Cerebral Cortex May 2009;19:1175–1185
doi:10.1093/cercor/bhn161
Advance Access publication September 26, 2008

A Dual Role for Prediction Error in Associative Learning

Hanneke E.M. den Ouden[1], Karl J. Friston[1], Nathaniel D. Daw[2], Anthony R. McIntosh[3] and Klaas E. Stephan[1,4]

[1]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK, [2]Department of Psychology, New York University, New York, NY 10003, USA, [3]Rotman Research Institute of Baycrest Centre, University of Toronto, Toronto, Ontario, Canada M6A 2E1 and [4]Branco-Weiss-Laboratory, Institute for Empirical Research in Economics, University of Zürich, Switzerland

Confronted with a rich sensory environment, the brain must learn statistical regularities across sensory domains to construct causal models of the world. Here, we used functional magnetic resonance imaging and dynamic causal modeling (DCM) to furnish neurophysiological evidence that statistical associations are learnt, even when task-irrelevant. Subjects performed an audio-visual target-detection task while being exposed to distractor stimuli. Unknown to them, auditory distractors predicted the presence or absence of subsequent visual distractors. We modeled incidental learning of these associations using a Rescorla–Wagner (RW) model. Activity in primary visual cortex and putamen reflected learning-dependent surprise: these areas responded progressively more to unpredicted, and progressively less to predicted visual stimuli. Critically, this prediction-error response was observed even when the *absence* of a visual stimulus was surprising. We investigated the underlying mechanism by embedding the RW model into a DCM to show that auditory to visual connectivity changed significantly over time as a function of prediction error. Thus, consistent with predictive coding models of perception, associative learning is mediated by prediction-error dependent changes in connectivity. These results posit a dual role for prediction-error in encoding surprise and driving associative plasticity.

Keywords: associative learning, cross-modal, dynamic causal modeling, effective connectivity, fMRI, Rescorla–Wagner model

Introduction

Among the fundaments of adaptive behavior is the ability to predict future events. This ability is crucial to functions ranging from sensory processing to decision making. In psychology and neuroscience, prediction has been studied most extensively in the context of Pavlovian and instrumental conditioning tasks, which measure how organisms anticipate (and act on) affectively significant events such as food delivery or electric shocks. A recent series of functional neuroimaging studies has investigated the neurophysiological basis of prediction and learning in humans. Using Pavlovian and instrumental conditioning tasks, these studies have identified several areas where blood oxygenation level-dependent (BOLD) signals correlate with trial-wise estimates from formal learning models like temporal difference (TD) learning (Sutton and Barto 1998) or the Rescorla–Wagner (RW) model (Rescorla and Wagner 1972). In particular, BOLD activity in areas including the striatum and the dorsolateral prefrontal cortex (DLPFC) (key dopaminergic targets) has been shown to covary with both *predictions* and *prediction errors* (Fletcher et al. 2001; McClure et al. 2003; Corlett et al. 2004; O'Doherty et al. 2004; Seymour et al. 2004; Turner et al. 2004; Gläscher and Büchel 2005; Pessiglione et al. 2006; Jensen et al. 2007).

In all of these previous studies, the learned associations had direct relevance for behavior, either because they were linked to rewarding or punishing outcomes (e.g., McClure et al. 2003; O'Doherty et al. 2004; Seymour et al. 2004) or because subjects received feedback on their performance (Fletcher et al. 2001; Aron et al. 2004; Corlett et al. 2004; Turner et al. 2004). In contrast, it is unclear whether incidental learning of stimulus–stimulus associations, i.e., learning of associations that are irrelevant for current behavioral goals, draws upon the same neuronal mechanisms. A paradigm that shows that these types of associations are learned is sensory preconditioning. Here, in a first stage, the subject is exposed to behaviorally meaningless CS_1–CS_2 associations and, in a second stage, to CS_1–US (unconditioned stimulus) pairings. In a third and final stage, the presentation of a CS_2 alone generates a conditioned response, indicating that the subject must have learned the initial CS_1–CS_2 association (Brogden 1939; Gewirtz and Davis 2000).

In this study we used a factorial design that extended the first stage of classical sensory preconditioning paradigms. Healthy volunteers performed an audio-visual target-detection task, while being exposed to a stream of concurrent audio-visual "distractor" stimuli (Fig. 1). These stimuli possessed statistical regularities, which enabled prediction of the visual distractor from the preceding auditory cue (Fig. 2). Critically, however, these statistical associations were completely irrelevant to the target-detection task. Any learning of these associations would therefore be of an incidental (task-unrelated) nature and, in the absence of behavioral responses to the learned associations, could only be inferred neurophysiologically. This paradigm capitalized on previous work by McIntosh et al. (McIntosh et al. 1998) who used positron emission tomography (PET) to show that learning of associations between sensory stimuli was reflected by activity in early visual cortex. However, the use of PET permitted only a simple conditioning scheme and precluded a full investigation of dynamic changes in the brain's representation of the learned association. Here, we employed a more refined conditioning scheme and used functional magnetic resonance imaging (fMRI) to study learning-dependent changes in brain activity over time. Additionally, we assessed learning-dependent changes in effective connectivity between auditory and visual cortex using dynamic causal modeling (DCM).

Using a 4-factorial design (c.f. Fig. 2), this study characterized learning in terms of the temporal evolution (learning; factor 1) of both brain activity and interregional connectivity in response to a visual stimulus whose presence or absence (V^+ vs. V^-; factor 2) was predicted in 2 contexts, established by 2 types of auditory conditioning stimuli (CS^+ vs. CS^-; factor 3), each of which could be present or absent on each trial (A^+ vs. A^-; factor 4). In other

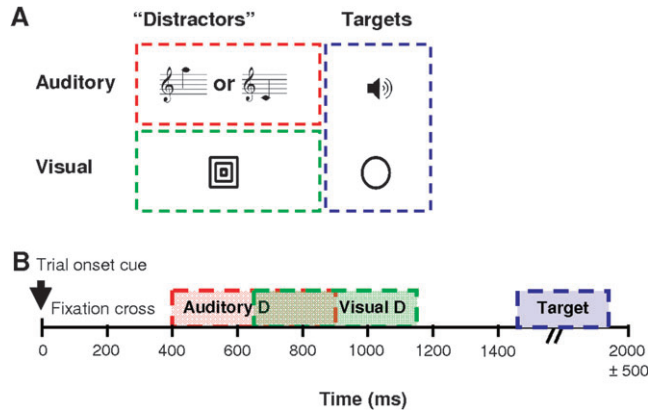

Figure 1. Experimental design. (*A*) stimuli presented during the experiment. The "distractor" stimuli, whose associations are being learned incidentally, comprised 2 auditory CS corresponding to high- and low-frequency tones and one visual US consisting of 3 concentric squares. The target stimuli, to which the subjects responded, comprised a white noise burst and a circle. (*B*) Temporal sequence of a single trial. The CS and US could be either presented or omitted. The average trial duration was 2 s. The TO cue was a small central dot (100 ms); the auditory CS was presented for 500 ms, starting 400 ms after TO. The visual stimulus was presented 750 ms after TO, also for 500 ms. The intertrial interval (ITI) was jittered, ranging from 350–1350 ms, and target stimuli were inserted only in the longest ITIs, lasting for 300 ms.

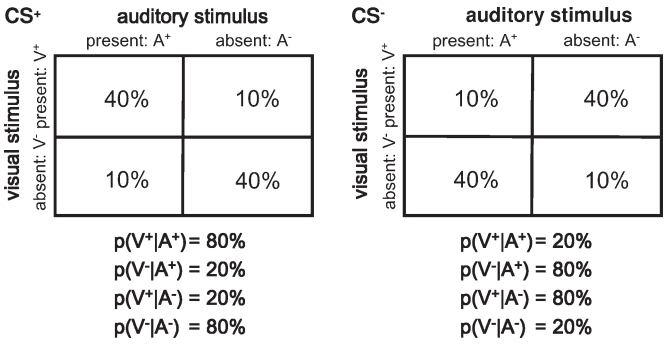

Figure 2. Probabilistic relationship between auditory and visual stimuli. Contingency tables showing the proportion of each trial type occurring during CS^+ and CS^- blocks respectively. Below the tables are the resulting conditional probabilities of the visual stimulus being present (or absent), given the presence (or absence) of the auditory CS; these probabilities can be inferred by comparing the frequencies within each column of the table.

words, in contrast to a classical sensory preconditioning paradigm, we could not only investigate differential learning, depending on CS type but could also assess whether the consequences of an absent CS were learned. It should be noted that both the CS^+ and CS^- context (or blocks) were balanced in terms of stimuli; the a priori probabilities of the auditory CS and of the visual stimulus occurring on a given trial were always 50%. Critically, the task was not related to these auditory and visual stimuli; subjects performed a target-detection task on unrelated stimuli that were presented sporadically.

