v_t - v_{t-1}\right)^2\right), \end{aligned}$$

where K is the prior variance of the volatility, i.e. the volatility's volatility.

The prior on K itself is supposed to be non-informative, i.e.:

$$p\left(K\right) \propto 1.$$

To summarize, the generative model assumes the following cascade of events (illustrated in the graph in Figure S1):

1- A value for the volatility variance K is randomly drawn from its prior pdf $p(K)$. Then, at each trial *t*:

2- This value determines the transition pdf of the volatility. Then, a first value v_t is randomly drawn from $p\left(v_t \middle| v_{t-1}, K\right)$.

3- Knowing the volatility v_t then allow us to derive the transition density for r_t. Then, a value for the cue-outcome association probability is drawn from $p\left(r_t \middle| r_{t-1}, v_t\right)$.

4- This finally defines the likelihood of the outcome itself: the first outcome u_t is then drawn randomly from $p\left(u_t \middle| u_t^c, r_t\right)$.

5- The steps 2, 3 and 4 are repeated in time, giving rise to three time series for the volatility v_t, the cue-outcome association probability r_t and the observed outcomes u_t.

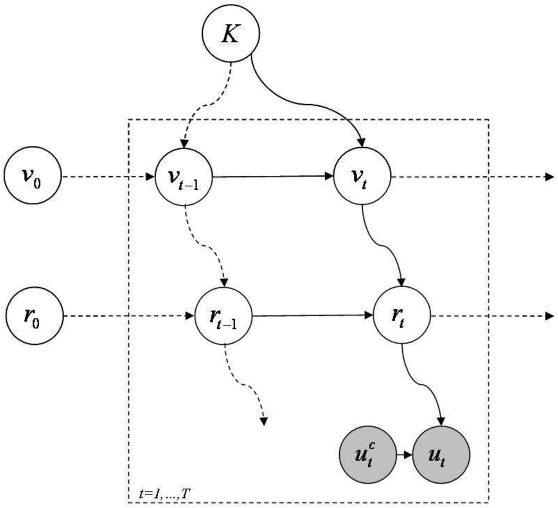

Figure S1. Graph illustration of the volatility model. u_t= observed outcome at trial *t*; r_t = cue-outcome association probability; v_t= volatility; *K* = variance of the volatility.

The model assumes that the observer updates its posterior belief on-line, in the light of incoming data, in a Kalman filter-like manner. The joint posterior pdf over the full set of unknown variables, namely $x = \{K, v, r\}$, then follows the following prediction and update steps:

$$\text{prediction: } p\left(r_t, v_t, K \middle| u_{1:t-1}\right) = \iint p\left(r_t \middle| r_{t-1}, v_{t-1}\right) p\left(v_t \middle| v_{t-1}, K\right) p\left(r_{t-1}, v_{t-1}, K \middle| u_{1:t-1}\right) dr_{t-1} dv_{t-1}$$

$$\text{update: } p\left(r_t, v_t, K \middle| u_{1:t}\right) = \frac{p\left(r_t, v_t, K \middle| u_{1:t-1}\right) p\left(u_t \middle| r_t\right)}{\iiint p\left(r_t, v_t, K \middle| u_{1:t-1}\right) p\left(u_t \middle| r_t\right) dr_t dv_t dK} \quad .$$

These two steps are iterated as long as new data are measured and, after each cue-outcome observation, yield estimates of both the current cue-outcome association probability r_t and the environmental volatility v_t, as well as an estimator of the static volatility's variance K, given all previously observed data. In the present study, the trajectory of these estimates as a function of time (trial t) served as predictors for behavioural data (response speeds) and neuroimaging data (fMRI data in SPM and DCM analyses).

3. Alternative learning models

In addition to the two learning models described in the main text (i.e. a linear model based on the true probabilities generating the stimulus sequence and a hierarchical Bayesian observer), we tested three further models, following suggestions by our reviewers. Firstly we employed a simple "model-free" reinforcement learning approach using a classical Rescorla-Wagner learning model. As there were no significant differences in behaviour for the two cues (see main text), the model fitted one joint learning rate for both cue types. The associative strengths V_t of the cues to the outcomes (face = 1, house = 0) were modelled according to the equation

$$V_t = V_{t-1} + \alpha(\lambda_{t-1} - V_{t-1})$$

where λ is the outcome (face our house) and α denotes the learning rate, which is fitted for each individual subject.

