Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions

P.R. Corlett,¹ G.K. Murray,¹,² G.D. Honey,¹ M.R.F. Aitken,³ D.R. Shanks,⁴ T.W. Robbins,³ E.T. Bullmore,¹,² A. Dickinson³ and P.C. Fletcher¹

¹Brain Mapping Unit, Department of Psychiatry, University of Cambridge, School of Clinical Medicine, Addenbrooke’s Hospital, Cambridge CB2 2QQ, ²CAMEO, Cambridgeshire and Peterborough Mental Health Partnership NHS Trust, Cambridge, CB1 5EE, ³Department of Experimental Psychology, University of Cambridge, Downing Street, CB2 3EB, UK and ⁴Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK

Correspondence to: Dr Paul C. Fletcher, Box 255, University of Cambridge, Department of Psychiatry, Addenbrooke’s Hospital, Cambridge CB2 2QQ, UK
E-mail: pcf22@cam.ac.uk

Delusions are maladaptive beliefs about the world. Based upon experimental evidence that prediction error—a mismatch between expectancy and outcome—drives belief formation, this study examined the possibility that delusions form because of disrupted prediction-error processing. We used fMRI to determine prediction-error-related brain responses in 12 healthy subjects and 12 individuals (7 males) with delusional beliefs. Frontal cortex responses in the patient group were suggestive of disrupted prediction-error processing. Furthermore, across subjects, the extent of disruption was significantly related to an individual’s propensity to delusion formation. Our results support a neurobiological theory of delusion formation that implicates aberrant prediction-error signalling, disrupted attentional allocation and associative learning in the formation of delusional beliefs.

Keywords: prediction error; associative learning; fMRI; delusions; psychosis

Abbreviations: rPFC = right prefrontal cortex

Received February 7, 2007. Accepted July 5, 2007

Introduction

Delusions are fixed and irrational beliefs. Establishing their neurobiological basis is a major challenge, given their complex and insidious nature (Gilleen and David, 2005; Corlett et al., in press). Advances in our understanding of the neural bases of learning and belief formation may help us to meet this challenge. Theories of psychiatric illness that implicate the formation of abnormal associations between ideas have a long history (Locke, 1690/1976; Hartley, 1801; Pavlov, 1928). Indeed, the earliest theories of schizophrenia suggest that the formation of such aberrant associations is a core disease process (Bleuler, 1911/1950; Schneider, 1930). These views have been refined and, latterly, embedded in the neural architecture of learning and association formation (Miller, 1976; Beninger, 1988; Gray et al., 1991; Vinogradov et al., 1992; Hemsley, 1994; Kapur, 2003; Corlett et al., in press).

Prediction error, the mismatch between expectancy and experience, plays a direct role in forming and strengthening associations (Rescorla and Wagner, 1972; Dickinson, 2001) and has an indirect impact upon learning through the allocation of attention to stimuli in the environment. Organisms attend to events that violate their expectancies, which in turn facilitates associative learning (Mackintosh, 1975; Pearce and Hall, 1980; Grossberg, 1982). Psychotic patients describe how, in the early stages of their illness, irrelevant details capture their attention and are imbued with inappropriate significance (Matussek, 1952; McGhie and Chapman, 1961; Chapman, 1966). The attempt to rationalize and account for these bizarre experiences may result in the formation of delusions (Maher, 1974; Gray et al., 1991; Hemsley, 1994; Kapur, 2003; Corlett et al., in press).

Data from laboratory animals indicates that prediction errors are signalled by midbrain dopamine neurons (Schultz, 2000; Waelti et al., 2001). These neurons send dense projections to the basal ganglia and prefrontal cortex, forming the mesocorticolimbic dopamine system (Mann, 1986). Given that this system is strongly implicated in the pathophysiology of schizophrenia (Robbins, 1990; Grace, 1993; Laruelle et al., 2003), and that delusions are theorised to result from abnormal formation of associations, it has been suggested that dysfunction of the mesocorticolimbic dopamine system causes delusion formation via disrupted
prediction-error signalling (Gray et al., 1991; Hemsley, 1994; Kapur, 2003; Corlett et al., in press).