One of the features of our factorial paradigm is that on half the trials the auditory CS is absent. This necessitates an additional cue that marks the beginning of each trial which was a visual trial onset (TO) cue. In other words, learning of stimulus associations in this paradigm has 2 components, one related to the auditory CS and another related to the visual TO cue. As a consequence, any model of the learning process must be able to formulate how a net prediction is computed from the associative strengths of the 2 cue components. Here we chose the RW model because it is the simplest and most generic model of associative learning that accounts for cue interactions (see Discussion for details). The RW model has been validated extensively, using behavioral data from both humans and animals and can account for many aspects of associative learning (Schultz and Dickinson 2000; Pearce and Bouton 2001). In our study, the trial-wise associative strength predicted by the RW model was used to construct regressors for a voxel-wise general linear model (GLM) of fMRI data and modulatory inputs for dynamic causal models (Friston et al. 2003) of the effective connectivity between auditory and visual areas. Specifically, we addressed the following 2 questions:

1) In the absence of any behavioral responses to the audiovisual stimulus associations, can we obtain neurophysiological evidence that the brain learns these associations? Specifically, can we find brain regions whose activity correlates with learning (throughout the paper, we will use the colloquial term "learning curve" to denote the vector of predicted associative strength over time, i.e., ϕ_t^j in eq. 1.) predicted by a generic model of associative learning (i.e., the RW model)? Candidate areas included early visual cortex and the striatum. Furthermore, do these areas show a response profile across cue–outcome combinations that reflects a match between prediction and outcome or rather a prediction-error response?

2) Because the predictive auditory cue temporally precedes the visual outcome, learning should modify neuronal activity in early visual cortex in response to auditory cues. Can these putative learning-related changes in visual cortex activity be explained by changes in the effective connectivity from auditory to visual cortex (c.f., (McLaren et al. 1989; McIntosh et al. 1998)? Specifically, do these changes conform to changes in associative strength under a RW model of learning?

Before describing our experiment, 2 important issues should be highlighted. First, the goal of this fMRI study was not to pinpoint the exact mathematical form of incidental learning by comparing different models of associative learning. Instead, we used the simplest (i.e., the RW) model of associative learning that could accommodate our paradigm. In the Discussion, we argue why the RW can be considered an appropriate *a priori* learning model for our particular paradigm, relative to other models of associative learning. Second, it is important to note that *within* a given experimental condition the predicted outcomes and prediction errors are perfectly anticorrelated (see Supplementary Material for details). This means they cannot be distinguished as alternative predictors of observed brain responses. However, with our factorial design one can analyze the pattern of parameter estimates *across* experimental conditions, contrasting expected and unexpected cue–outcome combinations. This enabled us to distinguish, voxel by voxel, brain responses that reflected a match between predicted and actual trial outcomes from responses that encode prediction error or surprise.

Methods and Materials

Subjects

Sixteen healthy volunteers, 25.3 ± 3.3 years of age, (mean age ± SD, 8 female) participated in the study. The subjects had no history of psychiatric or neurological disorders. Written informed consent

was obtained from all volunteers prior to the study, which was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee.

Experimental Design—fMRI

The central idea of this study was to present subjects with "distractor" stimuli that were linked by predictive associations: 2 auditory stimuli served as CS and differentially predicted whether or not a visual stimulus would follow. Critically, the volunteers performed an unrelated detection task on separate auditory and visual targets; for this task, the predictive relationships between the distractor stimuli were completely irrelevant. Stimuli were presented using Cogent2000 (www.vislab.ucl.ac.uk/Cogent/index.html). An initial sound matching task and the subsequent learning study (4 × 10 min) were all completed inside the scanner. Subjects were debriefed with a postscan questionnaire to assess whether they had learned the experimental contingencies.

Sound Matching

Preceding the learning experiment, subjects had to match the 2 CS (450 and 1000 Hz) and the auditory target stimulus (white noise burst) for perceived loudness. Stimuli were presented sequentially and dichotically. Subjects adapted the volume of the 1000-Hz tone to the 450-Hz tone until they perceived them to be of equal loudness. This procedure was repeated 8 times and the results averaged. Subsequently, subjects matched the perceived loudness of the white noise burst to the pure tones, each repeated 4 times. The adapted volumes, as a percentage of the volume of the low tone were 94.0 ± 6.2% (mean ± SD) for the high tone, and 104 ± 4.9% for the white noise burst.

Differential Conditioning

During the experiment, subjects were exposed to alternating blocks of trials in which one of 2 auditory CS (high and low tone) predicted the presence (CS^+) or omission (CS^-) of a subsequent visual stimulus with a fixed probability of 80% (Fig. 1 and 2). On each trial, a CS was presented (A^+) with 50% probability. On 50% of all trials, a visual stimulus was present (V^+). Every trial was preceded by a visual TO cue.

Our paradigm thus used a 4-factor design with the following factors for each trial: 1) *CS context* (CS^+ vs. CS^-), 2) *CS presence* (A^+ vs. A^-), 3) *visual outcome* (V^+ vs. V^-), and 4) *learning* (or time). We used a mixed event and epoch design in which CS type was blocked, whereas the presentation of the CS and visual outcome were randomized (event-related) within blocks. CS^+ and CS^- blocks were completely balanced so that in each block of 10 trials 5 CS and 5 visual stimuli were presented. Within each subject, the auditory CS^+ and CS^- and their probabilistic relation to subsequent visual stimuli were fixed throughout the experiment. The assignment of tones to the 2 CS was counterbalanced across subjects, that is, in half the subjects the high tone served as CS^+ (and the low tone as CS^-), and vice versa the other half of the subjects. Each of the 4 sessions consisted of 20 blocks of 10 trials, interspersed with periods of rest (12 s), in which subjects fixated on a fixation cross. Blocks and sessions were balanced across and within subjects.

Target-Detection Task

To ensure continuous attention to auditory and visual targets per se (but not their statistical associations), subjects performed a concurrent target-detection task. The target stimuli were randomly interspersed between trials and consisted of either a white noise burst or a circle. Target stimuli occurred on average once per block (at most 2 times). In total, 40 auditory and 40 visual target stimuli were presented, randomized within conditions and sessions.

fMRI Data Acquisition

A 3 Tesla Siemens Allegra MRI scanner (Siemens, Erlangen, Germany) was used to acquire T_1-weighted fast-field echo structural images and multislice T_2^*-weighted echo-planar volumes with BOLD contrast (time repetition = 2.08 s). For each subject, functional data were acquired in 4 scanning sessions of approximately 10 min each. 306 volumes were acquired per session (1224 scans in total per subject). The first 6 volumes of each session were discarded to allow for T_1 equilibrium effects. Each functional brain volume comprised 34 2-mm axial slices with a 2-mm interslice gap, and an in-plane resolution of 3 × 3 mm. The field of view covered the whole brain, except for the cerebellum and brainstem. The total duration of the experiment was approximately 60 min per subject.

Data Analysis

Functional Neuroimaging Analysis

fMRI data were analyzed using the statistical software packaged SPM5 (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). The 1200 images from each subject were realigned to correct for head movements, corrected for movement-by-distortion interactions (Anderson et al. 2001), spatially normalized to the Montreal Neurological Institute (MNI) template brain, smoothed spatially with a 3-dimensional Gaussian kernel of 8-mm full width half maximum and resampled to 3 × 3 × 3 mm voxels. The data were then modeled voxel-wise, using a GLM that included regressors for all experimental trials as well as regressors for the target-detection task. Trial-specific effects were modeled by trains of delta functions convolved with 3 hemodynamic basis functions (a canonical hemodynamic response function, and its temporal and dispersion derivatives). Additionally, the time-dependent associative strengths from the RW model ($\phi^j_{i,t}$; see eq. 1) and their partial derivatives with respect to learning rate (see next section) were used as parametric modulators of each trial-specific regressor. The data were high-pass filtered (cut-off 128 s) to remove low-frequency signal drifts, and a first-order autoregressive model was used to model the remaining serial correlations (Friston et al. 2002). Contrast images of parameter estimates encoding trial-specific effects were created for each subject and entered separately into voxel-wise one-sample *t*-tests (df = 15), to implement a second-level random effects analysis. We report regions that survive cluster-level correction for multiple comparisons (family-wise error, FWE) across the whole brain at $P < 0.05$. Because previous studies demonstrated the role of the striatum and the prefrontal cortex in associative learning (e.g., Fletcher et al. 2001; O'Doherty et al. 2004; Corlett et al. 2004), we performed an additional restricted search in these areas, using anatomical masks generated from the PickAtlas toolbox (Maldjian et al. 2003). Again, we only report activations that survived a small volume correction (SVC) at $P < 0.05$.

