The Role of the Lateral Frontal Cortex in Causal Associative Learning: Exploring Preventative and Super-learning

Danielle C. Turner[1], Michael R.F. Aitken[2], David R. Shanks[3], Barbara J. Sahakian[1], Trevor W. Robbins[2], Christian Schwarzbauer[4] and Paul C. Fletcher[1]

[1]Department of Psychiatry, University of Cambridge, School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK, [2]Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK, [3]Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK and [4]MRC Cognition and Brain Sciences Unit, Cambridge CB2 2EF, UK

Prediction error — a mismatch between expected and actual outcome — is critical to associative accounts of inferential learning. However, it has proven difficult to explore the effects of prediction error using functional magnetic resonance imaging (fMRI) while excluding the confounding effects of stimulus novelty and incorrect responses. In this event-related fMRI study we used a three-stage experiment generating preventative- and super-learning conditions. In both cases, it was possible to generate prediction error within a causal associative learning experiment while subtracting the effects of novelty and error. We show that right lateral prefrontal cortex (PFC) activation is sensitive to the magnitude of prediction error. Furthermore, super-learning activation in this region of PFC correlates, across subjects, with the amount learned. We thus provide direct evidence for a brain correlate of the surprise-dependent mechanisms proposed by associative accounts of causal learning. We show that activity in right lateral PFC is sensitive to the magnitude, though not the direction, of the prediction error. Furthermore, its activity is not directly explicable in terms of novelty or response errors and appears directly related to the learning that arises out of prediction error.

Keywords: associative learning, fMRI, PFC, prediction error

Introduction

Humans are quick to learn causal associations between co-occurring environmental stimuli. Traditional theories of human causal inference are based on statistical comparisons of co-occurrence rates across learning experiences (Cheng, 1997). Alternative accounts draw upon associative learning theories, postulating that causal inference is based upon the formation of associations between representations of events and their outcomes (Dickinson, 2001). In such theories, it is not merely co-occurrence, but also unpredictability, that governs the formation of these associations (Rescorla and Wagner, 1972; Schultz and Dickinson, 2000). As an example of this distinction, consider a person who suffers an allergic reaction every time they eat chicken. Even if several meals consisting of chicken and potatoes are eaten, and result in an allergic reaction, a causal link between the potatoes and the allergy is unlikely to form because the allergy is already fully predicted by the presence of chicken in the meal.

In previous functional neuroimaging studies of associative learning, unpredictability and novelty (of experimental trial structure) have been correlated (Ploghaus *et al.*, 2000; Fletcher *et al.*, 2001; McClure *et al.*, 2003; O'Doherty *et al.*, 2003). An important challenge, if we are to provide neurobiological support for associative theories, lies in their experimental dissociation. We have achieved this using preventative and super-learning tasks. In this setting we have been able to characterize the relationship between brain activity and the magnitude and direction of prediction error and to relate this to learning-dependent behavioural change.

Previously, we showed that activity in human dorsolateral prefrontal cortex (DLPFC) correlates with the surprise-dependent learning of a cue–outcome relationship (Fletcher *et al.*, 2001). Right DLPFC activation was greater when outcomes were unexpectedly present or absent. However, the unpredicted events were necessarily configured to be different from control trials in terms of event configuration and relative novelty. To delineate more accurately the functional neuroanatomy of prediction-error based learning it is necessary to match the activation and control events precisely for their configuration and familiarity. The use of compound cues in a causal inference task enables manipulation of the magnitude of prediction error across conditions whilst holding these factors constant.

It has been shown that, if the repeated co-occurrence of two stimuli with an outcome strongly defies the prediction of one or other of the stimuli, this association will strengthen to an unusual extent (Aitken *et al.*, 2000). This increase in prediction error is the basis for *super-learning* (Aitken *et al.*, 2000) and is analogous to the 'super-conditioning' of responding first described by Rescorla (1971). In order to generate super-learning, two prior learning stages must occur (Fig. 1). In the first, a subject learns that a given stimulus is positively associated with an outcome (A+). In the next stage – preventative learning – the familiar stimulus and a novel stimulus are seen together, with no outcome (AB–). This generates a negative prediction error (an unfulfilled expectation of an outcome). Stimulus B is attributed negative causal potential, i.e. preventative learning, because it prevents the allergy expected from stimulus A. In the third stage, stimulus B is presented together with a new stimulus and an outcome occurs (BC+). At this stage, the presence of stimulus B generates an expectancy that no outcome will occur and, therefore, the occurrence of the outcome generates an extra large positive prediction error. The strong and rapid learning that is generated by this error is known as super-learning. It occurs because stimulus C overcomes the preventative effect of the stimulus B, and is thus attributed greater causal significance than an appropriate control cue (Aitken *et al.*, 2000; Dickinson, 2001). Thus, super-learning may be conceived of as a special case of error dependent learning in which greater learning is engendered by a greater prediction error.