The Rescorla Wagner model, like the Bayesian learning model, has no explicit knowledge of the task structure. It is conceivable, however, that over the course of the experiment the subjects learned to recognise the discrete levels of predictability of the cues. Therefore, we tested two additional models that did represent the underlying structure of the task. These models consisted of two variants of a first order hidden Markov model (HMM), which were used to model the sequences of observed cue-outcome combinations (Rabiner, 1989). An HMM is a set of hidden states, each of which is probabilistically associated with an observable output (in our case a cue-outcome combination). Transitions between states occur stochastically; the conditional probabilities (transition probabilities) are such that the state at time t depends only on the state at time t-1 (Markov property). In the HMMs used here, five hidden states were used to represent the five discrete levels of associative cue strengths. This knowledge about the levels of associative strengths in the experiment enabled the model to represent the subject's potential "metalearning" about the levels that the strengths could change to, allowing for sudden jumps between levels when associations changed rather than having to learn them anew each time.

We used the Baum-Welch algorithm (also known as forward-backward algorithm (Baum et al., 1970)) to find the transition matrix that best explained the observed cue-outcome combinations. In a first version of the HMM ("HMM fixed" in Fig. S1 and Table S1), the optimal transition matrix was learned from the entire observed sequence. However, this equates inference with learning, and assumes that subjects know the structure of the task from the start of the experiment. An alternative scheme ("HMM learn") is to update the transition matrix each time an observation is made, that is, to run the Baum-Welch algorithm at each trial anew, using the trial sequences $\{[1,2],[1,2,3],\ldots,[1\ldots N]\}$. A priori, i.e. at the start of the experiment, there was an equal belief to be in each of the 5 states.

Figure S2 shows the estimated probability of observing a face following one of the CS in block 3, (cf. figure 1C main text) as computed by each of the four learning models and juxtaposes these trial-by-trial estimates to the true probabilities. For each of these models the log model evidence was calculated as described in section S1, taking into account the additional parameter of the RW model (i.e. the learning rate) by means of the Bayesian Information Criterion.

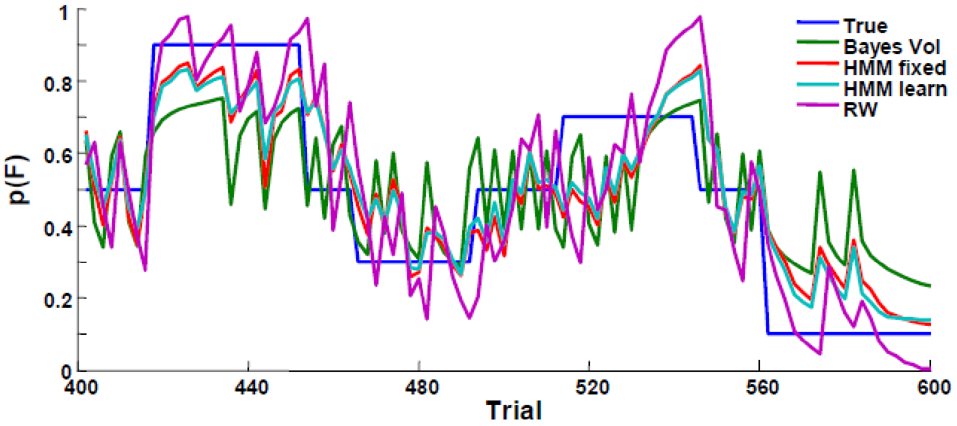

Figure S2. *p*(F|CS) as estimated by the various models.

The model evidences were then used for a random effects Bayesian model selection across subjects (see main text). The hierarchical Bayesian learner turned out to be superior to any of the other tested models (Table S1), with an exceedance probability of 66% that this model was more likely than any other model considered.