RW Model

We used a RW model of associative learning to generate predictors of learning-dependent changes in brain activity (as indexed by the BOLD signal) and inter-regional connectivity over time. The basic principle of this model is that the size of the trial-specific prediction error, that is, the degree of surprise incurred by an event, determines the change in associative strength. From the train of observed events a learning curve was computed and fitted to the fMRI data. Trial-specific cueing was modeled by means of 2 separate components (see Fig. 1): the visual TO cue, which was present on every trial and the auditory CS per se, which was present on half the trials. This allowed us to model learning effects on trials where no CS was present. In the RW framework, the predicted outcome on trial t, ϕ^j_t, is the sum of the associative strengths of each cue component:

$$\phi^j_{i,t+1} = \phi^j_{i,t} + \varepsilon_i\left(\lambda_t - \phi^j_t\right) \times u_{i,t} \quad (1)$$

where

$$\phi^j_t = \sum_i \phi^j_{i,t} \times u_{i,t} \quad (2)$$

On each trial t, equation (1) is calculated separately for each cue component, indexed by i (i.e., the auditory CS, and TO), whereas $u_{i,t}$ indexes which of the cue components is actually present on trial t (see the Supplementary Material). λ_t indicates the actual outcome at trial t, being 1 for V^+ and 0 for V^-; ε_i is the learning rate that determines how strongly the prediction error affects the update of the prediction. Separate components are summed in equation (2), where ϕ^j_t is the summed prediction of whether a visual stimulus will be presented at trial t, and j indexes whether this is a CS^+ or CS^- trial. (When considered for a single cue per trial, eq. 1 can also be seen as a simple model of Hebbian or associative plasticity. In this context, $\phi^j_{i,t}$ encodes the

associative strength, which changes according to the second term in eq. 1. This associative term comprises a (presynaptic) input $u_{i,t}$ encoding the outcome on any trial, and a (postsynaptic) prediction error.)

A challenge when applying the RW model to our experiment was to determine an appropriate learning rate. In principle this could be done by fitting the model to behavioral data and using the resulting learning rate to construct regressors for the fMRI analysis. However, our experimental design deliberately precluded behavioral responses; instead, learning could only be assessed neurophysiologically in terms of changes in cortical activity and inter-regional connectivity. Alternative strategies are to choose the learning rate based on principled considerations (e.g., O'Doherty et al. 2004) or using model comparison (Gläscher and Büchel 2005). Because we knew from a previous study that learning should occur in the visual cortex (McIntosh et al. 1998), we adopted the approach by Gläscher and Büchel (2005) of optimizing the value of ε_i to best explain putative learning-induced responses within the main area of interest, the visual cortex. Given our volunteers did not notice the statistical associations (and thus learning was presumably slow) and given that another study of perceptual association learning showed small learning rates ε_{CS} below 0.1 (Gläscher and Büchel 2005), we tested the following values of ε_{CS} in separate models: 0.01, 0.025, 0.05, 0.075, 0.1. We found that $\varepsilon_{CS} = 0.075$ gave the best fit to the data in primary visual cortex for the main contrast of interest (i.e., the 4-way interaction in a random effects second-level analysis); this learning rate was then used for further analysis across the entire brain and for the connectivity analyses described below. Importantly, we used a first-order Taylor expansion around the learning rate $\varepsilon_{CS} = 0.075$ to make the model less dependent on the particular choice of learning rate and to account for intersubject variability in the shape of the learning curves. This was implemented by including the partial derivative of the learning curve ϕ_t^j with respect to the learning rate ε_i as an additional parametric modulator in the GLM for the fMRI data.

These analyses assumed that the optimal learning rate was identical for CS^+ or CS^- trials. In additional analyses suggested by our reviewers, we tested this assumption. We examined whether 1) a selective decrease of the learning rate for CS^- trials improved our ability to detect learning effects during this trial type, and, more generally, whether 2) trial-type specific tests of the partial derivatives indicated a learning rate that was different from $\varepsilon_{CS} = 0.075$. As detailed in the Supplementary Material, neither of these analyses provided any evidence for a differential learning rate over stimuli or regions.

Because of its short duration and small size, the TO cue is less salient than the CS. Because in the RW model the learning rate reflects stimulus properties including salience (Rescorla and Wagner 1972), ε_{TO} can be assumed to be considerably smaller than ε_{CS}. In this study ε_{TO} was assumed to be 4 times smaller than the ε_{CS}. It should be noted that violations of this assumption are unlikely to have a dramatic effect because the inclusion of the derivatives enables the model to cope with deviations from the assumed learning rates (see above). The resulting learning curves are shown in Figure 3 (see Supplementary Fig. 1*A* for a breakdown of the learning curves with regard to the 2 cue components).

Statistical Analysis of Learning Effects

In our factorial design, learning is reflected by time-evolving, context-dependent brain responses to visual stimuli. Specifically, over time, learning should change how differential brain responses to visual stimuli depend on the presence of an auditory CS and whether it is presented in a CS^+ or CS^- context. Furthermore, the emergence of differential responses should follow the time-course predicted by the RW model. In other words, learning is expressed as a 4-way interaction *CS type* × *CS presence* × *visual outcome* × *RW learning*. (Note that when the *CS* is absent on a specific trial, this trial can be assigned unambiguously to the CS^+ or CS^- factor because this factor was blocked.) The primary goal of our GLM analyses was therefore to test this interaction. To establish which CS was driving this interaction, we also tested, the simple (3-way) interactions *CS presence* × *visual outcome* × *RW learning* within each CS type. Finally, to test for responses reflecting the prediction (ϕ_t^j) entailed by the auditory CS, independently of the prediction error ($\lambda_t - \phi_t^j$) elicited by the visual outcome, we tested the simple 3-way interaction *CS type* × *CS presence* × *RW learning*, which is independent of visual outcome.

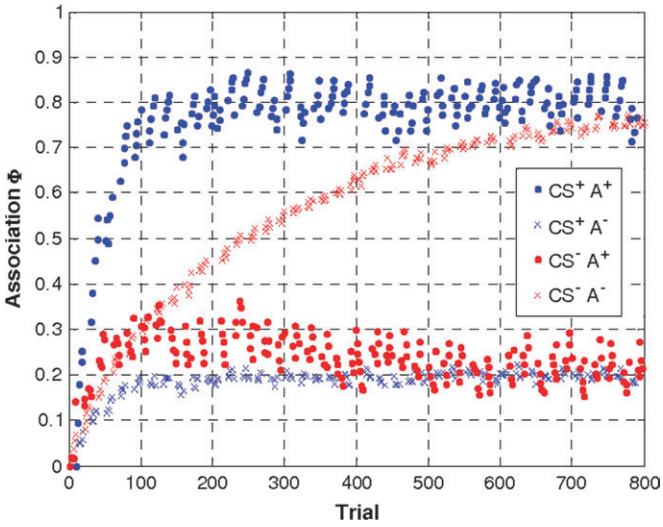

Figure 3. Compound learning curves. Learning curves were calculated separately for trials on which the auditory CS was present (dots) and absent (crosses), during CS^+ (blue), and CS^- (red) blocks. Note that learning is slower in the absence of an auditory CS than in its presence and faster for CS^+ than for CS^- trials.