We have previously identified a reliable marker for prediction-error processing in right prefrontal cortex (rPFC) using fMRI (Fletcher et al., 2001; Corlett et al., 2004; Turner et al., 2004) and have recently shown that ketamine, a drug which causes a transient psychosis, perturbs this brain response in healthy individuals. Furthermore, the magnitude of the prediction-error response under placebo predicts an individual’s likelihood of experiencing delusional beliefs under ketamine (Corlett et al., 2006). While these data are consistent with a link between disrupted prediction-error signalling and delusional beliefs, direct evidence is needed from individuals suffering a psychotic illness. In the present study, we therefore used the associative causal learning task that we used previously (Corlett et al., 2004, 2006) to evaluate prediction-error signal in such individuals, relating this signal, on an individual basis, to delusion severity.

**Experimental procedure**

**Subjects**

Twenty-eight subjects were recruited for the study. Fourteen healthy volunteers were identified from within the local community by advertisement. Fourteen psychotic patient volunteers were identified from the Cambridge-based CAMEO early intervention in psychosis service (http://www.cameo.nhs.uk). All patients had a diagnosis of first-episode psychosis according to the following criteria: clinical presentation with psychotic symptoms for the first time; presentation with psychotic symptoms with suspected untreated episodes in the past, or following previous treatment with antipsychotic medication for less than 6 months (Barnett et al., 2005).

Two healthy volunteers and two CAMEO patients were unable to comply with the scanning procedure and so each group was comprised of 12 subjects (Healthy control; 8 males and 4 females and CAMEO patients; 7 males and 5 females). All subjects gave written informed consent prior to their involvement in the study and received an honorarium for their participation. The study was approved by the Addenbrooke’s NHS Trust Local Research Ethics Committee and was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). The groups were matched for age, handedness and IQ as measured using the National Adult Reading Test (Nelson, 1982). The mean age was 26 years (standard deviation, SD = 3) for both groups, the mean IQ for the control group was 116 (SD = 5); for the patient group it was 111 (SD = 12). Both psychotic and healthy subjects had normal structural MR brain scans, as confirmed by neuroradiological assessment. Control subjects were without a history of psychiatric or physical illness (particularly cardiovascular or neurological disorders), head injury, any history of drug or alcohol dependence, based on subjects’ responses during a locally devised telephone screening interview. Both patient and control subjects were without contra-indications for fMRI scanning.

**Patient clinical information**

At the time of scanning, patients underwent a psychiatric assessment using the Brief Psychiatric Ratings Scale (BPRS; Ventura et al., 1993). This assessment revealed both positive symptoms (e.g. hallucinations, delusions and thought disorder, BPRS positive symptom score, 14 ± 3) and negative symptoms (e.g. anhedonia and affective flattening, BPRS negative symptom score, 6 ± 3). Eight of the 12 patients were stabilized on atypical antipsychotic medication; see Table 1 (chlorpromazine equivalent dose, 181 ± 70 mg/day (Woods, 2003).

**Associative learning task**

We used the task reported by Corlett et al. (2004), a retrospective revaluation paradigm in which engendered expectations are violated to produce a prediction error. The task design is summarized in Table 2. Subjects were asked to imagine themselves working as an allergist confronted with a new patient ‘Mr X’, who suffers allergic reactions following some meals but not others. Their task was to work out which foods caused allergic reactions by observing the consequences of eating various foods. The task consisted of a series of trials each of which had the general structure summarized in Fig. 1. Trials comprised presentation of a food picture (representing a meal eaten by Mr X), a predictive response by the subject and, following this, an outcome. Subjects’ behavioural responses indicated both the outcome that they predicted for a particular meal (which button that they pushed) and their confidence in that prediction (how long they held down the button).

