



Atypical processing of uncertainty in individuals at risk for psychosis

David M. Cole^{a,b,1,*}, Andreea O. Diaconescu^{a,c,d,1,*}, Ulrich J. Pfeiffer^e, Kay H. Brodersen^a, Christoph D. Mathys^{a,f,g}, Dominika Jolkowski^e, Stephan Ruhrmann^e, Leonhard Schilbach^{h,i,j,k,l}, Marc Tittgemeyer^{m,n}, Kai Vogeley^{e,o}, Klaas E. Stephan^{a,m,p}

^a Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland

^b Department of Psychiatry, Psychotherapy and Psychosomatics, University of Zurich, Psychiatric Hospital of the University of Zurich, Zurich, Switzerland

^c Department of Psychiatry (UPK), University of Basel, Basel, Switzerland

^d Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), University of Toronto, Toronto, Canada

^e Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany

^f Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy

^g Interacting Minds Centre, Aarhus University, Aarhus, Denmark

^h Independent Max Planck Research Group for Social Neuroscience, Max Planck Institute of Psychiatry, Munich, Germany

ⁱ Graduate School for Systemic Neuroscience, Munich, Germany

^j International Max Planck Research School for Translational Psychiatry, Munich, Germany

^k Ludwig-Maximilians-Universität München, Munich, Germany

^l Kliniken der Heinrich-Heine-Universität/LVR-Klinik Düsseldorf, Düsseldorf, Germany

^m Max Planck Institute for Metabolism Research, Cologne, Germany

ⁿ Cologne Cluster of Excellence in Cellular Stress and Aging associated Disease (CECAD), Germany

^o Institute for Neuroscience and Medicine – Cognitive Neuroscience (INM3), Research Center Juelich, Juelich, Germany

^p Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom

ARTICLE INFO

Keywords:

At-risk mental state
Computational psychiatry
Decision-making
Hierarchical Bayesian learning
Prodromal
Volatility

ABSTRACT

Current theories of psychosis highlight the role of abnormal learning signals, i.e., prediction errors (PEs) and uncertainty, in the formation of delusional beliefs. We employed computational analyses of behaviour and functional magnetic resonance imaging (fMRI) to examine whether such abnormalities are evident in clinical high risk (CHR) individuals.

Non-medicated CHR individuals ($n = 13$) and control participants ($n = 13$) performed a probabilistic learning paradigm during fMRI data acquisition. We used a hierarchical Bayesian model to infer subject-specific computations from behaviour – with a focus on PEs and uncertainty (or its inverse, precision) at different levels, including environmental ‘volatility’ – and used these computational quantities for analyses of fMRI data.

Computational modelling of CHR individuals’ behaviour indicated volatility estimates converged to significantly higher levels than in controls. Model-based fMRI demonstrated increased activity in prefrontal and insular regions of CHR individuals in response to precision-weighted low-level outcome PEs, while activations of prefrontal, orbitofrontal and anterior insula cortex by higher-level PEs (that serve to update volatility estimates) were reduced. Additionally, prefrontal cortical activity in response to outcome PEs in CHR was negatively associated with clinical measures of global functioning.

Our results suggest a multi-faceted learning abnormality in CHR individuals under conditions of environmental uncertainty, comprising higher levels of volatility estimates combined with reduced cortical activation, and abnormally high activations in prefrontal and insular areas by precision-weighted outcome PEs. This atypical representation of high- and low-level learning signals might reflect a predisposition to delusion formation.

* Corresponding author at: Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric Hospital of the University of Zurich, Lenggstrasse 31, PO Box 1931, 8032 Zurich, Switzerland.

E-mail address: dcole.neurosci@gmail.com (D.M. Cole).

¹ These authors contributed equally.