Table S1. BMS results for behavioural model comparison

	Dirichlet parameters α	**Exceedance probability φ**
True categorical model	1.00	0.00
Bayesian volatility model	8.99	0.66
HMM (fixed)	6.53	0.22
HMM (learn)	3.39	0.02
Rescorla Wagner	5.11	0.09

4. Additional SPM results

The main text deals only with key questions of interest for this study, namely characterization of stimulus-independent and stimulus-specific surprise responses and connectivity changes. For completeness, the results of additional analyses are reported here; these include a detailed analysis of the main effects of the stimuli as well as an analysis of regional responses associated with the estimated volatility of the probabilistic associations.

Stimulus main effects in FFA and PPA

As expected, the mid fusiform gyrus was activated more strongly to *face* stimuli than to *house* stimuli (FFA, Table S1), and the parahippocampal gyrus showed the opposite effect (PPA, Table S1). At the group level the FFA activation was significant at whole brain corrected level only in the right hemisphere, but the left FFA activation was significant within an anatomically defined ROI for the fusiform gyrus (table S1).

Volatility dependent brain activations

Although this was not the focus of this study, for completion we also tested in which areas activity increased or decreased with the trial-by-trial volatility estimates. Following the results by Behrens et al., (2007), who demonstrated that ACC activity correlated with volatility estimates during reward learning, we tested whether volatility encoding in the ACC would also be present in our purely perceptual paradigm which did not include any rewards. Figure S3A shows the trial-by-trial estimates of the volatility for session 3 (compare figure 1C in the main text for the parallel probability estimates). Indeed, activity in the dorsal and rostral ACC and the ventromedial prefrontal cortex correlated significantly with the volatility estimates (Table S2 and Fig. S3B-C).

Although the use of a volatile environment was not a phenomenon of primary interest for this study, but merely a means of enforcing continuous learning (and thus maximising induction of synaptic plasticity and hence changes in connectivity), it is noteworthy that our analysis of volatility effects replicated previous results (Behrens et al., 2007).

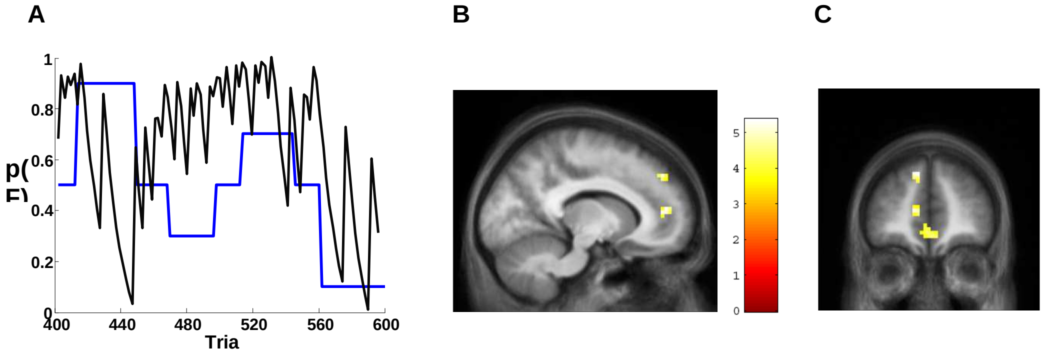

Figure S3. Volatility effects. A) Trial-by-trial estimates of the volatility. Blue line = true probabilities, black = volatility. **(B,C)** The anterior cingulate cortex and ventromedial prefrontal cortex show a positive correlation with the volatility estimate of the Bayesian observer model, here shown in a sagittal **(A)** and axial **(B)** slice

Table S2. MNI coordinates and Z-values for significantly activated regions

	MNI coords.			
Foci of activation	x	y	Z	Z score
Main effects of sensory stimulation				
House>Face				
R parahippocampal gyrus*	30	-51	12	7.01
L parahippocampal gyrus*	-24	-57	-18	6.70
Face> House				
R mid fusiform gyrus*	45	-57	-24	5.42
L amygdala*	-21	-12	-9	4.31
L mid fusiform gyrus**	-45	-54	-21	3.47
Volatility effects				
positive correlation				
Ventromedial prefrontal ctx*	3	48	-9	3.64
ACC**	-12	45	9	4.11
Ventral ACC / subgenual ctx **	-6	36	-3	3.57
L caudate/thalamus*	-21	-9	9	4.32
negative correlation				
No significant activations.				