An important feature of our factorial design is that it enabled us to determine whether the responses of a particular brain region reflected the prediction of the visual target or the prediction error. This is important because one cannot include separate regressors based on predictions and prediction errors in the same design matrix. This is due to the form of the RW equation, in which predictions and prediction errors are perfectly correlated (*within* a given experimental condition), after mean-correction (see Supplementary Materials for details). However, in a factorial design like ours such a distinction can be made by analyzing the pattern of parameter estimates *across* conditions, contrasting conditions that correspond to expected and unexpected cue–outcome combinations. Specifically, our factorial design provided us, in a mirror-symmetric fashion, with 2 expected outcomes and 2 unexpected outcomes for each CS type. For example, on CS^+ trials, A^+V^+ and A^-V^- trials represented expected cue–outcome combinations (conditional probability = 80%) whereas A^+V^- and A^-V^+ trials consisted of unexpected cue–outcome combinations (conditional probability = 20%); c.f. Figure 2. This means one can effectively compare expected and unexpected trials (with low and high prediction error, respectively), with a contrast that is orthogonal to the presence or absence of the visual outcome and its prediction. This enabled us to distinguish, voxel by voxel, brain responses that reflected *expected visual outcomes* from those that represented *unexpected* or *surprising outcomes*. During learning, brain regions encoding prediction errors should show increasing activation on trials where the outcome was unexpected according to the learned contingencies and decreasing (or nonchanging) activation on trials where the outcome was expected. We will call such an activation pattern a "prediction-error response"; this activation pattern would be expected if surprise was the driving force for learning. In this case, surprising events, or prediction errors, signal the need for learning in order to update predictions. This idea is not only a core component of associative learning models (Shanks 1995; Schultz and Dickinson 2000), but is also central to predictive coding theories of perception (Rao and Ballard 1999; Friston 2005): that the brain should concentrate resources on representing surprising sensory events.

Note that our factorial analysis was not geared towards detecting prediction-error responses only. It was equally capable of finding opposite activation patterns, that is, increasing activation on trials where the prediction based on the learned contingencies matched the outcome, and decreasing (or nonchanging) activation on trials where

the prediction did not match the outcome (c.f. Baier et al. 2006). Notably, for our particular design, both types of responses could be identified by the same statistical test, that is, the 4-way interaction *CS type* × *CS presence* × *visual outcome* × *learning* (see above). Because it is only the direction of the interaction that differs between the 2 types of responses, our factorial design enabled an analysis that simultaneously tested for these 2 aspects of associative learning.

Dynamic Causal Modeling

In DCM, the states of multiple interacting brain regions are modeled as a set of coupled bilinear differential equations (Friston et al. 2003). The neuronal states, which represent the neuronal population activity of the modeled brain regions, change in time according to the system's connectivity and experimentally controlled inputs u. These inputs can enter the model in 2 different ways; they can either elicit responses through direct influences on specific regions ("driving inputs," e.g., sensory inputs) or they can change the strength of connections between regions ("modulatory inputs," e.g., task effects or learning). The hidden neural dynamics (i.e., not directly observed by fMRI) are modeled by the following bilinear differential equation:

$$\frac{dz}{dt} = \left(A + \sum_{j=1}^{m} u_j B^{(j)}\right) z + Cu \qquad (3)$$

Here, z is the state vector (with each state variable representing the population activity of one region in the model, in this study the auditory and visual cortex), t is continuous time, and u_j is the j-th input to the modeled system (here the stimuli and learning curve). In this state equation, the A matrix represents the fixed (endogenous) strength of connections between regions and the $B^{(1)} \ldots B^{(m)}$ matrices represent the modulation of these connections by (exogenous) inputs (in this case, learning), as an additive change. Finally, the C matrix represents the influence of exogenous inputs on each area (here the auditory and visual stimuli). Note that DCM allows one to make inferences about changes in effective connections between areas, which do not necessarily correspond to direct anatomical connections but may be via intermediary regions.

In DCM, the hidden neuronal dynamics described by equation (3) is linked to predicted BOLD responses by a hemodynamic forward model (Friston et al. 2003). Given measured BOLD responses, maximum a posterior estimates of the parameters in equation (3) can be obtained through an optimization scheme based on variational Bayes (Friston et al. 2003).

Choice of areas and time series extraction. The goal of the present DCM analysis was to explain the (3-way) simple interaction *CS presence* × *visual outcome* × *RW learning* for CS^+ trials in V1 (see SPM findings in the Results section) by a simple model, in which the strength of the A1 → V1 connection was modulated as a function of the RW predictions, ϕ_t^j (i.e., learning curves; Fig. 3). Representative A1 time series were chosen by testing for the main effect of CS presence, and V1 time series were selected by testing for the simple interaction described above. (The goal of DCM is to explain regional effects [as detected in a voxel-wise GLM analysis] in terms of interregional connectivity and its experimentally induced changes. This puts congruence constraints on the contrast used to identify a regional time series and the mechanisms in a DCM that are proposed to model this time series. Therefore, different contrasts are typically required for selecting time series representing the different areas in a model; c.f. Stephan, Harrison, et al. 2007.) We did not model the 4-way interaction with DCM because the SPM analysis showed that the learning effect was driven by the CS^+ (see Results section).

As the exact locations of activation maxima varied over subjects, we ensured the comparability of our models across subjects by using combined anatomical–functional constraints in selecting the subject-specific time series (c.f. Stephan, Marshall, et al. 2007). Specifically, we thresholded the subject-specific SPMs at $P < 0.05$ and chose the local maximum within 8 mm of the group activation maxima in primary auditory cortex (A1) and primary visual cortex (V1) as inferred by a probabilistic cytoarchitectonic atlas in MNI space (Eickhoff et al. 2005). As a summary time series, we computed the first eigenvector across all suprathreshold voxels within a radius of 4 mm around the chosen local maximum. Overall, we were able to extract time series in 14 out of 16 subjects. In 2 subjects, V1 could not be defined due to the lack of a significant interaction that met the anatomical and functional criteria described above. These 2 subjects were excluded from the DCM analysis.

DCM specification. The question addressed by DCM was whether learning effects in V1 could be explained by changes in the connectivity of a simple auditory–visual network. Our DCMs modeled the entire time series, so data from all trials or conditions, trying to explain regional activations by condition-dependent changes in connectivity. We tested 3 simple models that could potentially account for the interaction we found in V1. These models were fitted separately to each subject's data and compared using Bayesian model selection (Penny et al. 2004). In these models, auditory and visual stimuli from all trials elicited activity directly in their respective primary sensory areas (see Fig. 4). These driving inputs were modeled as individual events. The first model only had a connection from A1 to V1, whereas the second and third models included the reciprocal connection (see Fig. 5). The A1 → V1 connection in model 1 and 2, and the V1 → A1 connection in model 3 were modulated by the Hadamard product (point-wise multiplication) of the RW associative strength ϕ_t^j and a vector encoding visual outcome (1 for visual stimulus present, −1 for visual stimulus absent) during CS^+ trials. In the first 2 models, this modulatory effect corresponds to the interaction of the auditory CS^+ prediction with the visual outcome and models a learning-dependent contribution from CS^+ responses in auditory cortex to visual cortex responses that depends on whether the visual stimulus was present or not (c.f., a prediction error that rests on top-down signals from auditory areas). In the third model, which represented a control suggested by

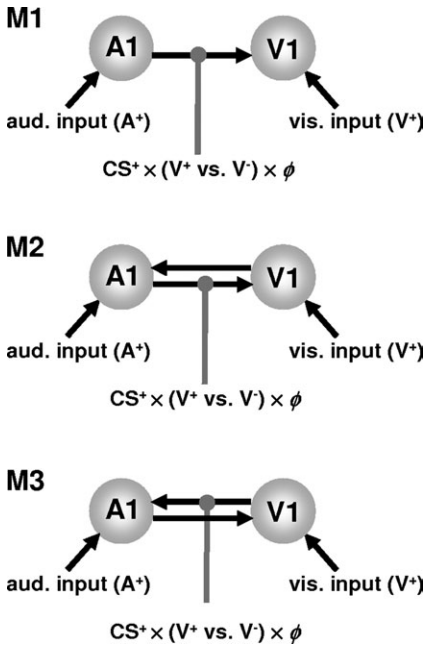

Figure 4. Dynamic causal models of learning effects on audio-visual connectivity. For all 3 models, the primary auditory (A1) and visual (V1) areas are both driven by their respective sensory inputs. The first model tested had a single connection from A1 to V1 (M1). In model 2 (M2) the V1 → A1 connection was added. In both M1 and M2, the A1 → V1 connection was allowed to change during CS^+ trials as a function of the visual outcome (V^+ vs. V^-) and the RW learning curve (ϕ). This modulatory effect corresponds to the interaction of the auditory CS^+ prediction with the visual outcome and models a learning-dependent contribution to V1 responses from CS^+ responses in A1; and this contribution depends on whether the visual stimulus was present or not (c.f., a prediction error mediated by top-down signals from A1). In the third model, suggested as a control by one of the reviewers, instead of the A1 → V1 connection, the V1 → A1 connection is modulated by the learning signal.

one of our reviewers, this modulatory effect acted on the reverse connection, V1→A1.