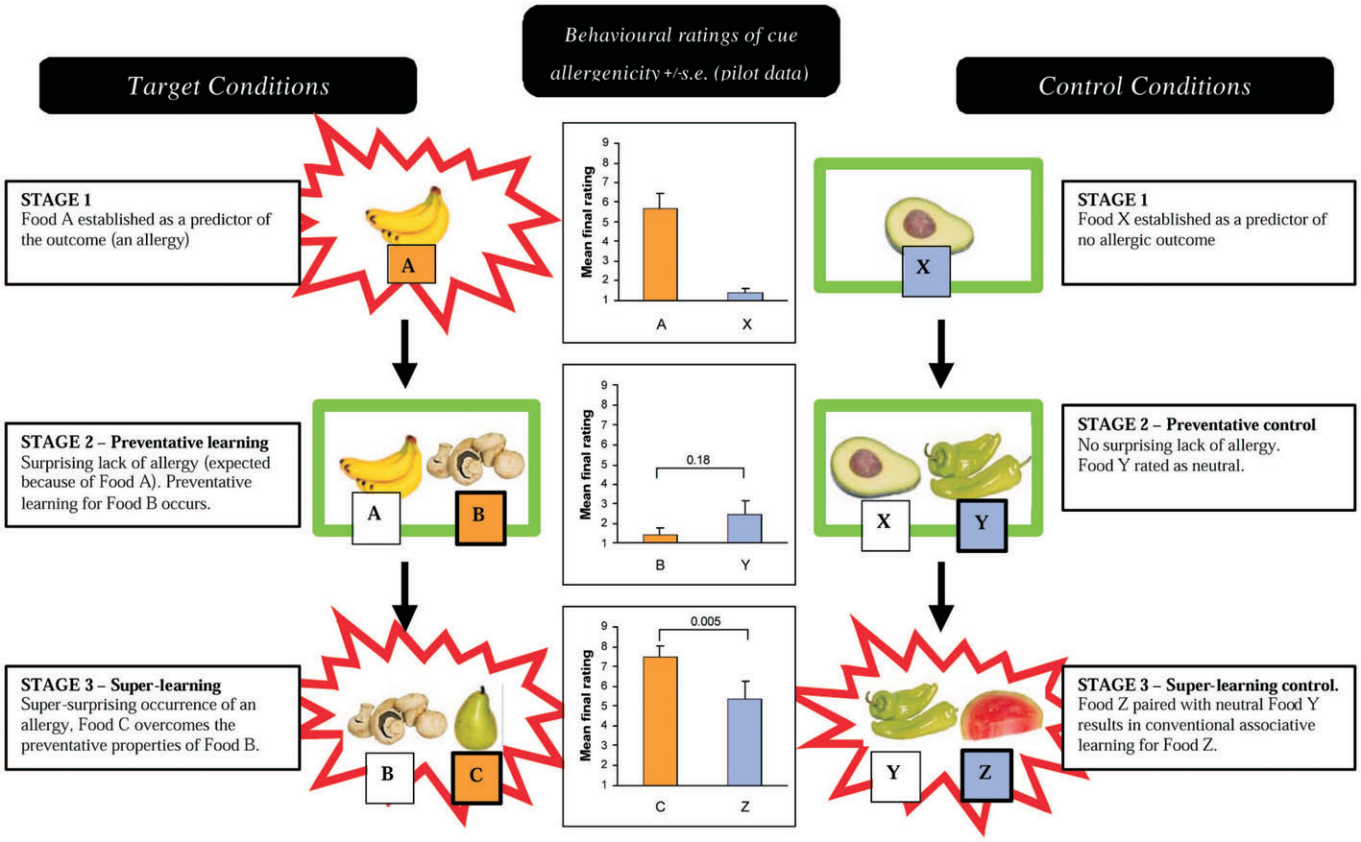

Figure 1. Illustration of the design of the experiment, showing the procedure used to generate super-learning and its appropriate controls. The pilot behavioural ratings obtained immediately on completion of the study (outside of the scanner) are illustrated in the centre panel.

We report an event-related fMRI study of super-learning using the three experimental stages required to produce this phenomenon (Fig. 1). As Figure 1 shows, this experimental procedure allows us to generate control trials that are matched for cue–outcome configuration and for stimulus novelty.

Materials and Methods

Subjects

Thirteen healthy, right-handed volunteers (eight female, five male) with a mean age of 27 ± 3 years and a mean predicted verbal IQ of 121 ± 3 (as indexed by the National Adult Reading Test from Nelson, 1982) were recruited from within the local community by advertisement. Exclusion criteria included a history of psychiatric or physical illness (particularly cardiovascular or neurological disorders), head injury, any history of drug or alcohol dependence, left-handedness or the possibility of magnetic metal being present in their body. All experiments were performed in compliance with the relevant laws and institutional guidelines. The study was approved by the Local Research Ethics Committee and written informed consent was given by all subjects prior to imaging. One female subject failed to perform the task and her data were therefore excluded from further analysis.

Learning Task Stimuli and Trial Structure

This task, and the instructions that preceded it, were based on an existing associative learning paradigm (Aitken *et al.*, 2000, 2001). In brief, before entering the scanner subjects were told to imagine themselves as food allergists whose goal it was to ascertain which of an array of presented foods would cause allergic reactions in an imaginary patient. They were presented with a series of trials in which an initial stimulus (a picture of a single food or a pair of foods – see Fig. 1 for example) informed them which food their imaginary patient had eaten. They were then required to predict whether an allergic reaction occurred using a two-choice button push, following which they were shown whether an allergic reaction had indeed occurred. If it had, a red jagged line appeared encircling the word 'Allergic Response', if it had not, a smooth green box appeared around the words 'No Response'. Stimuli for learning were presented on a screen using DMDX (K.I. Forster and J.C. Forster, University of Arizona), viewed via a mirror comfortably situated within the subject's field of view. Each trial lasted a total of 4 s, with the food presented for 3 s (during which time subjects made their predictive response) and the outcome ('Allergic Response' or 'No Response') for 1 s. Trials ran successively with occasional (1 per 20 trials) baseline events in which subjects viewed a fixation cross for periods of between 10 and 20 s.

The trial structure (stimulus–prediction–outcome) was comparable to that used in our previous associative learning study (Fletcher *et al.*, 2001). However, we used foods rather than fictitious drugs and syndromes to facilitate subjects' learning since multiple stimuli were required within a learning-session.

Learning Stages

The study employed a within-subjects design in which each subject was trained concurrently on a number of different contingencies between the food and allergic reaction. Learning occurred over three stages (Fig. 1).

Stage 1

This was the first of the two set-up phases. Subjects were presented with a total of six single foods across 60 trials (10 presentations of each food and its outcome in a randomized order). Two of the foods were invariably paired with an allergic response; two were invariably paired with no response. In addition, two foods were presented with a variable outcome (allergic response in 50% of cases, no response in the other 50%) to encourage subjects in the belief that causal contin-

gencies might vary for a given stimulus. Although this was primarily a set-up phase, functional imaging data were acquired and used to define a 'mask' of the learning system that was used to constrain the spatial analyses of subsequent effects and thus reduce the number of voxel-wise comparisons.