* significant at p<0.05 FEW cluster-level corrected across the whole-brain

** significant at p<0.05 cluster-level corrected for a priori region of interest

5. Optimisation of fixed connections (DCM)

In order to optimise the fixed connections, a basic model was defined that included the minimal number of connections necessary to test the hypothesis outlined in the main text. The endogenous connectivity of this ‘minimal’ model was then optimised by systematically adding connections. A random effects Bayesian model selection (BMS) procedure was then used to select the optimal model at the group level (Stephan et al., 2009). This procedure quantifies the relative goodness of models in terms of exceedance probabilities ϕ_i which denote the probability that model i is superior to any other model considered, given the data from all subjects. Note that exceedance probabilities are a function of model space; for example, because they sum to unity over all models considered, they decrease monotonically when increasing the set of alternative models. They are thus to be interpreted in relative, not absolute terms.

A minimal model (Figure S4, m_1) included only the endogenous connections from the sensory areas to the PMd, and these connections were modulated by the activity from the putamen. An additional six models (m_2-m_7) were derived from this basic architecture in two steps. In a first step, we compared all combinations of endogenous connections between PPA, FFA and PMd (m_1-m_4) and two models (m_5, m_6) with connections to the putamen from the visual and premotor cortex, respectively. Model m_5 tested whether there was any direct influence of FFA and PPA on the putamen. Model m_6 included a direct connection from the PMd to the putamen (Fig. S4, m_6) because there exist direct projections from the premotor cortex to the putamen (Takada et al., 1998;Leh et al., 2007). Based on a suggestion by one of our reviewers we also included 2 models in which Face and House stimuli directly entered PPA and FFA, respectively. For this, we tested a model with (m_7) and without (m_8) reciprocal connections between the PPA and FFA.

Note that comparing DCMs with additional connections is not equivalent to testing whether these connections do or do not exist anatomically, but whether these connections play a functional role in the process modelled. Comparing all six models (m_1-m_6) against each other using random effects BMS, model m_4 turned out to be the best model, i.e. a model with full reciprocal connectivity between PPA, FFA and PMd (but not direct connections from either visual areas nor PMd to the putamen). The exceedance probability for model m_4 was $\varphi_4 = 0.44$, surpassing the exceedance probabilities of all other models (which ranged from 0.01 to 0.28; see Table S3).

Once we had identified the most likely pattern of connections among the areas, we constructed an additional model (m_{pm}) and compared this to model m_4 in order to verify the specificity of the modulatory influence exerted by the putamen (note that in the main text m_4 is referred to as m_{pt}). Since the putamen and the PMd showed similar prediction error related activations (Fig. 3, main text), we wished to demonstrate that putamen activity gated visuomotor connections (Fig. 4C, main text), instead of premotor activity gating visuostriatal connections (Fig. 4B, main text). Indeed, BMS showed that the reversed model (m_7, with PMd as source of modulatory effects) was clearly inferior to the original model (m_4, with the putamen as source of modulatory effects), with an exceedance probability of 99% in favour of the latter (see Fig. 4D, main text).

Table S3. BMS results for models m_1-m_8

	Dirichlet parameters α	**Exceedance probability φ**
m_1	1.68	0.01
m_2	4.11	0.17
m_3	1.70	0.01
m_4	5.63	0.44
m_5	3.06	0.07
m_6	4.81	0.28
m_7	1.02	0.00
m_8	1.00	0.00