**Experimental structure**

The key manipulations relevant to the question under study are summarized in Fig. 2. Each subject was trained concurrently on a number of different contingencies between foods and allergic reactions. Learning occurred over three stages: Stage 1, Training; Stage 2, Retrospective revaluation and, finally, Stage 3, Violation. This design is clarified with examples in Fig. 2 and Table 2. In summary expectancies were set using an initial training phase. Retrospective revaluation (unovershadowing and backward blocking) occurred, engendering revalued expectancies about particular foods and, at the critical stage, we explored the impact on brain activity of violation of those revalued expectancies, relative to events that confirmed subjects’ expectancies (Table 2), in which a consistent relationship between a stimulus and an outcome (or non-outcome) was maintained throughout the course of the experiment.
fMRI data acquisition

A Bruker MedSpec 30/100 (Ettlingen, Germany) operating at 3 Tesla was used to collect imaging data. Gradient-echo echo planar T2*-weighted images depicting BOLD contrast were acquired from 21 non-contiguous near axial planes: TR = 1.1 s, TE = 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. A total of 705 volumes per subject were acquired in stage 1 and 893 volumes per subject across stages 2 and 3 (21 slices each of 4 mm thickness, interstice gap 1 mm). The first 6 volumes were discarded to allow for T1 equilibration effects leaving 887 volumes.

fMRI data analysis

fMRI data were analysed using statistical parametric mapping in the SPM2 programme (Wellcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm). Images were realigned, spatially normalized to a standard template and spatially smoothed with a Gaussian kernel (8 mm). The time series in each session were high-pass filtered (to a maximum of 1/120 Hz) and serial autocorrelations were estimated using an AR(1) model. The average haemodynamic response to each
The occurrence of no allergic reaction following unovershadowing should fulfill the retrospective revaluation during stage 2. Items associated with unovershadowing should violate the expectation engendered by the revaluation process in stage 2. Thus the trials on which no outcome was presented following the critical stage of the experiment. It involved items that were absent at stage 2 but subject to unovershadowing (represented here by the banana). Half of these items were presented in association with an allergic reaction, half with no allergic reaction. Critically, the trials on which no outcome was presented following unovershadowing should violate the expectation engendered by retrospective revaluation during stage 2. Items associated with an allergic reaction following unovershadowing should fulfill the prediction engendered by the revaluation process. Thus the occurrence of no allergic reaction following banana should be surprising (i.e. should engender a prediction error).

Fig. 2 Experimental stages. Stage 1—Training: This preliminary stage set up the expectancies. The key trial types in this were pairs of foods (in this case, hamburger paired with banana) in which subjects learned to expect an allergic response. Stage 2—Unovershadowing: In the unovershadowing condition, one item from the pair (here, Hamburger) that had previously been paired with an allergy was presented without an allergy outcome. The aim was to engender an augmented expectancy that the other cue from the pair (Banana) would cause an allergy. This process of increasing the expectancy associated with the absent food is called unovershadowing. Stage 3—Violation of learned expectancies: This was the critical stage of the experiment. It involved items that were absent at stage 2 but subject to unovershadowing (represented here by the banana). Half of these items were presented in association with an allergic reaction, half with no allergic reaction. Critically, the trials on which no outcome was presented following unovershadowing should violate the expectation engendered by retrospective revaluation during stage 2. Items associated with an allergic reaction following unovershadowing should fulfill the prediction engendered by the revaluation process. Thus the occurrence of no allergic reaction following banana should be surprising (i.e. should engender a prediction error).
measure of subjects’ prediction-error response (Corlett et al., 2006), which is spatially consistent with prediction-error responses engendered in the context of a range of human causal learning phenomena (Fletcher et al., 2001; Corlett et al., 2004; Turner et al., 2004).

**Investigating the relationship between delusion scores and brain response to associative learning in psychotic patients**

Demonstrating that prediction-error-driven associative learning is disrupted in psychotic patients would provide some support for associative theories of symptom formation. However, more compelling evidence in favour of such theories would be the demonstration of a relationship between the extent to which prediction-error processing is disrupted in psychotic patients and the severity of their symptomatology. The relationship between positive and negative symptoms at the time scanning, and brain responses to simple associative learning, unovershadowing and prediction-error processing were assessed using backward multiple regression (Altman, 1991). This process begins by fitting the full model of all potentially explanatory symptom variables and proceeds by removing unimportant variables, one at a time, until all of those remaining in the model contribute significantly. The symptom variables entered into the model were BPRS unusual thought content; grandiosity; suspiciousness; hallucinations; conceptual disorganization and negative symptom scores (Ventura et al., 1993).