The three types of foods (allergic, non-allergic and variable) were used to set the scene for subsequent preventative and super-learning (stages 2 and 3, respectively). In order to ensure continuity across the stages (i.e. to prevent subjects from seeing successive stages as three separate studies and ignoring what had previously been learned), these foods were presented (in pairs) as 'fillers' during subsequent stages with their predictive relationships preserved. Thus, if subjects learned that 'bananas' and 'cake' both separately caused allergies during stage 1, then they would also see compound stimuli ('banana' plus 'cake' predicting an allergic response) in stages 2 and 3. This inclusion of filler cues to preserve experimental continuity is used in the behavioural studies upon which the current study is based. The filler cues in stages 2 and 3 were not included in the fMRI comparisons.

Stage 2 — Preventative Learning

During this stage, compound cues (pairs of foods) were presented. Once again, this stage may be considered a set-up phase for the evocation of super-learning in stage 3 although, in addition, it gave rise to preventative learning trials (Fig. 1). Foods in which a positive causal relationship with the allergic response had been established during stage 1 were now paired with novel foods and a 'No Response' outcome. Since these foods were associated with a strong expectancy of an allergic response, this non-outcome would be surprising. The result of the mismatch is preventative learning for the novel food, i.e. this item is considered to overcome the learned allergenicity of the familiar item. The comparison cues comprised a novel food and a familiar food that had been learned, during stage 1, to produce no response. Brain regions responsive to preventative learning events were isolated by a direct comparison of these two trial types. A total of 12 trials for each association were produced by stage 2.

As with the fillers from stage 1, which continued into stages 2 and 3, we also included further preventative learned cues from stage 2 during the subsequent (super-learning) stage. Once again, the purpose of this was to preserve continuity so that subjects did not view the stages as separate studies. Our intention had not been to use these trials in the fMRI analysis but we subsequently did so in order to ascertain that the effects associated with the super-learning trials were not attributable to the fact that these compound cues contained items that had been preventatively learned (see below).

Stage 3 — Super-learning

Super-learning was generated by pairing novel foods with familiar foods that had been presented in stage 2 [and thereby subjected to preventative learning (Fig. 1)]. Seeing the latter preventative food, a subject was likely to strongly predict a 'No Response' outcome. The expectancy violation, when an outcome occurred, would therefore be large and super-learning for this novel food would be generated. The control pairs, similarly, comprised the familiar item from stage 2 plus a novel item, followed by an allergic response. In this case, the response was also unexpected (the familiar item had previously been paired with no allergic response) and would therefore generate associative learning. However, the expectancy violation was not as great as in the super-learning condition [in the formalization of the Rescorla–Wagner theory (Rescorla and Wagner, 1972), the violation is 2λ in the super-learning condition versus λ in the control condition, where λ measures the maximum strength of an associative link]. Thus, the contrast between these two trials enables us to determine brain systems whose activity is greater when the expectancy–outcome mismatch is greater in the setting of trials that are well-balanced for familiarity/novelty and cue–outcome configuration.

Behavioural Measures

Prior to scanning, each participant was asked to rate the likelihood that each of the stimulus foods would produce an allergic reaction, in order to ascertain that they had no strong preconceptions about the foods that they would later be required to learn about. The row of numerical keys on the computer keyboard corresponded to an attached scale showing the likelihood of an allergic reaction occurring, ranging from 1 (definitely not) to 9 (definitely). Analysis of the pre-ratings by food type revealed no systematic effect of item on initial causal ratings, with a mean initial allergy rating for the target cues of 0.26 ± 0.38.

As well as the subjective ratings of allergenicity for each food, we recorded on-line predictive responses as a measure of the extent to which subjects changed and established their expectancy of a given food pairing causing an allergy across each of the learning stages.

Scanning

Imaging data were collected using a Bruker MedSpec 30/100 (Ettlingen, Germany) scanner operating at 3 Tesla. A T_R of 1.1 s allowed an acquisition of 1554 volumes (21 slices each of 4 mm thickness, interslice gap 1 mm) per subject. Gradient-echo echo planar T_2^*-weighted images depicting BOLD contrast were acquired from 21 noncontiguous near axial planes: T_E = 27.5 ms, flip angle = 66°, in-plane resolution = 3.1 × 3.1 mm, matrix size 64 × 64, field of view 20 × 20 cm, bandwidth 100 kHz.

Analysis of fMRI Data

All data analysis was carried out using statistical parametric mapping (Friston *et al.*, 1995) in the SPM 99 programme (Wellcome Department of Cognitive Neurology, London, UK). This included reorientation, slice acquisition time correction, within-subject image realignment, spatial normalization to a standard template (Cocosco *et al.*, 1997) and spatial smoothing using a Gaussian kernel (8 mm). The time series in each session was high-pass filtered (to a maximum of 1/120 Hz).

The average haemodynamic responses to each event type (designated as occurring at the presentation of the outcome stimulus) were modelled using a canonical, synthetic haemodynamic response function (Friston *et al.*, 1998). This function was used as a covariate in a general linear model and a parameter estimate was generated for each voxel for each event type. The parameter estimate, derived from the mean least squares fit of the model to the data, reflects the strength of covariance between the data and the canonical response function for a given condition. Individuals' contrast images, derived from the pairwise contrasts between parameter estimates for different events, were taken to a second level group analysis in which *t*-values were calculated for each voxel treating inter-subject variability as a random effect. The *t*-values were transformed to unit normal *Z* distribution to create a statistical parametric map for each of the planned contrasts.

Masking was used. We used stage 1 to identify a learning system [by comparing all trials to the baseline fixation task, False Discovery Rate (FDR) thresholded at $P < 0.05$ (Genovese *et al.*, 2002)]. This was primarily to ensure that all regions reported in subsequent contrasts of interest were those that showed activation relative to a low-level baseline task. Subsequent to this masking procedure, statistical thresholding for the contrasts of interest used a small volume correction based upon the an area (20 × 30 × 30 mm) encompassing the right prefrontal cortex (PFC) activation identified by our previous study (Fletcher *et al.*, 2001). All of the right frontal activations reported and discussed below survived a small volume FDR correction (Genovese *et al.*, 2002). This was motivated by a desire to maximize sensitivity (in the face of the limited power generated by subtle manipulations and necessarily few repetitions of each event) without inflating type II error. Of course, the use of such an approach is highly exclusive and it remains possible that regions outside the masks show task-dependent activity that will be of interest to subsequent researchers. For completeness, therefore, we report all regions for the important contrasts (preventative learning versus its control and super-learning vs its control) with a low threshold $P < 0.01$, uncorrected for multiple comparisons (see supplementary tables).