Results

The postscan debriefing questionnaire showed that none of the subjects had become aware of the contingencies between the auditory and visual stimuli. Prior to the fMRI data analysis we verified subjects' performance on the target-detection task. On average, subjects responded to 93 ± 3% of the target stimuli. Following Gläscher and Büchel (2005) we determined an optimal learning rate for the RW model, evaluating the primary contrast of interest (i.e., the 4-way interaction in a random effects second-level analysis) under different learning rates in the primary visual cortex (as defined by a probabilistic cytoarchitectonic atlas (Eickhoff et al. 2005). Model fits under 5 different learning rates, suggested $\varepsilon_{CS} = 0.075$ was the optimal learning rate (see Fig. 3 and Methods section for details).

Statistical Parametric Mapping

First, we examined the 4-way interaction *CS type* × *CS presence* × *visual outcome* × *RW learning*. We found learning-dependent responses in the primary visual cortex and putamen that survived whole-brain correction for multiple comparisons (see Fig. 5*A*,*B*). To characterize the nature of this interaction, we tested the simple interaction (*CS presence* × *visual outcome* × *RW learning*) within each CS type. This showed that the 4-way interaction was driven mainly by learning during the CS^+ blocks (see Supplementary Fig. 1*B* for the parameter estimates). As shown in Figure 5*A*,*B*, testing the simple interaction for CS^+ trials afforded almost identical results in the visual cortex and the putamen as the 4-way interaction (see also Table 1). In contrast, no evidence of learning, that is, no significant interaction of CS presence and outcome with learning, was found for CS^- trials.

The nature of the simple 3-way interaction was such that V1 and the putamen showed an increased response when an expected visual stimulus was omitted, or when an unexpected visual stimulus was presented (i.e., A^+V^- and A^-V^+ trials). Critically, this response to surprising visual outcomes increased over time as the association was learned, following the form of the RW learning curve. Conversely, V1 responses to predicted stimuli diminished during learning. The putamen showed the same pattern of responses bilaterally; this activation extended into the insula bilaterally (see Table 1).

Because previous studies have implicated the right DLPFC in prediction (error) processing (Fletcher et al. 2001; Corlett et al. 2004), we used an anatomically defined fronto-striatal mask to test the 3-way interaction *CS type* × *CS presence* × *RW learning*, which characterizes responses to the prediction entailed by the auditory CS, independent of the visual outcome. During learning, the right DLPFC became increasingly active when a visual stimulus was predicted compared to when it was not; activity was higher for CS^+A^+ and CS^-A^- trials compared with CS^+A^- and CS^-A^+ trials (compare the probabilities in Fig. 2). As above, we characterized the nature of the 3-way interaction by testing the associated simple interactions, confirming it was also driven by CS^+ trials (Fig. 4*C*). The same pattern of activation was found in the left putamen, but this activation did not survive correction for multiple comparisons.

Learning-Dependent Changes in Connectivity

Because the learning effect was mainly driven under CS^+ blocks, we focused on changes in connectivity between auditory and visual cortices during incidental learning of the predictive attributes of CS^+ trials (see Fig. 6). Bayesian model comparison showed that a DCM with a single connection from A1 to V1 (model 1) was superior to alternative models with reciprocal connections (group Bayes factor in favor of model 1: 2.1×10^{17} and 2.2×10^{18} when compared with model 2 and model 3, respectively). Across subjects, the A1 → V1 connection in the optimum model had an average strength of 0.10 s^{-1} ($p = 0.003$, df = 13, $t = 3.57$). During CS^+ trials, this connection was significantly modulated by learning, depending on whether the visual stimulus was present or not (i.e., $CS^+ \times (V^+ \text{ vs. } V^-) \times \phi$ in Fig. 6). Note that the modulatory variable in the DCM corresponds to the interaction of the auditory prediction with the visual outcome during CS^+ trials. It accounts for a learning-dependent contribution from CS^+ responses in auditory cortex to visual cortex responses that depends on whether the visual stimulus was present or not (c.f., a prediction error mediated by top-down signals from auditory areas). Quantitatively, the strength of this modulation was –0.01 s^{-1} ($p = 0.028$, df = 13, $t = 2.49$). This corresponds to learning-induced changes in connectivity ranging from 2% (for

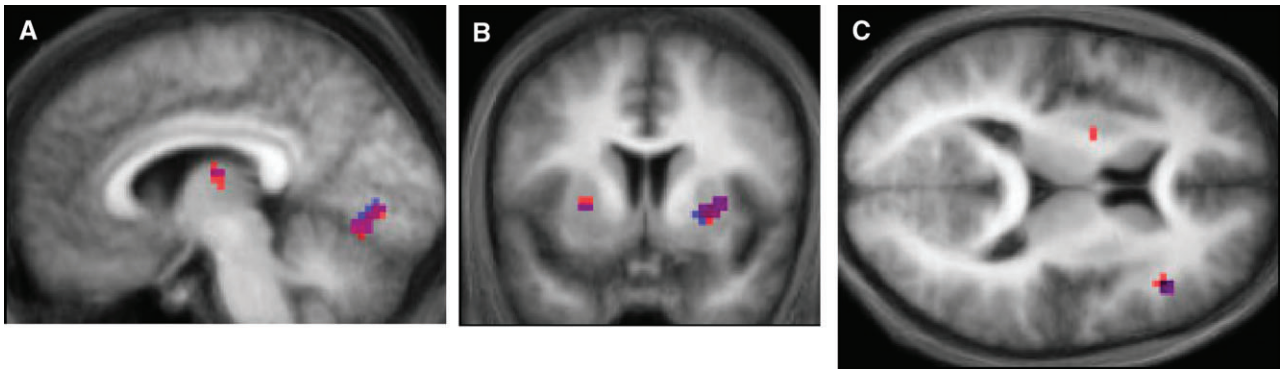

Figure 5. fMRI results. (A) Significant activations in V1 as a function of RW learning, for both the 4-way interaction (*CS type* × *CS presence* × *visual outcome* × *RW learning*; red), and the simple (3-way) interaction (blue), which is restricted to the CS^+ trials ($x = -6$, also showing the caudate activation) and (B) in the putamen bilaterally ($y = 6$), displayed on the mean structural image across all subjects. (*C*) $z = 12$. Significant 3-way interaction *CS type* × *CS presence* × *RW learning* in the DLPFC and left putamen (red). This interaction is driven by the CS^+ trials, as shown by the simple interaction in blue.

Table 1
MNI coordinates and *Z*-values for significantly activated regions

Foci of activation	MNI coordinates			*Z* value	Cluster size
	x	*y*	*z*		
Four-way interaction: CS type × CS presence × visual outcome × RW learning					
L occipital lobe*	−6	−75	−9	4.25	41
L insula and putamen*	−30	18	6	4.84	84
L putamen**	−24	12	6	3.85	20
R insula and putamen*	36	12	3	4.72	82
R putamen**	27	6	−3	4.48	35
L caudate/thalamus*	−9	−15	15	4.70	40
L SII cortex*	−51	−27	24	4.39	93
L middle temporal gyrus*	−57	−39	−3	3.88	26
Simple (3-way) interaction: CS presence × visual outcome × RW learning (restricted to CS^+)					
L occipital lobe*	−9	−78	−3	4.31	36
L insula and putamen*	−33	12	3	4.55	57
L putamen**	−27	12	6	3.63	10
R insula and putamen*	36	12	3	3.98	57
R putamen**	27	9	0	3.94	32
L caudate/thalamus*	−21	−9	9	4.32	54
L caudate**	−15	−9	21	4.19	14
R caudate**	15	12	18	4.24	7
L SII cortex*	−60	−33	15	4.15	87
L middle temporal gyrus*	−57	−36	−6	4.30	34
R posterior insula*	39	12	−12	5.01	38
Three-way interaction: CS type × CS presence × RW learning					
R inferior frontal gyrus**	42	27	12	4.39	10

*Significant at $P < 0.05$ (FWE whole-brain cluster-level corrected).
**Significant at $P < 0.05$ (SVC).

CS^+A^- trials) to 8% (for CS^+A^+ trials) (Fig. 6). (As shown by eq. 3, the overall strength of a connection, given a single modulatory parameter, is the sum of the intrinsic connection strength [*A*] and the modulatory parameter [*B*] multiplied with its associated input [*u*]. In the present case, the asymptotic magnitude of the input function is 0.8 for CS^+A^+ trials and 0.2 for CS^+A^- trials [see Fig. 5].)