<https://doi.org/10.1016/j.nicl.2020.102239>

Received 9 December 2019; Received in revised form 24 February 2020; Accepted 6 March 2020

Available online 07 March 2020

2213-1582/ © 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

In standard classification schemes such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Statistical Classification of Diseases and Related Health Problems (ICD), schizophrenia and related psychotic disorders are defined as syndromes, i.e., combinations of clinical symptoms and signs. It is widely acknowledged that this phenomenological definition of schizophrenia (and other mental disorders) amalgamates heterogeneous groups of patients with possibly different pathophysiological mechanisms (van Os and Tamminga, 2007; Insel, 2010; Owen, 2014; Stephan et al., 2016). This may explain the diversity of clinical trajectories and treatment responses across patients and calls for new approaches to dissect the schizophrenia spectrum into subgroups or dimensions (Stephan et al., 2009a; Insel et al., 2010; Kapur et al., 2012; Cuthbert and Insel, 2013). One potential avenue to obtaining a formal understanding of differences in disease mechanisms across patients is the deployment of mathematical (specifically: generative) models that can be applied to non-invasive measures of behaviour and brain activity (Frässle et al., 2018); in psychiatry, the clinical application of this translational neuroimaging approach is referred to as “computational psychiatry” (Montague et al., 2012; Wang and Krystal, 2014; Stephan et al., 2015; Adams et al., 2016). A computational approach to phenotyping patients in a more fine-grained manner may be particularly relevant for the early detection of individuals at risk, an increasingly important domain of psychosis research (Klosterkotter et al., 2011; Fusar-Poli et al., 2013a, 2015; Moorhead et al., 2013; Koutsouleris et al., 2015; McGuire et al., 2015; Nieman and McGorry, 2015).

The clinical high risk (CHR) state is defined by the presence of either attenuated psychotic symptoms, brief and self-limiting psychotic symptoms, or a significant reduction of function under a family history of schizophrenia (Fusar-Poli et al., 2013a). It is a construct pertaining to the pre-psychotic phase, before a formal diagnosis (according to DSM or ICD) can be made (Yung et al., 2005). Experimental studies and clinical trials of CHR individuals are often logistically challenging, for several reasons. For example, a diversity of strategies for diagnosis and clinical management exists across centres, and the unresolved question of whether CHR symptoms mandate the use of antipsychotics (Wood et al., 2011; Kahn and Sommer, 2015; Nieman and McGorry, 2015) limits the availability of non-medicated CHR individuals for research. Furthermore, not all CHR individuals seek medical help and, if they do, they may not always be recognised due to frequent comorbidity, e.g., mood impairments may overshadow psychotic symptoms (Falkenberg et al., 2015a).

These issues lead to considerable difficulties in recruiting sufficiently large, homogeneous samples of medication-free (particularly antipsychotic-naïve) patients for clinical studies (Ruhrmann et al., 2010; Klosterkotter et al., 2011; Fusar-Poli et al., 2013a, 2013b; Nieman and McGorry, 2015). As a consequence, we still have a very limited understanding of the neurobiological mechanisms that produce CHR symptoms and the subsequent transition to full-blown psychosis (Tsuang et al., 2013). A formal computational account of the cognitive and neurophysiological aberrancies underlying the CHR state would be highly beneficial, both for future research on pathophysiology and clinical studies.

In recent years, theories on the development and maintenance of psychotic symptoms have become increasingly enriched by testable computational mechanisms and are beginning to converge on a few central themes (Corlett et al., 2010). One major theory postulates that the attribution of “aberrant salience” to objectively uninformative or neutral events fuels the formation of delusions (Kapur, 2003). This framework posits a key role for dopamine (DA) in mediating the misattribution of salience in psychosis, consistent with longstanding theories of dopaminergic dysfunction in schizophrenia (Grace, 1993; Laruelle et al., 1996; Howes and Kapur, 2009) and specifically the idea that contextually inappropriate phasic DA release triggers maladaptive

plasticity and learning (King et al., 1984; Heinz and Schlagenhauf, 2010; Roiser et al., 2013; Winton-Brown et al., 2014). This notion is consistent with a number of experimental findings in schizophrenia, most of which correlate with positive symptoms, including enhanced learning for neutral cues (assessed via behavioural and autonomic responses) and increased activation of dopaminergic and dopaminoceptive regions, including the ventral striatum and mid-brain, in response to neutral cues (Jensen et al., 2008; Murray et al., 2008; Roiser et al., 2009; Romaniuk et al., 2010; Diaconescu et al., 2011).