Main Effect of Associative Learning to Single Foods (Stage 1)

This effect was explored through a comparison of all associative learning trials to the randomly occurring fixation events. Its purpose was to define a set of brain regions sensitive to the associative learning task in order that subsequent analyses of preventative learning and super-learning could be confined to this system as described above.

Main Effects of Compound Cue Associative Learning, Preventative Learning and Super-learning Compared to Fixation Baseline
These analyses were carried out in order to establish the broad brain system activated in association with compound cue, surprise-dependent associative learning. The threshold for this analysis was, therefore, set at $P < 0.05$, FDR correction (Genovese *et al.*, 2002).

Effects of Preventative Learning
A direct comparison of preventative learning events (compared to the fixation baseline) with the appropriate control events (again in comparison with the fixation baseline), as illustrated in Figure 1, was carried out within the masked area defined by analysis of stage 1. FDR threshold $P < 0.05$ was set for this contrast. For regions of right PFC, a threshold of $P < 0.05$ (FDR corrected) was set using a small volume correction based upon the ROI as defined above.

Effects of Super-learning
A direct comparison of super-learning events with the appropriate control events (Fig. 1) was carried out, initially thresholded at $P < 0.05$ (FDR corrected). For regions of right PFC, a threshold of $P < 0.05$ (FDR corrected) was set using a small volume correction based upon the ROI as defined above.

Exploration of the Correlation Between Behavioural Change and Magnitude of Super-learning Related Brain Activation
In order to evaluate the extent to which an increase in the magnitude of activation (super-learning versus control associative learning task) predicted a greater change in predictive response (from negative to positive predictions) we calculated, for each subject, an index of this change. Average tendency to predict an allergic response in the first third of stage 3 was subtracted from that in the last third of stage 3 for each subject and the resulting values were regressed upon magnitudes of voxel activation within the learning system mask. Thresholding was as above.

Results

Behavioural Results

Consistent with the analogous behavioural studies (Aitken *et al.*, 2000; Le Pelley and McLaren, 2003), the predictive responses made by subjects in response to successive trials showed adaptation to the prevailing contingencies. All subjects made a greater number of 'yes' responses on trials with the outcome (filled symbols) than on trials without the outcome (open symbols) at the end of each stage of training (Fig. 2).

Scanning demands meant that individual food ratings, similar to those taken prior to scanning, could not be recorded immediately at the end of the task. However, equivalent data from a group of 10 volunteers, who piloted the task and completed rating scales immediately on finishing the task, showed the expected effects. Subjects rated the super-learned foods as more allergenic than the control foods [$t(9) = 3.63$, $P = 0.005$]. The ratings provided by the pilot data have been included in the central panel of Figure 1 for illustration purposes.

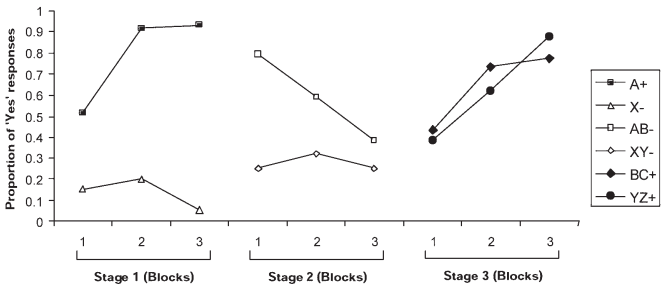

Figure 2. The change in predictive responses for each contingency across the stages. Subjects predictive responses across successive trials showed adaptation to the prevailing contingencies. The stages and symbols used to denote the different contingencies are the same as those used in Figure 1. Each stage was divided into three blocks illustrating the change in predictive responding over time. By the end of each learning stage, all subjects were making a greater number of 'yes' responses on trials in which an allergy was caused (filled symbols, denoted with a '+' in the key) than on trials where no outcome occurred (open symbols, denoted '–' in the key).

Scanning Results

A comparison of all single cue–outcome learning events with the fixation baseline events in stage 1 showed activation of a broad system comprising bilateral dorsolateral and ventrolateral PFC, anterior cingulate cortex, bilateral occipital and parietal cortex, cerebellum and medial temporal cortex including the hippocampus (Fig. 3). All of the comparisons reported below were masked by this analysis. Supplementary tables available online describe in detail the regions activated by each main learning event (preventative learning, compound cue associative learning and super-learning) compared to fixation baseline (supplementary Tables 1–3).

A direct comparison of the preventative learning events with the appropriate control events (as described in the methods) was carried out within the masked area defined by the contrast from stage 1. Discrete areas of superior, middle and inferior frontal gyri were activated (Fig. 4*a*). Similarly, a direct comparison of super-learning events with the appropriate control events (Fig. 4*b*) was carried out within the masked area. Table 1 (parts a and b) provides the coordinates for the areas of activation observed.