**Effects of medication**

Of the 12 psychotic patients included in the final analysis, 8 were medicated on stable doses of atypical antipsychotics, 4 were unmedicated. The effect of antipsychotic medication upon brain activation in the context of this task was determined by computing the correlation, across subjects, between brain activation in the regions of interest at stages 1, 2 and 3 and chlorpromazine equivalent dose of medication (Woods, 2003).

**A dynamic assessment of brain changes in response to changing prediction error**

Prediction errors change with training and, in a secondary analysis of the data, we explored further the nature of aberrant prediction-error processing in individuals with psychosis by exploring the relationship between trial-by-trial behavioural predictions and brain responses to outcome presentation. Taking the data from the learning stage, we calculated, for each individual and each trial type, trial-by-trial estimates of prediction error based upon individually specified behavioural responses. Brain responses were modelled as canonical haemodynamic responses occurring at the presentation of feedback in each trial parametrically modulated by the behavioural response and the confidence asserted by the individual with respect to this response. Put simply, this analysis was based upon the hypothesis that prediction error should diminish as, across trials, individuals made correct responses with increasing confidence and that brain regions responding to prediction error should show decreases with this increasing confidence. For patients and controls separately, we analysed the nature of relationship between activity [in regions within the previously specified mask (see earlier)] and this trial-by-trial regressor using one-sample t-tests. We then compared the relationships across groups using a two-sample t-test. Our hypothesis was that, complementary to the findings from the subtraction analysis above, there would be a significant relationship between prediction error and rPFC response in the controls and that this would be attenuated in patients. In a further analysis, confined to the patient group, we explored whether the relationship was modulated by the BPRS measure of unusual thought content. In this case, we sought to determine whether a perturbed prediction-error response in rPFC was correlated with delusional symptoms.

**Results**

**Behavioural results**

Figure 3 illustrates subjects’ behavioural responses across the three training phases. It is clear from the plots that both patients and controls rapidly learn to predict the outcome of each meal, with increasing confidence. Analyses of variance with group (patient and control) as a between-subjects factor and training trial as a repeated measure, were conducted upon these data using SPSS 13.0 for Windows. The stage 1 analysis revealed a significant effect of training trial \[(F(5,118) = 57.9, P = 0.0001)\] but no group by training interaction \[(F(5,118) = 0.642, P = 0.681)\].

The findings were similar at stage 2; a significant effect of training \[(F(4,85) = 28.3, P < 0.0001)\] but no significant group by training interaction \[(F(4,85) = 0.31, P = 0.873)\].

These findings suggest that both psychotic patients and controls did indeed acquire the associative relationships between foods and outcomes.

Unovershadowing was similar in both groups, confirmed by subjects’ initial predictive responses at stage 3. Subjects’ initial choice (allergy or no allergy) and their confidence in that choice, about revalued items at trial 1 of stage 3 relative to well-learned control items, provides a measure of what they learned about the absent food. These data are depicted in Fig. 3. Analysis of variance revealed that unovershadowing had indeed taken place, since both patients and control subjects predict an allergic response following unovershadowed items relative to well-learned items [main effect of task, \[F(1,22) = 84.37, P < 0.001\]]. There was no significant group by learning interaction \[(F(1,22) = 0.357, P = 0.556)\].

Overall, the behavioural observations show that both patients and controls were capable of acquiring basic
associations and of revaluating these associations as a result of additional information. This matched performance is important to interpreting any brain differences since, clearly, it would be difficult to interpret differences in prediction-error responses if subjects were making different predictions (Price and Friston, 1999).

**Neuroimaging results**

**Patient versus control analyses**

**Stage 1 (training).** Table 3 summarizes the group differences in brain response to stage 1 training. This stage represents the basic acquisition of associative relationships between causal food cues and allergy/no allergy outcomes. As Fig. 4 shows, psychotic patients failed to activate the left caudate relative to controls.