Computational treatments of aberrant salience have examined this phenomenon in relation to prediction errors (PEs) about rewarding or novel outcomes. This is motivated by the putative relation of outcome-related PEs to phasic DA release and possible involvement in dysfunctional learning in psychosis and schizophrenia (Schultz et al., 1997; Pessiglione et al., 2006; Corlett et al., 2007, 2009a, 2010; Murray et al., 2008; Gradin et al., 2011; Adams et al., 2013; Ermakova et al., 2018). The dopaminergic PE signal is thought to represent a neural response to deviation from an expected outcome (of rewards but also sensory features; Iglesias et al., 2013; Gardner et al., 2018; Suarez et al., 2019) and likely supports the updating of beliefs about the environment by induction of synaptic plasticity (Montague et al., 2004), for example, via modulation of NMDA receptors (Gu, 2002). However, predictions are inevitably uncertain, and PEs should carry different weight, depending on the precision (the mathematical inverse of uncertainty) of the underlying prediction. Computationally, Bayesian frameworks offer a formal account of this intuitive notion (Rao and Ballard, 1999; Friston, 2008). These theories view perception and learning as a hierarchically organised process, in which beliefs at multiple levels, from concrete (e.g., specific stimuli) to more abstract aspects of the environment (e.g., probabilities and volatility), are updated based on level-specific PEs and precisions. Specifically, under fairly general assumptions (i.e., for all probability distributions from the exponential family) a ratio of precisions (of bottom-up input vs. prior beliefs) serves to scale the amplitude of PE signals and thus their impact on belief updates (Mathys et al., 2011; Mathys, 2013); see Eq. (1) below.

Recent theories of perceptual abnormalities in schizophrenia have built on Bayesian accounts of this sort, enriching traditional concepts of aberrant salience with the crucial role of uncertainty (Stephan et al., 2006; Corlett et al., 2009a, 2010; Fletcher and Frith, 2009; Adams et al., 2013). One specific suggestion from these accounts is that chronically over-precise low-level PE signals may be the starting point of delusion formation, as they continue to induce unusual belief updates, without diminishing over time. Put differently, persistently surprising events may eventually require adoption of extraordinary higher-order beliefs to be explained away (Corlett et al., 2007, 2010). Alternatively, high-order beliefs may be of abnormally low precision (Adams et al., 2013; Sterzer et al., 2018; Diaconescu et al., 2019), leading to lack of regularisation, which renders the environment seemingly unpredictable (e.g., extremely volatile) and enhances the weight of low-level precision-weighted PE. Notably, these explanations are not exclusive but could co-exist (specifically, they relate to the numerator and denominator of the precision ratio in Eq. (1) below).

Here we investigated the presence of these putative abnormalities during learning and decision-making under environmental uncertainty (volatility) in the behaviour and brain activity of CHR individuals. To this end, we combined fMRI of a probabilistic learning task under volatility with computational modelling, hierarchical Gaussian filtering (HGF), which emphasises the importance of uncertainty for updating a hierarchy of beliefs via precision-weighted PE signals (Mathys et al., 2011, 2014).

2. Methods and materials

2.1. Participants

Thirteen CHR individuals (mean age 21.2 years \pm 3.3 s.d.; 4 female) and thirteen healthy controls (mean age 27.2 \pm 2.7; 7 female) were included in the study. Given the slight differences in group composition, all statistical comparisons between the CHR participants and controls in the fMRI analyses described below (Section 2.7) were conducted with age and sex as covariates, in order to control for possible confounding influences on neurobiology. Notably, all of the CHR individuals were completely antipsychotic-naïve; none had any lifetime history of exposure to any antipsychotic medications at the time of testing. Four of the CHR individuals (2 female) had previously received medication for other mental health issues unrelated to psychosis. No CHR individuals had any history of neurological disorder, and none of the healthy controls had any history of neurological or psychiatric disorder. All CHR individuals were part of the sample seeking help at the early detection and intervention centre (FETZ) at the University Hospital of Cologne. In order for an individual to be included in the study CHR group, one or more of the following criteria of the Comprehensive Assessment of At-Risk Mental States (CAARMS; for details see Yung et al., 2005) had to be fulfilled: attenuated positive symptoms (APS), brief limited intermittent positive symptoms (BLIPS), or at least one of two risk factors for psychosis (i.e., a positive family history of psychosis or the patient had schizotypal personality disorder as defined by DSM-IV) in combination with a significant decline in psychosocial functioning maintained over the last month. The exclusion criteria were past or present diagnosis of psychotic disorders, substance abuse or dependence according to DSM-IV criteria, past or present use of antipsychotics, and any contraindications related to magnetic resonance imaging measurements. A summary of demographic and clinical data is provided in Table 1. All participants provided written, informed consent to participate in the experiment, which was approved by the local ethics committee of the Medical Faculty of the University of Cologne (Cologne, Germany).