We next explored the extent to which super-learning (versus its control) correlated with the behavioural changes observed across subjects, using change in averaged predictive responses from the first to the last third of the learning phase (stage 3). Regions identified by this comparison included right lateral

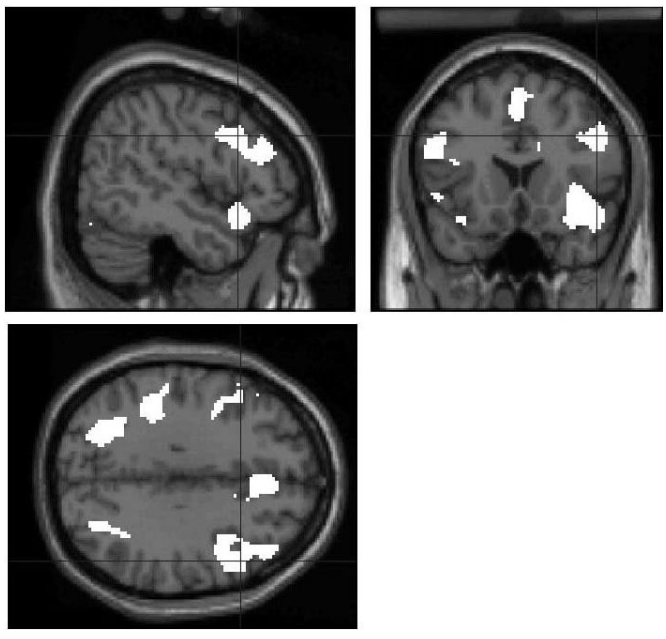

Figure 3. A comparison of all single cue–outcome learning events with the fixation baseline events in stage 1. The main effect of associative learning to single foods from stage 1 using SPM ($P < 0.001$) rendered onto structural MRI in standard space with sections at $x, y, z = 50, 18, 32$. This contrast was used as a mask for all further comparisons.

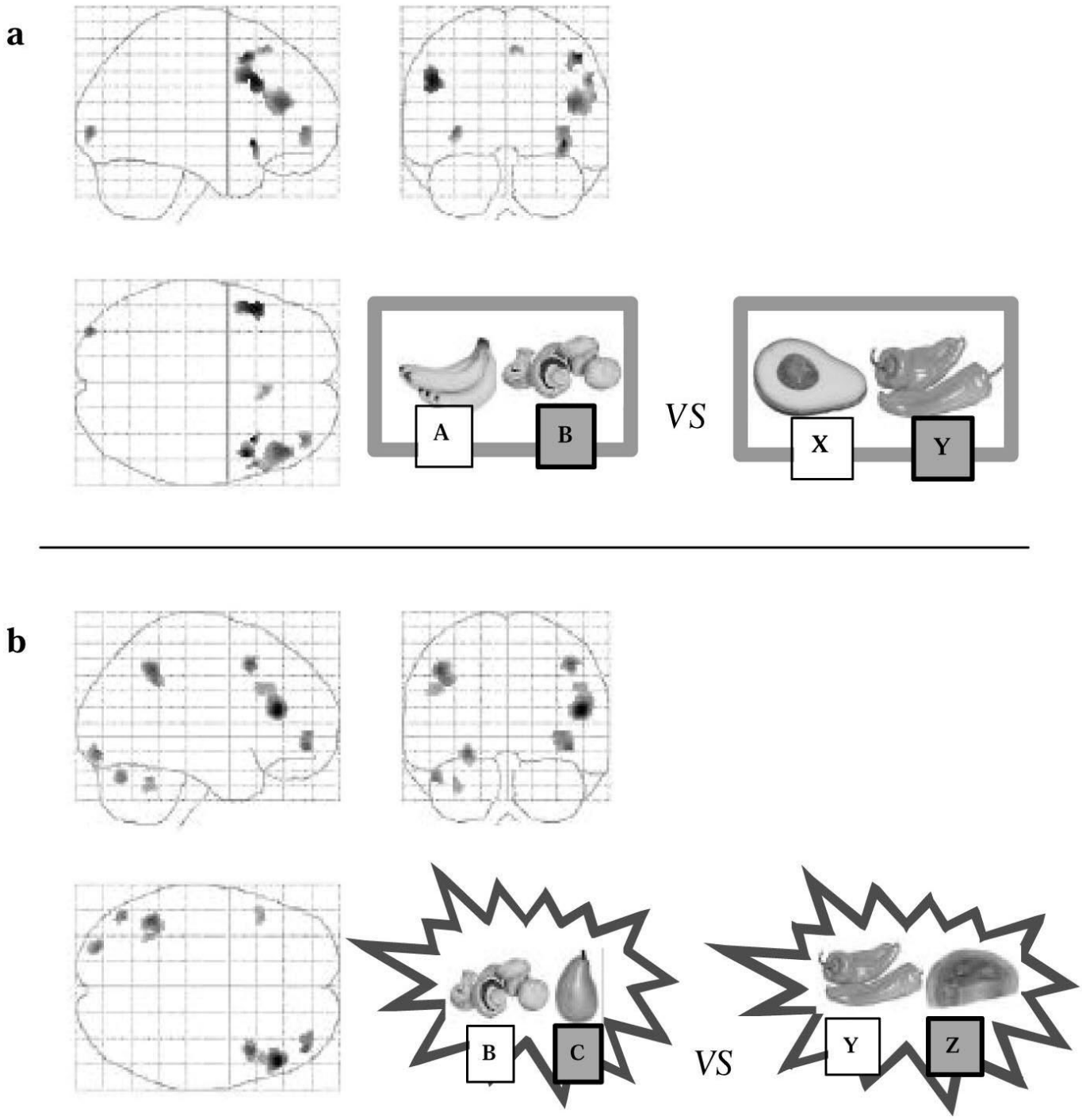

Figure 4. Comparison with control events. (*a*) The activations obtained from direct comparison of preventative learning trials with the appropriate control event. (*b*) The activations from super-learning versus its control.

PFC (Fig. 5) and are described in Table 1 (part c) ($P = 0.017$). Additionally, we explored the extent to which the size of behavioural change in the preventative learning condition correlated with magnitude of activation for this condition. A correlation was noted ($x, y, z = 42, 18, 46$, $P < 0.05$). That is, right lateral PFC showed greater activity in those subjects demonstrating a greater change as a result of preventative learning.

The degree of overlap between the preventative and super-learning conditions is noteworthy and is illustrated in Figure 6. This may suggest that the right lateral PFC activation in response to surprise dependent learning is independent of the direction of the prediction error and of whether subjects are learning a positive (causative) or negative (preventative) contingency between cues and outcome (though see discussion).