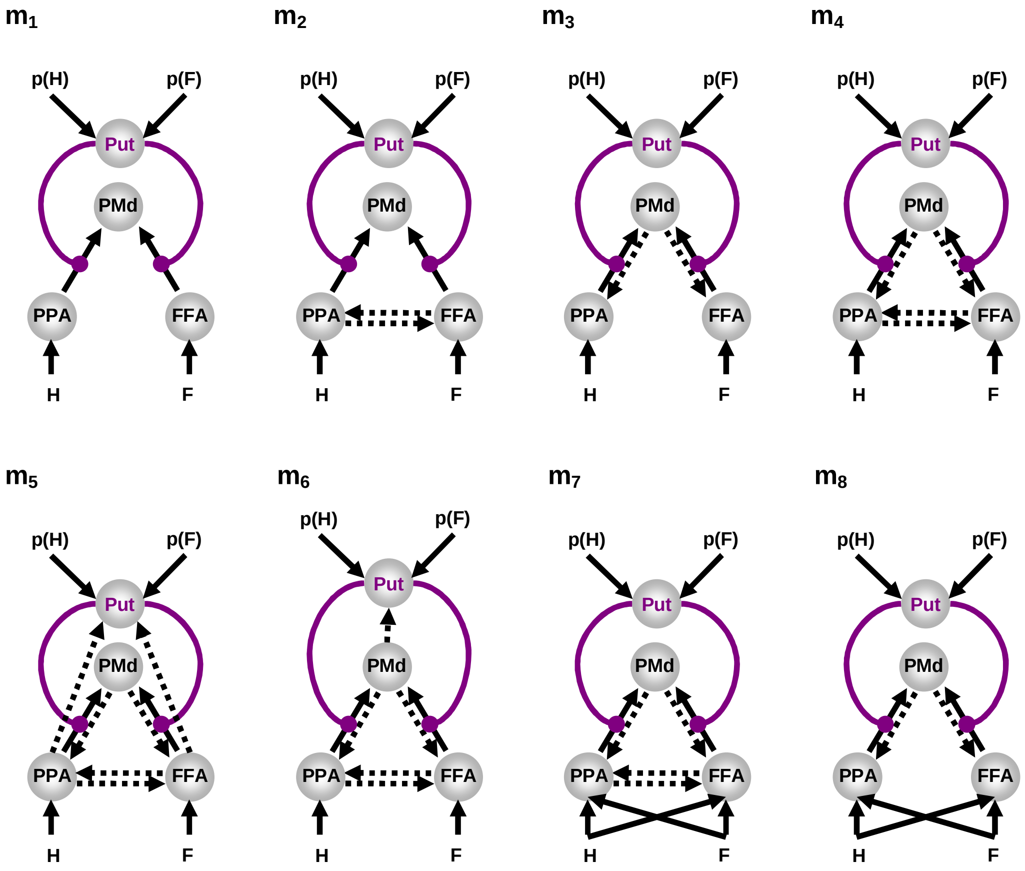

Figure S4. Alternative DCMs. M_1 includes the minimal connections needed to model the observed modulatory effects in the premotor cortex (PMd); this model includes endogenous connections from the PPA and FFA to the PMd, and these connections are modulated by output activity from the putamen. Models 2-4 then add or exclude connections between the sensory and premotor areas. Model 5 includes direct connections from the sensory areas to putamen, and model 6 includes a connection from PMd to the putamen. Model 7 and 8 include direct inputs of both visual stimulus types to both the PPA and FFA. In an additional model (m_{pm}) the role of the putamen and the PMd were swapped such that PMd activity modulated visual afferents to the putamen; this model is shown in Figure 4C in the main text.

Reference List

Baum LE, Petrie T, Soules G, Weiss N (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Ann Math Statistics 1: 164-171.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10: 1214-1221.

Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2007) Variational free energy and the Laplace approximation. Neuroimage 34: 220-234.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS ONE 2: e943.

Leh SE, Ptito A, Chakravarty MM, Strafella AP (2007) Fronto-striatal connections in the human brain: a probabilistic diffusion tractography study. Neurosci Lett 419: 113-118.

Rabiner LR (1989) Proc IEEE 77: 257-286.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46: 1004-1017.

Takada M, Tokuno H, Nambu A, Inase M (1998) Corticostriatal projections from the somatic motor areas of the frontal cortex in the macaque monkey: segregation versus overlap of input zones from the primary motor cortex, the supplementary motor area, and the premotor cortex. Exp Brain Res 120: 114-128.