The clinical status of CHR individuals was established via a checklist of inclusion criteria and clinical tests including the Structured Interview for Prodromal Symptoms (SIPS; Miller et al., 2002, 2003). One male individual in the patient group did not have his CHR status upheld upon further clinical examinations at a standardised follow-up period of two years; his data were thus excluded from all between-group fMRI analyses. Of the other twelve CHR individuals, three had transitioned to a diagnosed psychotic disorder by follow-up, while three more had withdrawn from the study before their follow-up status could be determined. Scores on sub-scales of the SIPS assessing positive symptoms and global assessment of functioning (SIPS-GAF) were compared with data from the behavioural task and with a representative fMRI measure

of aberrant PE encoding in CHR, described below (Section 2.8), via Pearson correlation analyses with one-tailed hypothesis testing, due to the explicit directionality of our predictions that atypical neural representation of PEs would be associated with worsened clinical functioning under CHR.

2.2. Behavioural task

While undergoing fMRI, each participant completed a probabilistic learning task (Fig. 1A), which required trial-wise binary decisions between two fractal stimuli. On each trial, the same pair of fractal stimuli were presented, each paired with a reward value (points) that varied independently of the task contingency structure. Accrued points were converted to monetary reward at the end of the experiment. During cue presentation, participants had a maximum of 4 s to make a decision, followed by a 5-second delay displaying the choice and a 2-second presentation of the decision outcome (correct or incorrect). If no response was made during the decision period, a time-out occurred and a blank screen was displayed for 8 s, instead of the delay and outcome screens. The inter-trial interval was jittered between 5 and 7 s. The across-trial task structure incorporated ‘switches’, or reversals, in a block-wise fashion in terms of which of the cues was most likely (80% probability) or least likely (20%) to be the correct, rewarded choice on that trial. It also incorporated a change of contingencies over time; this volatility (or variance per unit time) induces higher-order (environmental) uncertainty – about the probabilistic structure of the task, in addition to informational uncertainty by trial-wise outcomes – and determines the temporal evolution of a subject's learning rate (Behrens et al., 2007; Mathys et al., 2014). The task consisted of 160 trials, with contingency blocks comprising between 14 and 46 trials depending on which volatility pseudo-block they were contained within (Fig. 1B). The task structure was fixed across all participants in terms of the correct, rewarded choice on a given trial. Two early participants (one patient and one control, both male) completed a slightly varying version of the task that differed only in that there was no time-out for responses; this did not lead to longer data acquisition in the majority of trials, neither did it impact in any way the behavioural modelling described in Section 2.3.

2.3. Behavioural modelling

The computational framework adopted in this study was guided by Bayesian theories of brain function that suggest that the brain maintains and continuously updates a generative model of its sensory inputs (Dayan et al., 1995; Rao and Ballard, 1999; Friston, 2005). In other words, individuals are thought to update their beliefs about states of the external world based on the sensory inputs they receive (perceptual model); these beliefs, in turn, provide a foundation for making decisions (response model; see Daunizeau et al., 2010).

A number of different hypotheses about how humans learn about probabilistic stimulus-outcome contingencies were embodied in the following model space (Fig. 2A). With regard to the perceptual model, our main question was whether subjects' models of reward probabilities based on stimulus-outcome associations had a hierarchical structure and accounted also for the volatility of these associations. We thus compared (i) the classical Rescorla-Wagner (RW) reinforcement learning model (Rescorla and Wagner, 1972), in which predictions evolve as a function of PE and a constant learning rate (model M_1) to three Bayesian models of learning, which included (ii) a non-hierarchical model, based on a reduced Hierarchical Gaussian Filter (HGF) that assumes that subjects do not infer on the volatility of stimulus-outcome probabilities (M_2 ; see Diaconescu et al., 2014), (iii) a three-level ‘canonical’ HGF ($M_{3,4}$; see Mathys et al., 2011), and (iv) a three-level ‘mean-reverting’ HGF in which volatility estimates drift towards a subject-specific equilibrium ($M_{5,6}$).