Table 1
Coordinates of activation foci together with *Z* scores and an estimate of where the activations lie in anatomical terms is presented for each contrast

		x	*y*	*z*	*Z* score
(a) Preventative learning Events AB vs XY (masked)					
Middle frontal gyrus	left	–48	18	34	2.69
Middle frontal gyrus	right	48	12	50	2.65
		44	36	18	2.37
		54	18	34	2.21
		58	24	24	2.02
Inferior frontal gyrus	right	38	16	–8	2.76
		38	50	2	2.20
Occipital cortex		–34	–90	0	2.33
Superior frontal gyrus		4	24	54	2.04
(b) Super-learning Events BC vs YZ (masked)					
Middle frontal gyrus	right	46	20	34	1.92
		42	14	48	2.37
		40	52	–8	2.36
Inferior frontal gyrus	right	50	32	18	2.83
		36	52	2	2.13
Inferior parietal lobe		–44	–50	44	2.45
Occipital cortex		–26	–86	–10	2.32
Cerebellum		–46	–70	–28	2.16
		–34	–52	–28	1.96
		–44	22	32	1.92
(c) Super-learning and behavioural change Correlation with events BC					
Middle frontal gyrus	left	–52	20	30	1.80
Middle frontal gyrus	right	44	36	26	2.39
		42	28	22	1.72
		50	22	28	2.25
Medial frontal cortex		0	34	34	2.22
Fusiform gyrus		18	–92	–10	2.16
Inferior frontal gyrus		36	18	–6	2.08

Coordinates highlighted in bold are those falling within the region of interest defined by our previous study (Fletcher *et al*., 2001) and surviving a small volume correction for multiple comparisons on the basis of these previous data.

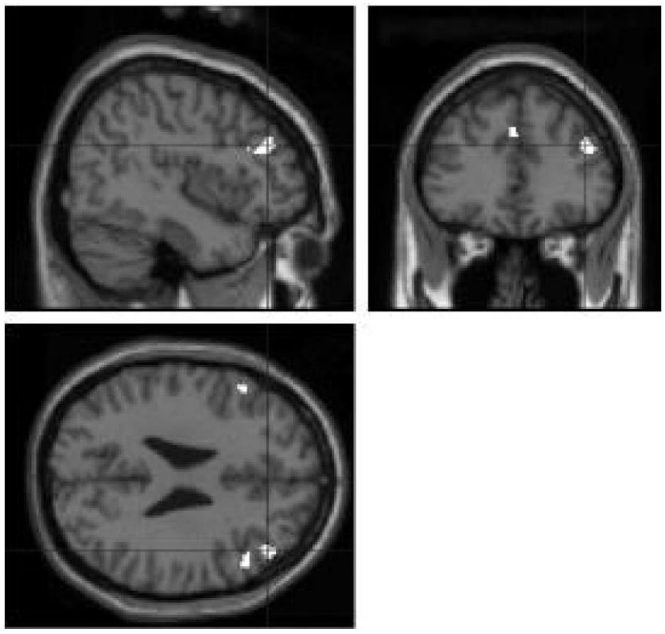

Figure 5. Areas sensitive to the correlation between behavioural change and the magnitude of super-learning related brain activation.

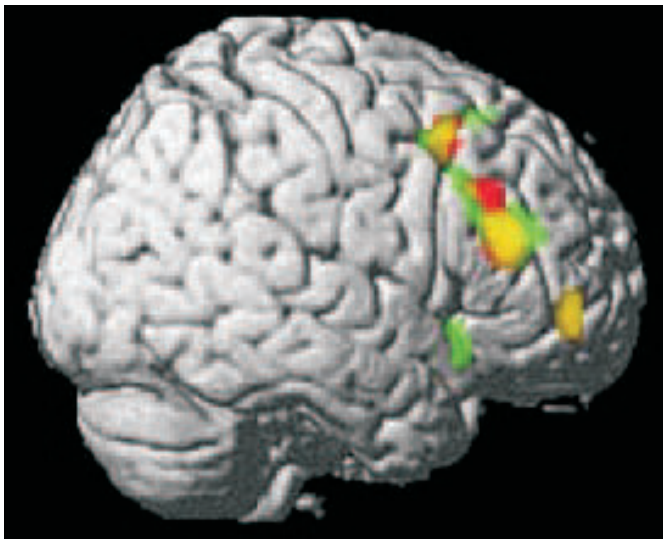

Figure 6. A graphical rendering of the activations from preventative learning versus its control (green) and super-learning versus its control (red). Yellow denotes overlap between the two conditions.

Note that, with respect to a more precise localization of the right frontal response to preventative and super-learning, the foci mainly fall in middle frontal gyrus and may therefore be designated as DLPFC. However, since most are in close proximity to the inferior frontal sulcus we shall refer to lateral PFC activations in the interests of caution.

Additional Comparison to Ensure That the Super-learning Activation Does Not Reflect the Presence of Preventatively Learned Items in Super-learning Trials

While the super-learning trials and their controls are matched for familiarity and outcome, one way in which they do differ is that super-learning compound cues contain items that have previously been subject to preventative learning. While the presence of such an item is critical for super-learning to take place, it could be argued that the presence of a preventatively learned item alone could produce right frontal activation. This could then account for the activations attributed here to super-learning. In order to ensure that this was not the case, we compared super-learning trials with filler trials in stage 3 containing preventatively learned items (see above). This contrast was carried out purely as a check and the results are not shown. They indicate that right frontal activation in super-learning trials was significantly greater than that seen for well-learned preventative trials. While this was not an initially planned contrast and the super-learning and preventative learning trial types are not matched for cue–outcome configuration and novelty we believe that the result indicates that the presence of preventatively learned items cannot be invoked to

account for the right frontal activation seen in the super-learning versus control contrast.