Critically, the negative sign of the modulatory parameter reflects the nature of the visual responses to auditory afferents under CS^+ trials: V1 responses to predicted visual stimuli diminished during learning and the DCM explained this through a decrease in the strength of the A1 → V1 connection. This is exactly consistent with an increase in the "explaining away" of predicted visual input under predictive coding; in other words, if top-down predictions ϕ_t^j (see eq. 2) from auditory cues decrease the amplitude of V1 prediction error $|\lambda_t - \phi_t^j|$, a better prediction corresponds to a decrease in effective connectivity. Conversely, V1 responses to unpredicted (i.e., absent) visual stimuli increased during learning. This was modeled in the DCM through an increase in the A1 → V1 connection strength; again this is consistent with an increase in V1 prediction-error amplitude $|\lambda_t - \phi_t^j|$, when predictions are violated. In summary, A1 → V1 influences depended on whether the visual outcome was expected or surprising and were consistent with an "explaining away" role. The emergence of this effect conformed to the learning curve provided by the RW model.

Discussion

McIntosh and colleagues showed that after a predictive relationship between an auditory stimulus and a visual stimulus had been learned, the auditory stimulus alone was able to evoke responses in the visual cortex (McIntosh et al. 1998). The current study extended this work, pairing a visual stimulus with a predictive auditory stimulus in a 4-factorial design, with the factors CS type (CS^+, CS^-), CS presence (A^+, A^-), visual stimulus presence (V^+, V^-), and learning (over time). Both CS^+ and CS^- blocks were exactly balanced in terms of sensory stimulation, so that the *a priori* probabilities of the auditory CS and of the visual stimulus occurring on a given trial were always 50%. Critically, the volunteers did not make any responses to the stimuli whose associations were being learned; instead, they performed a target-detection task on unrelated stimuli. Our factorial design enabled us 1) to characterize changes in neurophysiological responses due to learned associations that were incidental to behavior, and 2) to investigate whether activity in specific brain areas, and the connection strengths amongst them, reflected a match between predictions and outcome or prediction errors, respectively.

Our results demonstrate that during incidental learning of audio-visual associations changes in both regional activity and underlying connectivity reflect prediction errors. Furthermore, we show that learning-dependent responses in visual cortex can be elicited, even in the absence of visual stimuli. This finding can be explained by changes in top-down influences from auditory regions that are consistent with predictive coding models of perceptual inference.

RW Model: Predictions and Prediction Error

The goal of this study was not to pinpoint the exact mathematical form of learning by comparing different models of associative learning. Instead, we focused on changes in regional activity and interregional connectivity that could be explained by a specific learning model, namely the RW model. The RW model is a generic and well-established model of associative learning that has been successful in modeling a wide range of learning processes (Rescorla and Wagner 1972; Schultz and Dickinson 2000; Pearce and Bouton 2001). We chose this model because it is the simplest learning model appropriate for our particular paradigm. In the absence of interactions among multiple cues per trial, the RW model is mathematically equivalent to a Hebbian model of associative learning (Montague and Berns 2002). A crucial aspect of our paradigm, however, is that on each trial the net prediction resulting from 2 interacting cue components (the auditory CS and the visual TO cue) must be considered (see Methods sections for details). This excludes the use of any associative learning model that cannot accommodate cue interactions (e.g., Hebbian models). In contrast, the RW model accommodates this aspect gracefully. Another learning model, TD learning, can also deal with multiple cues and their temporal relationships; however, under our design with temporally overlapping cue and outcome, the TD model is effectively equivalent to the simpler RW model. Finally, the associative learning models of Pearce and Hall (1980) and Mackintosh (1975) assume that prediction errors affect the amount of attention that is allocated to stimuli and that the more attention is allocated to a specific stimulus, the more strongly it becomes associated with an outcome or reinforcer. This is not relevant to our experimental paradigm in which attention is actively directed away from the stimuli whose associations are learned.

The RW model has one problematic limitation, however: as detailed in the supplementary materials, its equation uses both predictions and prediction errors that are perfectly correlated under mean-correction. In situations where mean-correction is mandatory (e.g., when using them to form interaction terms) this makes it impossible to disambiguate/interpret their contributions to a dependent variable. However, the factorial

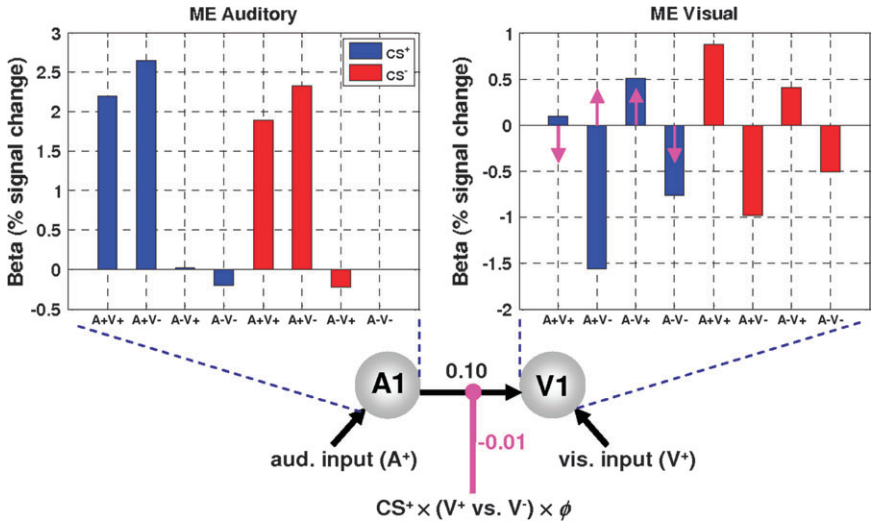

Figure 6. Learning effects on audio-visual connectivity. Bayesian model comparison showed that the DCM with a single connection from A1 to V1 was superior to the other models. Across subjects, there was a significant "endogenous" or "fixed" strength of the A1 → V1 connection (0.10 s^{-1}, $P = 0.003$) and a significant learning-induced modulation (magenta arrows) of this connection ($P = 0.028$). The insets show the parameter estimates for the main effects in both A1 and peripheral V1. The magenta arrows indicate how the main effect in peripheral V1 is modulated by changes in connectivity from A1 to V1 during CS^+ trials: over time the response to surprising visual outcomes is upregulated, whereas the response to unsurprising visual outcomes is downregulated. Note that in this plot the magenta arrows designate the direction in which V1 responses change due to modulation of connectivity; for quantitative information on this modulatory effect, see the main text.

design in our study allows us to circumvent this problem, as it comprises conditions that correspond to congruent and incongruent prediction/outcome combinations, respectively. Analyzing the 4-way interaction between our experimental factors, we found that responses in the primary visual cortex and the putamen were sensitive to surprising events; over time, these areas became significantly more active when presented with a surprising cue–outcome combination. Learning was stronger for the CS^+ blocks than for the CS^- blocks, which is in line with previous behavioral evidence (Wasserman et al. 1993; Fletcher et al. 2001). Previous fMRI studies in humans have demonstrated that BOLD activity in the striatum is correlated with (signed) prediction errors during reinforcement learning (O'Doherty et al. 2003; McClure et al. 2003; O'Doherty et al. 2004; Seymour et al. 2004; Jensen et al. 2007; Menon et al. 2007) and other associative learning tasks (Corlett et al. 2004). In these studies, the learned associations, and the sign of the resulting prediction errors, were of direct relevance for behavior. The current study shows that the putamen is sensitive to unexpected outcomes even when the cue-stimulus association is learned incidentally and has no relevance to behavior. However, in contrast to the previous studies, the pattern of putamen activity does not appear to be sensitive to the direction of the prediction error, only to its amplitude. This difference may reflect the fact that learning was perceptual as opposed to operant. In other words, the occurrence of an unpredicted or surprising event may play the role of negative reward, irrespective of whether the surprising event entailed the presence of absence of a stimulus. This issue will be discussed further in the section on predictive coding below.