**Stage 2 (retrospective revaluation).** Between-group differences in the neurophysiological response to unovershadowing are detailed in Table 3 and Fig. 5. In psychotic patients, bilateral substantia nigra and rPFC were not significantly more active for unovershadowed items than well-learned control items.

**Stage 3 (violation of expectancies).** The characteristic response to prediction error in rPFC (Fletcher et al., 2001; Corlett et al., 2004; Turner et al., 2004; Corlett et al., 2006) was perturbed in patients. Figure 6 and Table 3 depict this effect. The significant group-by-expectancy-violation interaction seems to be being driven by two effects; first, an attenuation of the rPFC activation to the...
unexpected event and secondly, an augmentation of the rPFC response to predictable events.

Relating brain prediction-error responses to delusions in the patient group

Brain response to expectancy violation is disrupted in first-episode psychosis patients, consistent with the associative theory outlined earlier. This finding was further supported by an observation that, following backward linear regression analysis as described earlier, the current level of unusual thought content on BPRS was the only surviving symptom correlating with rPFC response to expectancy violation at stage 3. Specifically, the greater the level of current unusual thought content, the less likely that rPFC activation distinguished violation and fulfillment of expectancy (Fig. 7).

Effects of medication

There was a significant \( (P < 0.05, r = -0.6) \) correlation between activation in left ventral striatum activation during the acquisition of causal associations at stage 1 and chlorpromazine equivalent dose of antipsychotic medication. There were no other significant correlations between brain activity and medication level at stages 2 and 3. Thus, the group difference noted at stage 1 should be treated with caution as a potential effect of medication. It will not be discussed further, particularly given that the key experimental manipulations were at stages 2 and 3 (Table 2).

Post hoc analyses

Regions more active in patients than controls in response to expectancy violation

Since the behavioural performance of individuals with psychosis and healthy controls did not differ significantly (see earlier), it is possible that individuals with psychosis engaged compensatory mechanisms. To examine this possibility, we contrasted the brain responses of patients and controls to expectancy violation at stage 3, weighting the contrast to identify regions that were more active in patients than controls. At a whole-brain uncorrected threshold of \( P < 0.05 \) we found a region of right middle temporal cortex that was more active in patients than controls. However, activation in this region did not correlate with behavioural performance on the task, as measured by subjects’ behavioural confidence and prediction about the first presentation of the item that had been subject to unovershadowing (Pearson’s \( r = 0.173, P = 0.59 \)).

Relating behavioural performance with delusion severity

We examined the possibility of a relationship between subject’s learning performance and their delusion severity, regressing patients’ delusion scores (as measured by their BPRS unusual thought content score) on their behavioural

### Table 3

<table>
<thead>
<tr>
<th>Stage</th>
<th>Region</th>
<th>( t )-value</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L Striatum</td>
<td>3.15</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>L substantia nigra</td>
<td>3.02</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>R substantia nigra</td>
<td>3.20</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>R prefrontal cortex</td>
<td>2.94</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>R prefrontal cortex</td>
<td>2.74</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note: The table depicts the results of region of interest analyses. Only significant \( P < 0.05 \) Bonferroni corrected differences between patients and controls are reported.
**Fig. 5** Group differences in brain response to retrospective revaluation of causal associations. The maximum intensity projection depicts between-group differences in brain responses to unovershadowing. The rendered images show the position of specific regional differences and next to those renderings, the plots display average parameter estimates from those regions for the controls (C) and patients (P), for unovershadowed (Unover) and control items.

**Fig. 6** Group differences in brain response to prediction error. The maximum intensity projection depicts between-group differences in brain response to prediction error. The rendered image shows the specific position of regional differences. In the adjacent panel, the plot displays average parameter estimates from those regions for the controls (C) and patients (P) for violation and control items.
performance (as measured by their prediction and confidence about the unovershadowed item at the first trial of stage 3). There was no significant correlation between subjects behavioural learning and their delusion severity (Pearson’s $r = 0.125$, $P = 0.698$).