Discussion

Our results indicate that right lateral PFC is sensitive not just to prediction error on each learning trial, but also that its activity is augmented in those situations when prediction error is greater, either negatively (as in preventative learning) or positively (super-learning). This error-dependent activation predicts the degree to which learning occurs (as measured by the adjustment of the predictive responses in both the super-learning and the preventative learning conditions). Of course, our inference that right PFC is sensitive to prediction error is based upon the supposition that we have isolated this phenomenon from confounding factors that would normally correlate with prediction error. The use of compound cues has indeed allowed us to isolate brain regions whose activity reflects prediction-error dependent learning from those areas reflecting changes in stimulus novelty or task performance. We believe that this is a novel dissociation in a causal associative learning task. Furthermore, in other studies of reward–punishment-based associative learning (e.g. Ploghaus *et al.*, 2000; O'Doherty *et al.*, 2003), prediction-error dependent trials must also occur as relatively novel occurrences. While this is less of a problem when carrying out direct comparisons between positive and negative prediction error trials, it does make interpretation of common effects difficult – a problem that we have overcome here.

Compound cues enable a precise and pure manipulation of prediction error, both in its magnitude and its direction. Super-learning and preventative learning were matched with their respective control conditions in terms of event configuration and degree of item familiarity. In each case, cues comprised one familiar and one novel food. Both the super-learning trials and super-learning control trials were succeeded by an outcome (allergy). Similarly, for preventative learning and its control, no outcome (no allergy) occurred. In addition, for the super-learning condition, very similar changes in predictive behaviour occurred for both the target and control conditions (Fig. 2). Error rates did not differ across these different events and cannot, therefore, account for the activation differences.

An emerging functional neuroimaging literature suggests that frontal cortex is an important mediator of many aspects of human memory function (Fletcher and Henson, 2001). However, activations of the regions seen in the current study are by no means unique to explicit memory tasks. Studies exploring a variety of processes may include, as a component of the activation task, manipulations that are likely to produce ongoing inferential associative learning. Studies of functions as diverse as working-memory (Rypma and D'Esposito, 2003; van den Heuvel *et al.*, 2003), attentional control (Milham *et al.*, 2003), reversal learning (Cools *et al.*, 2002), set-shifting (Konishi *et al.*, 2002, 2003) and reward expectation (Ramnani and Miall, 2003) have all been shown to recruit prefrontal cortical areas overlapping with the ones reported here. Similarly, an fMRI study exploring the dynamic processing of sequences showed that unexpectedly violating sequential patterns also evoked similar patterns of activity in prefrontal and interconnected subcortical regions (Heuttel *et al.*, 2002). Thus, for example, in exploring inhibitory processes and target detection (Coull *et al.*, 1996; Menon *et al.*, 2001; Ramnani and Miall, 2003) the aim is frequently to explore the brain response to items that occur relatively rarely compared to background/baseline items. In such studies, an outcome–expectancy mismatch will occur, initially at least, as a result of this relative rarity and, in light of the current data, this must be considered as a plausible explanation for observed activations in such tasks.

Prediction error is increasingly becoming a focus for functional neuroimaging studies (Ploghaus *et al.*, 2000; Pagnoni *et al.*, 2002; Braver and Brown, 2003; McClure *et al.*, 2003; O'Doherty *et al.*, 2003), although this has largely focused upon emotionally salient learning, or conditioning. O'Doherty *et al.* (2003), for example, explored the evolving prediction error during the formation of an association between a conditioned stimulus (abstract visual stimulus) and an unconditioned stimulus (a juice reward). Considering within-trial prediction error patterns, they observed a positive and attenuating response to the unconditioned stimulus (US) in the ventral striatum and orbitofrontal cortex, an evolving positive response at the time of the conditioned stimulus (CS) and a deactivation at the point at which the reward would have been expected in subsequent surprise omission trials. This pattern was seen in more dorsal regions of the striatum by McClure *et al.* (2003) and is precisely that predicted by the temporal difference (TD) model (Schultz *et al.*, 1997). Elsewhere, a negative prediction error (unexpected omission of the US) is associated with increased blood oxygen level dependent (BOLD) responses: Pagnoni *et al.* (2002) showed that the nucleus accumbens responds to unexpected reward omission. With respect to aversive stimuli, the picture is less clear. Ploghaus *et al.* (2000) showed that direction of BOLD signal responses to the unexpected occurrence or omission of painful heat appeared highly variable across brain areas and subjects.

Our study focused upon the magnitude of prediction error (in both negative and positive directions). The comparison of activation events with a fixation baseline highlighted a system including frontal, parietal and medial temporal regions (Fig. 3); presumably an effect of the cognitive, as distinct from emotional, salience of our stimuli and design. Aside from our previous study (Fletcher *et al.*, 2001) some work has implicated the PFC directly in prediction error. O'Doherty *et al.* (2003) showed that a number of prefrontal regions were sensitive to the magnitude of reward prediction error, though not its direction. Ploghaus *et al.* (2000) also observed frontal responses to pain prediction error although, as mentioned, this varied markedly among subjects. It is important to note that our study did not systematically manipulate reward or punishment. This could account for the absence of striatal activation in our surprise events. We subsequently compared super-learning with its control condition at a much reduced threshold ($P < 0.05$, uncorrected). We noted bilateral caudate and nucleus accumbens activations in association with super-learning. This suggests that striatal regions may be sensitive to error-dependent learning even in the absence of reward or punishment, but, of course, we must be cautious in proposing this in view of the subtlety of the effects in these regions.

Our study provides unambiguous evidence for a specific response to prediction error, where activation and control events are balanced for configuration and familiarity. The critical question, therefore, lies in the precise function of the DLPFC in cognition. It is a region that has been activated in many studies of human memory – including working memory,

episodic memory encoding and retrieval (see Fletcher and Henson, 2001 for review). It is also frequently activated in attentionally demanding conditions (D'Esposito *et al.*, 1995; Coull *et al.*, 1996; Braver *et al.*, 1997) and in tasks requiring the production of non-automatic responses (Carter *et al.*, 1998; Botvinick *et al.*, 1999). While many of the observed patterns of prefrontal response are consistent with a role in novelty processing of unfamiliar stimuli (Ranganath and Rainer, 2003), novelty *per se* is an insufficient explanation for the DLPFC activation seen here, for the reasons described above. Rather, our findings are more consistent with models that propose PFC function to be central to learning processes – the ability to adapt behaviour in response to new information (Miller, 2000; Miller and Cohen, 2001). This new information initially presents itself as a mismatch between expectancy and outcome: a mismatch that forms the basis for change. Attentional modulation is posited as the initial response to prediction error by one influential model of associative learning (Pearce and Hall, 1980). This model suggests that the attentional modulation is not influenced by the direction of the prediction error, which is consistent with our demonstration of highly comparable patterns of right lateral PFC activity for both positive and negative surprise. We offer this interpretation with a degree of caution however since it is possible that there is a difference in localization between responses to preventative learning and super-learning trials but that this difference lies below the spatial resolution of the fMRI technique as used here. It is possible that, at an increased spatial resolution, differences in activation, within lateral PFC, reflecting the direction of prediction error might be observed.