Role of Prediction Errors Beyond Reinforcement Learning

Our finding that learning-induced responses in primary visual cortex and the putamen reflected prediction errors accords with a basic principle emerging from many previous studies: prediction errors, or surprise, constitute a driving force for learning because they signal the need for learning in order to update predictions (Shanks 1995; Schultz et al. 1997; Schultz and Dickinson 2000). Although the role of prediction errors has been mainly explored for reinforcement learning so far, there is growing evidence that prediction errors may be equally important for learning statistical relationships that are affectively neutral and behaviorally irrelevant. In other words, the same mechanisms that optimize the learning of stimulus–response links may operate during the perceptual learning of stimulus–stimulus associations (Rao and Ballard 1999; Friston 2005). Evidence that organisms learn predictive associations between initially neutral stimuli is seen in classical conditioning effects such as sensory preconditioning (Brogden 1939). Some forms of sensory learning also exhibit such features, for example, the mismatch negativity (MMN) paradigm, in which responses to sensory stimuli decrease with predictability (Friston 2005; Baldeweg 2006), regardless of whether stimuli are attended. A mechanism similar to predictive coding has been proposed in the motor domain for cancellation of self-generated events (Wolpert et al. 1995; Blakemore et al. 1998; Shergill et al. 2005). Moreover, the learning of predictive relationships that are affectively neutral and task-irrelevant may engage similar computational and neural mechanisms as those for predicting significant events (Zink et al. 2006; Wittmann et al. 2007).

The results of the present study support the notion that the role of prediction errors in learning transcends the simple reinforcement of stimulus–response links and plays a more pervasive and general role in various forms of learning. Indeed a hallmark of adaptive systems is their ability to minimize surprising exchanges with their environment (Friston et al. 2006). This entails adjustments to their internal models of the environment so that potentially surprising event can be predicted. Almost universally, this adjustment involves changes

in the system's connections; it is therefore perhaps a little surprising that most previous imaging studies on learning and conditioning have exclusively searched for brain areas whose activity correlated with specific variables of a particular learning model (e.g., prediction or prediction error), but have not investigated how these variables change interactions among areas (but see McIntosh et al. 1998; Büchel et al. 1999). Functional interactions are central to the physiological implementation of learning; it has long been suggested that plasticity in connection strengths between neurons underlies the learning of predictive associations (Hebb 1949). Put simply, 2 neural units encoding associated entities increase their synaptic connections to encode the learned associative strength of the stimuli. More precisely, for RW and similar "caching" models (Daw et al. 2005) the connection strength at time t should carry the predicted association at time t (McLaren et al. 1989; Schultz and Dickinson 2000). This hypothesis requires models of effective connectivity, in which connection strengths vary as a function of the associative strength predicted by the learning model. To our knowledge, the present study has implemented this approach for the first time, modeling how learning, as described by a RW model, modulates the effective connectivity, as assessed by a DCM, between primary auditory and visual areas.

Changes in Connectivity between Auditory and Visual Areas

In accordance with the considerations above, we investigated whether the learning-related changes in visual cortex responses could be explained by a simple model of effective connectivity, in which the strength of A1 → V1 connection changed as a function of the associative strength predicted by the RW model. We modeled observed responses in the primary visual cortex by means of a simple 2-area DCM in which activity in the visual cortex was modeled by 2 components, 1) a direct effect of visual stimulation and 2) a modulation of the A1 → V1 connection by the interaction of the time-evolving prediction with the visual input (in CS^+ blocks; see Fig. 6). Across subjects, this DCM showed a significant change in the strength of the A1 → V1 connection congruent with the pattern of responses in V1: the A1 → V1 connection strength increased on trials where the visual outcome did not match the auditory prediction and decreased on trials where prediction and outcome matched. In other words, the learning-induced changes in A1 → V1 connection strength reflected the same pattern of surprise or prediction errors as the regional activity in V1. This demonstrated that the response of V1 to visual stimuli was modulated by learning-dependent changes in top-down auditory influences that were consistent with the notion of predictive coding, a general framework for perceptual inference and learning that is discussed in the next section (Friston 2005).

Although connections in models of effective connectivity do not need to correspond to monosynaptic anatomical connections, it is of interest to note that the surprise-related response in visual cortex appears to be in the peripheral visual field (Fig. 3*A*), and anatomical connections from primary auditory cortex to peripheral visual cortex have been demonstrated in recent monkey studies (Falchier et al. 2002; Rockland and Ojima 2003). Additionally, numerous fMRI studies have demonstrated that auditory stimulation or auditory attention affect activity in visual cortices during simultaneous processing of visual stimuli (e.g., McIntosh et al. 1998; Baier et al. 2006; Watkins et al. 2006).

Predictive Coding in Visual Cortex

In previous neurophysiological studies of reinforcement learning, a negative prediction error, in the form of unexpected absence of a reinforcer (e.g., a reward), often led to a decrease in neuronal or BOLD activity (Schultz 1998; McClure et al. 2003; Tobler et al. 2007). Such directed excursions are thought to reflect the fact that the prediction error is a signed quantity: it signals not just that predictions need to be updated, but in which direction. In contrast, in our study we found an increase in striatum and visual cortex activity not only for unexpectedly presented stimuli, but also for the unexpected absence of a stimulus. Similarly, the strength of the A1 → V1 connection decreased whenever the visual outcome was expected, and it increased whenever the outcome was surprising.

A useful perspective that explains our 2 main findings, the implicit encoding of surprise by V1 responses and its mediation by learning-dependent changes in input from the auditory cortex, is provided by the framework of predictive coding. Predictive coding posits a hierarchy of connected brain areas in which each level strives to attain a compromise between information about sensory inputs provided by the level below and predictions (or priors) provided by the level above (Rao and Ballard 1999; Murray et al. 2002; Friston 2003; Summerfield et al. 2006). The central learning principle is to establish a good model of the world, which is achieved by changing connection strengths such that prediction errors are minimized at all levels of the hierarchy. The hierarchy of a predictive coding architecture is often defined anatomically (in terms of forward and backward connections) and within one sensory modality, but it is equally possible to examine cross-modal predictive coding relationships (c.f. von Kriegstein and Giraud 2006). In the present study, a temporal hierarchical relation between auditory and visual areas is induced by presenting the auditory cue prior to the visual stimulus.

Predictive coding may be a general principle of brain function in which statistical relationships in the world are monitored, even when they are not attended and not relevant for ongoing behavior. This would allow the brain to ignore predictable and therefore uninteresting events in the environment, thereby enhancing the saliency of unexpected events. A good example of this notion is given by the mismatch negativity (MMN), the difference between the event-related potential to an unexpected "deviant" and predictable "standard" stimuli (Naatanen et al. 2001). Importantly, the relationship between the MMN and learning was not established on the basis of behavioral data; in fact, it was initially not even recognized (Naatanen et al. 1978). This relationship was only subsequently inferred from striking relationships between the probability of deviants and neurophysiological time series (e.g., Csepe et al. 1987; Pincze et al. 2002). Current theories of MMN, which interpret it as a paradigmatic example of learning based on predictive coding (Friston 2005; Baldeweg 2006), have recently received empirical support by DCM studies of electroencephalographic measurements (David et al. 2006; Garrido et al. 2007). These studies demonstrated that MMN can be understood as a prediction-error signal, which results from deviant-induced changes in inter-regional connection strengths. A similar conclusion is offered by the present study. Here, we

found that, at least during CS^+ trials, BOLD responses in area V1 increased when the prediction provided by the auditory cue did not match the subsequent visual stimulus (analogous to MMN elicited by deviants). This surprise signal progressively increased as the predictive properties of the auditory cue were learnt. Moreover, in direct analogy to DCM studies of the MMN (David et al. 2006; Garrido et al. 2007), we found a decrease in the A1 → V1 connection strength on "standard" trials (where the prediction by the auditory cue was correct), and an increase on "deviant" trials where the visual outcome did not match the prediction by the auditory cue. In the context of predictive coding, learning involves a more efficient suppression of sensory events, which is manifest by an apparent reduction in evoked responses, mediated by top-down predictions (which explain away bottom-up sensory afferents). Within the framework of our bilinear DCM, this is modeled as a decrease in top-down effective connectivity for visual stimuli that match the current prediction.

Limitations and Future Directions

We conclude this article by discussing a number of limitations of the present study. First, because we wished to study brain responses to stimulus associations that were irrelevant to behavior, we did not obtain behavioral evidence for learning. Instead, as with the MMN paradigm described above, learning is characterized neurophysiologically as a change in activity over time. We are currently conducting similar experiments with stimuli that do require a behavioral response, providing us with a behavioral assessment of the learning process. It might be useful to emphasize that a neurophysiological characterization of incidental associative learning processes, only requires that the statistical associations between the CS/US stimuli are irrelevant for task performance. In contrast, it is not essential that the CS and US stimuli themselves are behaviorally irrelevant. In fact, in our experiment these stimuli have some behavioral relevance insofar as they constitute distractors to which responses must be suppressed.