**Analysis of brain changes in response to trial-by-trial changes in prediction-error: controls**

As expected, with decreasing prediction error (estimated by increasing confidence in correct responses), activity in rPFC was significantly reduced ($x, y, z = 42, 16, 20, z = 3.6; x, y, z = 54, 26, 18, z = 3.2$). This provides strong confirmatory evidence for the relationship between activity in rPFC and prediction error.

**Patients**

There was no evidence of a significant relationship between rPFC and the dynamic measure of prediction error in the patient group as a whole.

**Controls versus patients**

There was some, albeit subtle, evidence of a significant group by prediction-error interaction, with the relationship between rPFC and prediction error being significantly stronger in controls than patients ($x, y, z = 42, 16, 20, z = 2.3$ and $44, 14, 16, z = 2.9$). These data are summarized in Fig. 8a.

**Relationship between trial-by-trial prediction-error responses and delusional severity**

The BPRS score on unusual thought content, an estimate of delusional severity, was significantly related to the rPFC-prediction-error pattern in the patient group. Those individuals showing lack of the predicted relationship were the ones showing delusional scores of greatest severity on the day of scanning ($x, y, z = 34, 34, 26$, Pearson’s $r = 0.81$, $P < 0.001$). See Fig. 8b and c.

**Discussion**

We determined brain responses to the acquisition, revaluation and violation of associations in psychotic patients. In addition to a general attenuation in prediction-error-related signal in rPFC, the patient group showed a significant relationship between this attenuation and their current experience of odd beliefs (as measured by their unusual thought content score on the BPRS). Given the consistent relationship between rPFC response and the violation of learned expectancies in causal learning (Fletcher et al., 2001; Corlett et al., 2004; Turner et al., 2004; Corlett et al., 2006); this perturbation, in psychotic patients, of rPFC response and its correlation with delusion score provides experimental support for learning-based accounts of delusion formation (Beninger, 1988; Gray et al., 1991; Vinogradov et al., 1992; Hemsley, 1994; Kapur, 2003; Corlett et al., in press). Our findings are consistent with the prior suggestion of fronto-basal-ganglia disruption in psychosis (Robbins, 1990), a suggestion which has already received support from neuropsychological (Elliott et al., 1995; Joyce et al., 1996; Pantelis et al., 1997; Hutton et al., 1998) and functional brain imaging studies (Bertolino et al., 1999; Biver et al., 1995; Buchsbaum et al., 1999; Manoach et al., 2000; Meyer-Lindenberg et al., 2002). We extend these studies by demonstrating a relationship between a specific psychological process (prediction-error processing), its neural instantiation (rPFC) and a psychotic symptom (delusions).

It is important to note that we have taken a symptom-rather than syndrome-approach to psychosis. Consequently,
our findings are applicable to delusions in general but are not diagnostically specific. Our experimental approach to understanding delusions rests upon a series of studies in healthy controls implicating a frontostriatal system in prediction-error-dependent causal associative learning (Fletcher et al., 2001; Corlett et al., 2004; Turner et al., 2004; Corlett et al., 2006).

The present data are consistent with the possibility that in psychotic patients, prediction errors are signalled inappropriately and those errors maladaptively update the prefrontal representation of the world with irrelevant information. In this respect, it is noteworthy that brain responses to violations of learning (at stage 3) correlated with delusions. Furthermore, our use of individual- and trial-specific regressors to estimate dynamic brain changes corroborated the link between prediction error signal and delusions: those patients in whom there was a relatively preserved link between rPFC activation and prediction error showed lowest delusional scores (Fig. 8). This is further evidence that an aberration in this frontally mediated