We believe we have provided unambiguous evidence for the existence of a brain correlate of prediction error and have shown that this, in turn, predicts subjects' behavioural changes. This provides, to our knowledge, the first direct support for the mechanisms proposed by associative theories of causal learning in humans. These results confirm that the pattern of PFC response is consistent with the notion of prediction error (Friston, 2002), operationalized here as a discrepancy between the expected (on the basis of previous stimulus exposure) and the actual outcome. This error term provides an experimental framework within which to understand the existing data on prefrontal function. These data clearly show the effects of learning on prefrontal function in isolation from the other factors, such as increased novelty, that normally accompany it.

Notes

P.C.F. is funded by the Wellcome Trust. D.C.T. is funded by a MRC Research Studentship. The work was completed within the MRC Centre for Behavioural and Clinical Neuroscience. We are grateful to the staff at the Wolfson Brain Imaging Centre.

Address correspondence to Paul C. Fletcher, Department of Psychiatry, University of Cambridge, School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK. Email: pcf22@cam.ac.uk.

References

Aitken MR, Larkin MJ, Dickinson A (2000) Super-learning of causal judgements. Q J Exp Psychol B 53:59–81.

Aitken MR, Larkin MJ, Dickinson A (2001) Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. Q J Exp Psychol B 54:27–51.

Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402:179–181.

Braver TS, Brown JW (2003) Principles of pleasure prediction. Specifying the neural dynamics of human reward learning. Neuron 38:150–152.

Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC (1997) A parametric study of prefrontal cortex involvement in human working memory. Neuroimage 5:49–62.

Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280:747–749.

Cheng PW (1997) From covariation to causation: a causal power theory. Psychol Rev 104:367–405.

Cools R, Clark L, Owen AM, Robbins TW (2002) Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. J Neurosci 22:4563–4567.

Cocosco CA, Kollokian V, Kwan RKS, Evans AC (1997) Brainweb: online interface to a 3D MRI simulated brain database. Neuroimage 5:425.

Coull JT, Frith CD, Frackowiak RS, Grasby PM (1996) A fronto-parietal network for rapid visual information processing: a PET study of sustained attention and working memory. Neuropsychologia 34:1085–1095.

D'Esposito M, Detre JA, Alsop DC, Shin RK, Atlas S, Grossman M (1995) The neural basis of the central executive system of working memory. Nature 378:279–281.

Dickinson A (2001) The 28th Bartlett Memorial Lecture. Causal learning: an associative analysis. Q J Exp Psychol B 54:3–25.

Fletcher PC, Henson RN (2001) Frontal lobes and human memory: insights from functional neuroimaging. Brain 124:849–881.

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 (2002) Functional integration and inference in the brain. Prog Neurobiol 68:113–143.

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RS (1995) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-related fMRI: characterizing differential responses. Neuroimage 7:30–40.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Heuttel SA, Mack PB, McCarthy G (2002) Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nat Neurosci 5:485–490.

Konishi S, Hayashi T, Uchida I, Kikyo H, Takahashi E, Miyashita Y (2002) Hemispheric asymmetry in human lateral prefrontal cortex during cognitive set shifting. Proc Natl Acad Sci USA 99:7803–7808.

Konishi S, Jimura K, Asari T, Miyashita Y (2003) Transient activation of superior prefrontal cortex during inhibition of cognitive set. J Neurosci 23:7776–7782.

Le Pelley ME, McLaren IP (2003) Learned associability and associative change in human causal learning. Q J Exp Psychol B 56:68–79.

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

Menon V, Adleman NE, White CD, Glover GH, Reiss AL (2001) Error-related brain activation during a Go/NoGo response inhibition task. Hum Brain Mapp 12:131–143.

Milham MP, Banich MT, Claus ED, Cohen NJ (2003) Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. Neuroimage 18:483–493.

Miller EK (2000) The prefrontal cortex and cognitive control. Nat Rev Neurosci 1:59–65.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Nelson H (1982) National adult reading test manual. Windsor: NFER-Nelson.
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.
Pagnoni G, Zink CF, Montague PR, Berns GS (2002) Activity in human ventral striatum locked to errors of reward prediction. Nat Neurosci 5:97–98.
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.
Ploghaus A, Tracey I, Clare S, Gati JS, Rawlins JN, Matthews PM (2000) Learning about pain: the neural substrate of the prediction error for aversive events. Proc Natl Acad Sci USA 97:9281–9286.
Ramnani N, Miall RC (2003) Instructed delay activity in the human prefrontal cortex is modulated by monetary reward expectation. Cereb Cortex 13:318–327.
Ranganath C, Rainer G (2003) Neural mechanisms for detecting and remembering novel events. Nat Rev Neurosci 4:193–202.
Rescorla RA (1971) Variations in the effectiveness of reinforcement following prior inhibitory conditioning. Learn Motiv 2:113–123.
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: current theory and research (Black AH, Prokasy WF, eds), pp. 64–99. New York: Appleton-Century-Crofts.
Rypma B, D'Esposito M (2003) A subsequent-memory effect in dorsolateral prefrontal cortex. Brain Res Cogn Brain Res 16:162–166.
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.
van den Heuvel OA, Groenewegen HJ, Barkhof F, Lazeron RH, van Dyck R, Veltman DJ (2003) Frontostriatal system in planning complexity: a parametric functional magnetic resonance version of Tower of London task. Neuroimage 18:367–374.