A second limitation is that the magnitude of the learning effects (i.e., changes in A1 → V1 connection strength in the range of 2–8%) was rather modest at the single-subject level. This is likely to be due to the incidental nature of the learning in the present study, with attention being directed away from stimulus associations and none of the subjects noticing the contingencies. However, the expression of these learning effects was highly consistent across subjects.

Finally, the dynamic causal model presented here does not make any assumptions about where in the brain the predicted associative strength is calculated; that is, which brain area exerts the modulatory influence onto the A1 → V1 connection. Given the responses that we observed in the putamen, it is possible that the modulation of the A1 → V1 connection is mediated via this region. Testing this hypothesis, however, requires the inclusion of nonlinear terms in the neuronal state equation of DCM which goes beyond its bilinear mathematical framework. However, very recently, there has been methodological progress in nonlinear extensions of DCM (Stephan, Harrison, et al. 2007), and once this approach is firmly established and accepted, it should be possible to investigate the source of the modulatory influences we observed. Notwithstanding this limitation, the current study has presented a novel combination of dynamic system models and formal learning theory, which were used to model human neuroimaging data. This is a further step toward the long-term goal of constructing invertible models that unite the neurophysiological and computational aspects of learning (c.f. Stephan 2004).

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/

Funding

Wellcome Trust (ref: 0856780/Z/99/B); Wellcome Trust PhD studentship (ref: 078047/ZS/04/Z) supported H.D.O.; and University Research Priority Program "Foundations of Human Social Interactions" at the University of Zurich supported K.E.S.

Notes

We thank Quentin Huys for helpful discussions of the manuscript.
Conflicts of Interest: None declared.

Address correspondence to Hanneke den Ouden, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, UK WC1N 3BG. Email: h.denouden@fil.ion.ucl.ac.uk.

References

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

Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, Poldrack RA. 2004. Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. J Neurophysiol. 92:1144–1152.

Baier B, Kleinschmidt A, Muller NG. 2006. Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. J Neurosci. 26: 12260–12265.

Baldeweg T. 2006. Repetition effects to sounds: evidence for predictive coding in the auditory system. Trends Cogn Sci. 10:93–94.

Blakemore SJ, Wolpert DM, Frith CD. 1998. Central cancellation of self-produced tickle sensation. Nat Neurosci. 1:635–640.

Brogden WJ. 1939. Sensory preconditioning. J Exp Psychol. 25:323–332.

Büchel C, Coull JT, Friston KJ. 1999. The predictive value of changes in effective connectivity for human learning. Science. 283:1538–1541.

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.

Csepe V, Karmos G, Molnar M. 1987. Evoked potential correlates of stimulus deviance during wakefulness and sleep in cat—animal model of mismatch negativity. Electroencephalogr Clin Neurophysiol. 66:571–578.

David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ. 2006. Dynamic causal modeling of evoked responses in EEG and MEG. Neuroimage. 30:1255–1272.

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 8:1704–1711.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K. 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage. 25:1325–1335.

Falchier A, Clavagnier S, Barone P, Kennedy H. 2002. Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci. 22:5749–5759.

Fletcher PC, Anderson JM, Shanks DR, Honey R, Carpenter TA, Donovan T, Papadakis N, Bullmore ET. 2001. Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nat Neurosci. 4:1043–1048.

Friston K. 2003. Learning and inference in the brain. Neural Netw. 16: 1325–1352.
Friston K. 2005. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci. 360:815–836.
Friston K, Kilner J, Harrison L. 2006. A free energy principle for the brain. J Physiol Paris. 100:70–87.
Friston KJ, Glaser DE, Henson RN, Kiebel S, Phillips C, Ashburner J. 2002. Classical and Bayesian inference in neuroimaging: applications. Neuroimage. 16:484–512.
Friston KJ, Harrison L, Penny W. 2003. Dynamic causal modelling. Neuroimage. 19:1273–1302.
Garrido MI, Kilner JM, Kiebel SJ, Stephan KE, Friston KJ. 2007. Dynamic causal modelling of evoked potentials: a reproducibility study. Neuroimage. 36:571–580.
Gewirtz JC, Davis M. 2000. Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. Learn Mem. 7:257–266.
Gläscher J, Büchel C. 2005. Formal learning theory dissociates brain regions with different temporal integration. Neuron. 47:295–306.
Hebb DO. 1949. The organisation of behaviour. New York: John Wiley.
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. Hum Brain Mapp. 28:294–302.
Mackintosh NJ. 1975. A theory of attention: variations in the associability of stimulus with reinforcement. Psychol Rev. 82:276–298.
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.
McClure SM, Berns GS, Montague PR. 2003. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 38: 339–346.
McIntosh AR, Cabeza RE, Lobaugh NJ. 1998. Analysis of neural interactions explains the activation of occipital cortex by an auditory stimulus. J Neurophysiol. 80:2790–2796.
McLaren IP, Kaye H, Mackintosh NJ. 1989. An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In: Morris RGM, editor. Parallel distributed processing: implications for psychology and neurobiology. Oxford: Clarendon Press. p. 102–120.
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
Montague PR, Berns GS. 2002. Neural economics and the biological substrates of valuation. Neuron. 36:265–284.
Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL. 2002. Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci USA. 99:15164–15169.
Naatanen R, Gaillard AW, Mantysalo S. 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychol (Amst). 42: 313–329.
Naatanen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I. 2001. Primitive intelligence" in the auditory cortex. Trends Neurosci. 24: 283–288.
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'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.
Pearce JM, Bouton ME. 2001. Theories of associative learning in animals. Annu Rev Psychol. 52:111–139.
Pearce JM, Hall G. 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev. 87:532–552.
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.
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.
Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: current research and theory. New York: Appleton Century Crofts. p. 64–99.
Rockland KS, Ojima H. 2003. Multisensory convergence in calcarine visual areas in macaque monkey. Int J Psychophysiol. 50:19–26.
Schultz W. 1998. Predictive reward signal of dopamine neurons. J Neurophysiol. 80:1–27.
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science. 275:1593–1599.
Schultz W, Dickinson A. 2000. Neuronal coding of prediction errors. Annu Rev Neurosci. 23:473–500.
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.
Shanks DR. 1995. The psychology of associative learning. Cambridge, UK: Cambridge University Press.
Shergill SS, Samson G, Bays PM, Frith CD, Wolpert DM. 2005. Evidence for sensory prediction deficits in schizophrenia. Am J Psychiatry. 162:2384–2386.
Stephan KE. 2004. On the role of general system theory for functional neuroimaging. J Anat. 205:443–470.
Stephan KE, Harrison LM, Kiebel SJ, David O, Penny WD, Friston KJ. 2007. Dynamic causal models of neural system dynamics: current state and future extensions. J Biosci. 32:129–144.
Stephan KE, Marshall JC, Penny WD, Friston KJ, Fink GR. 2007. Interhemispheric integration of visual processing during task-driven lateralization. J Neurosci. 27:3512–3522.
Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J. 2006. Predictive codes for forthcoming perception in the frontal cortex. Science. 314:1311–1314.
Sutton RS, Barto AG. 1998. Reinforcement learning: an introduction. Cambridge (MA): MIT Press.
Tobler PN, O'Doherty JP, Dolan RJ, Schultz W. 2007. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol. 97:1621–1632.
Turner DC, Aitken MR, Shanks DR, Sahakian BJ, Robbins TW, Schwarzbauer C, Fletcher PC. 2004. The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. Cereb Cortex. 14:872–880.
von Kriegstein K, Giraud AL. 2006. Implicit multisensory associations influence voice recognition. PLoS Biol. 4:e326.
Wasserman EA, Elek SM, Chatlosh DL, Baker AG. 1993. Rating causal relations: Role of probability in judgments of response^outcome contingency. J Exp Psychol Learn Mem Cogn. 19:174–188.
Watkins S, Shams L, Tanaka S, Haynes JD, Rees G. 2006. Sound alters activity in human V1 in association with illusory visual perception. Neuroimage. 31:1247–1256.
Wittmann BC, Bunzeck N, Dolan RJ, Duzel E. 2007. Anticipation of novelty recruits reward system and hippocampus while promoting recollection. Neuroimage. 38:194–202.
Wolpert DM, Ghahramani Z, Jordan MI. 1995. An internal model for sensorimotor integration. Science. 269:1880–1882.
Zink CF, Pagnoni G, Chappelow J, Martin-Skurski M, Berns GS. 2006. Human striatal activation reflects degree of stimulus saliency. Neuroimage. 29:977–983.