**Fig. 8** Dynamic changes in rPFC with trial-by-trial changes in prediction error. (a) The region of right lateral PFC in which there was a significant relationship between prediction-error changes (as estimated by changing behavioural response) and activation in controls (shown in red). Superimposed on this is the region (shown in yellow) in which there were group differences in this relationship. Specifically, there was an attenuation of this right PFC-prediction-error relationship in the patient group. (b) A surface rendering showing the region of right lateral PFC in which there was a significant inverse correlation between PFC prediction-error response (using the trial-by-trial estimate as for Fig. 8a) and symptoms scores (unusual thought content on BPRS). The plot in Fig. 8c shows this relationship graphically. Specifically, we have plotted across individuals, the BPRS scores against the extent to which lateral PFC activity correlated with out estimate of trial-by-trial prediction error. As can be seen, individuals in whom this relationship was most strongly positive, showed lower scores on this symptom (Pearson’s r = 0.81).
inferential processing relates to delusion formation. We should note that it is perfectly possible that the rPFC is not the ‘site’ of prediction error per se but may be concerned rather with inferences that are made as a consequence of prediction-error signal. This observation may offer an explanation of why it is possible to have, as we observed here, apparently preserved learning of associations in the face of such frontal perturbation. Specifically, subjects could learn simple associative relationships but show a change in the extent or nature of inference about the food–allergy relationships as a whole. This would account for the apparently normal behavioural learning profiles under these conditions. Further work would be required to establish whether inferential processing in patients is truly altered in the patients.

Extending this argument, it is relevant to note that humans often employ short-cuts or heuristics when making causal judgements (Kahneman et al., 1982). One basis for causal inference is the ease with which a plausible scenario can be constructed mentally. The prefrontal associative learning mechanism outlined by Daw and colleagues might underlie such construction or simulation (Daw et al., 2005). In response to a prediction-error signal, the prefrontal cortex may interrogate the possible associative chains of causal cues that may have led to the unexpected outcome and drive, where necessary, the formation of novel cue–outcome associations. Hyperactive engagement of this interrogation process may lead psychotic patients to form and strengthen inappropriate causal associations during delusion formation (Maher, 1974; Killslstrom and Hoyt, 1988; Gray et al., 1991; Hemsley, 1994; Corlett et al., in press). Of course, the theory under test here is relevant to the early formation of delusions and it does not attempt to account for the complexity and fixity of such beliefs.

An alternative possibility, one which we attempt to address empirically, is that patients perform the task with an alternative approach, underpinned by different neural mechanisms, and, using this mechanism, they achieve matched performance with control subjects. We did identify a region of right middle temporal cortex that was more active in patients than in controls, providing a candidate alternative neural mechanism. However, this region was identified at a less stringent statistical threshold and its activity in patients was not related to successful task performance. Based on these limitations on interpretation, we maintain that fMRI responses to associative learning and expectancy violation provide a more sensitive assessment of performance than subjects’ behavioural predictions, a position at which we have arrived based on own empirical work using this task (Corlett et al., 2004, 2006) and the neuroimaging findings of others (see Fletcher, 2004).

Violation of learned expectancy is also inherent in quite different tasks that have been used to explore psychosis and schizophrenia. Studies exploring models of cognitive control emphasize the interaction between phasic responses in subcortical dopamine neurons and more sustained firing in the prefrontal cortex (Cohen et al., 1996; Braver and Cohen, 1999; Miller and Cohen, 2001; Rougier et al., 2005). The prediction-error signal from the midbrain is responsible for ‘gating’ the access of information to the prefrontal cortex. Noise in this gating system would cause disruptions in manipulation and maintenance of information necessary for goal-directed behaviour and would disrupt attention and motivation as is observed in schizophrenia (Braver and Cohen, 1999). Indeed the processing of cortical noise as relevant signal has been implicated as a pathophysiologically mechanism in schizophrenia (Winterer and Weinberger, 2004; Winterer, 2006).

There is a wealth of evidence implicating mesocortico-limbic dopamine neurotransmission both in prediction-error processing (Schultz, 2000; Waelti et al., 2001) and in psychosis (Carlsson and Lindqvist, 1963; Anden et al., 1970; Nyback and Sedvall, 1970; Creese et al., 1976; Seeman et al., 1976). However, glutamate function may also be involved in both the pathophysiology of psychosis (Javitt and Zukin, 1991; Krystal et al., 1994; Jentsch and Roth, 1999; Goff and Coyle, 2001; Laruelle et al., 2003) as well as in the signalling of prediction errors (Lavin et al., 2005). Furthermore, these two neurotransmitter systems are intimately involved with each other, such that a disruption to one can have profound effects on the other, precipitating psychotic symptoms (Grace, 1991, 1993; Moore et al., 1999). The pattern and regional specificity of disruption to prediction-error processing in psychotic patients (Fig. 6) is strikingly redolent of our previously observed (Corlett et al., 2006) disruption of rPFC response under low dose ketamine (an NMDA receptor antagonist that is increasingly being employed as an experimental model of psychosis, see Krystal et al., 2002, for review). The strong similarity between the current findings and our previous study of ketamine provide support for the NMDA receptor hypofunction model of psychosis (Javitt and Zukin, 1991; Olney and Farber, 1995; Carlsson et al., 1999; Olney et al., 1999). However, the inferences are not straightforward. A challenge is to understand why, if glutamate is critical to prediction-error signal, NMDA hypofunction is associated with abnormally high prediction-error-related firing. One possibility is that NMDA receptor hypofunction increases extracellular glutamate levels in PFC via GABAergic disinhibition of cortical afferents including subcortical and midbrain dopamine neurons (Moghaddam et al., 1997). This effect would induce stimulation of cortical AMPA receptors, a possibility that finds support in the observation that an NMDA antagonist increases the number of randomly distributed single spikes in prefrontal neurons of awake rats (Jackson et al., 2004). Jackson and colleagues posit that this increased random spiking is induced by AMPA receptor stimulation. Bringing these observations together, if it is indeed the case that NMDA hypofunction (under conditions of psychosis or ketamine), increases cortical noise via AMPA receptor stimulation thus impairing the filtering of
irrelevant information and promoting the transmission of misinformation (Lisman, 1997), these effects would be consistent with our current and previous low dose ketamine findings (Corlett et al., 2006): an augmentation of the rPFC response to task irrelevant control items (a result of the increased prefrontal noise due to random spiking).

While we have focused upon the relationship between prediction error and the earliest emergence of false beliefs, it should be noted that other broader models have also explained other symptoms of psychosis, notably hallucinations, in terms of comparable underlying mechanisms (e.g. Kapur, 2003). In this respect, it is noteworthy that while, in the current study, perturbed prediction-error signal was related statistically to delusions, the same relationship was not found for hallucinations. However, given that our experimental design and analysis focused primarily on false beliefs, we are very cautious about interpreting the latter negative finding.

In summary, we used a brain marker for prediction-error-dependent causal learning to explore aberrant responses in psychosis. Psychotic patients demonstrate a disruption in prediction error, the magnitude of which correlates, across patients, with the severity of their delusions. Our findings provide support for associative models of delusion formation and provide a possible mechanism for the disruptions that underlie the emergence of psychotic beliefs. This mechanism incorporates disrupted neurotransmission in the mesocorticolimbic dopamine system, particularly the mesocortical pathway from VTA to PFC, known to corelease both dopamine and glutamate at its terminals in response to salient environmental events (Lavin et al., 2005). The present data suggest that this signalling process may be impaired in individuals with psychosis, such that the prefrontal cortex responds to physiological noise as if it were salient biological signal and drives the attribution of salience and attention to irrelevant and inconsequential environmental events.

The observation that disrupted prefrontal processing of expectancy violations correlates with delusion severity in our patient sample [in a manner redolent of that observed in healthy individuals administered ketamine (Corlett et al., 2006)], implicates aberrant prediction-error processing in delusion formation and suggests that glutamatergic neurotransmission contributes to endogenous psychosis.

Acknowledgements
This work was supported by the Wellcome Trust and conducted within the University of Cambridge Behavioural and Clinical Neuroscience Institute, supported by a joint award from the Medical Research Council and the Wellcome Trust. GKM is supported by the Department of Health. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

References
Braver TS, Cohen JD. Dopamine, cognitive control, and schizophrenia: the gating model. Prog Brain Res 1999; 121: 327–49.
Fletcher PC, Anderson JM, Shanks DR, Honey R, Carpenter TA, Donovan T, et al. Responses of human cortical to surprising