With respect to the response model, we followed previous work

Table 1
Summary of demographic, behavioural and clinical variables (mean \pm s.d.).

| | Measure | CHR (n = 13) | Controls (n = 13) |
|------------------------|---------------------------|-------------------|-------------------|
| Demographics | Age | 21.2 \pm 3.3 | 27.2 \pm 2.7 |
| | Sex (M:F) | 9:4 | 6:7 |
| Behavioural parameters | κ | 1.07 \pm 0.20 | 1.13 \pm 0.19 |
| | ω | -5.48 \pm 1.74 | -5.05 \pm 2.27 |
| | θ | 0.30 \pm 0.04 | 0.31 \pm 0.05 |
| | β | 9.06 \pm 4.26 | 7.03 \pm 2.91 |
| | m_3 | 2.49 \pm 0.36 | 2.21 \pm 0.35 |
| | $\mu_3^{(k=0)}$ | 1.29 \pm 0.57 | 1.80 \pm 0.56 |
| Clinical variables | SIPS positive symptom sum | 6.85 \pm 3.05 | |
| | SIPS negative symptom sum | 11.62 \pm 3.66 | |
| | SIPS-GAF | 62.54 \pm 10.14 | |
| | SIPS total | 31.23 \pm 6.47 | |

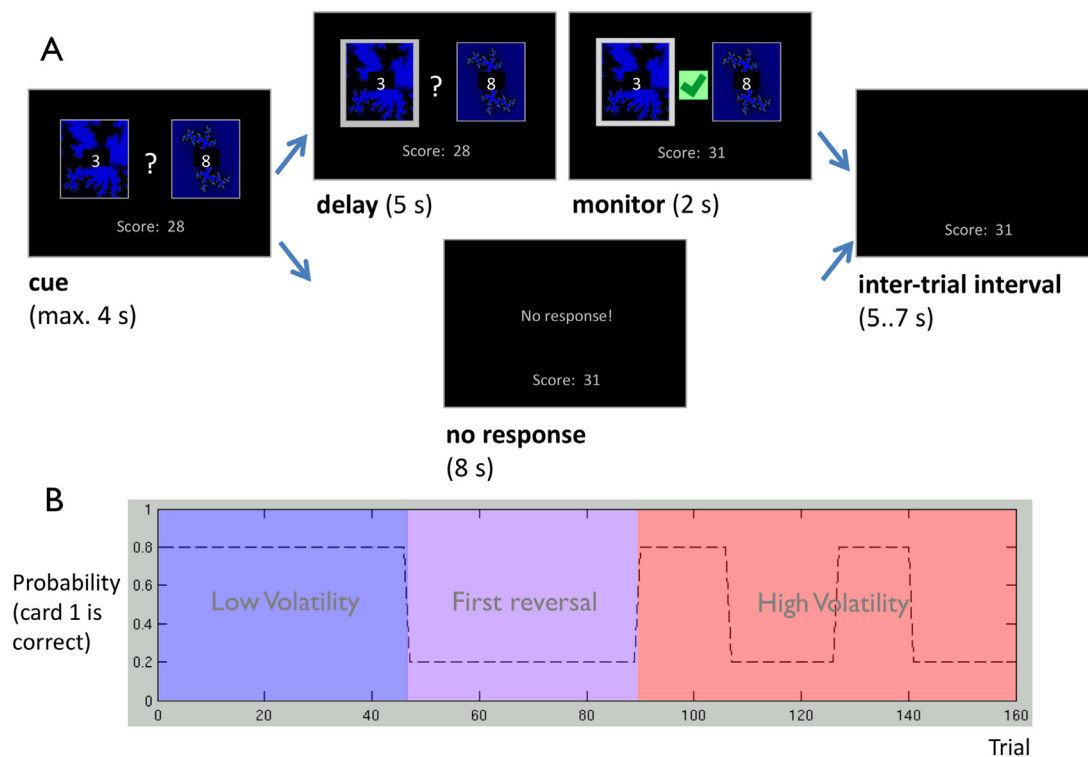


Fig. 1. Probabilistic reversal-learning task. The behavioural paradigm consisted of: (A) within trial, a pair of fractal stimuli, each paired with a reward value, requiring a decision from the participant via button press in order to obtain the reward; (B) across trials, the probabilistic contingency (dotted line) of which of the two fractal cues was most likely to yield a reward occasionally underwent ‘reversal’, the regularity of which engendered pseudo-blocks of volatility modulation (blue, violet and red panels). The reward values within trials were entirely independent of the stimulus-outcome contingencies.

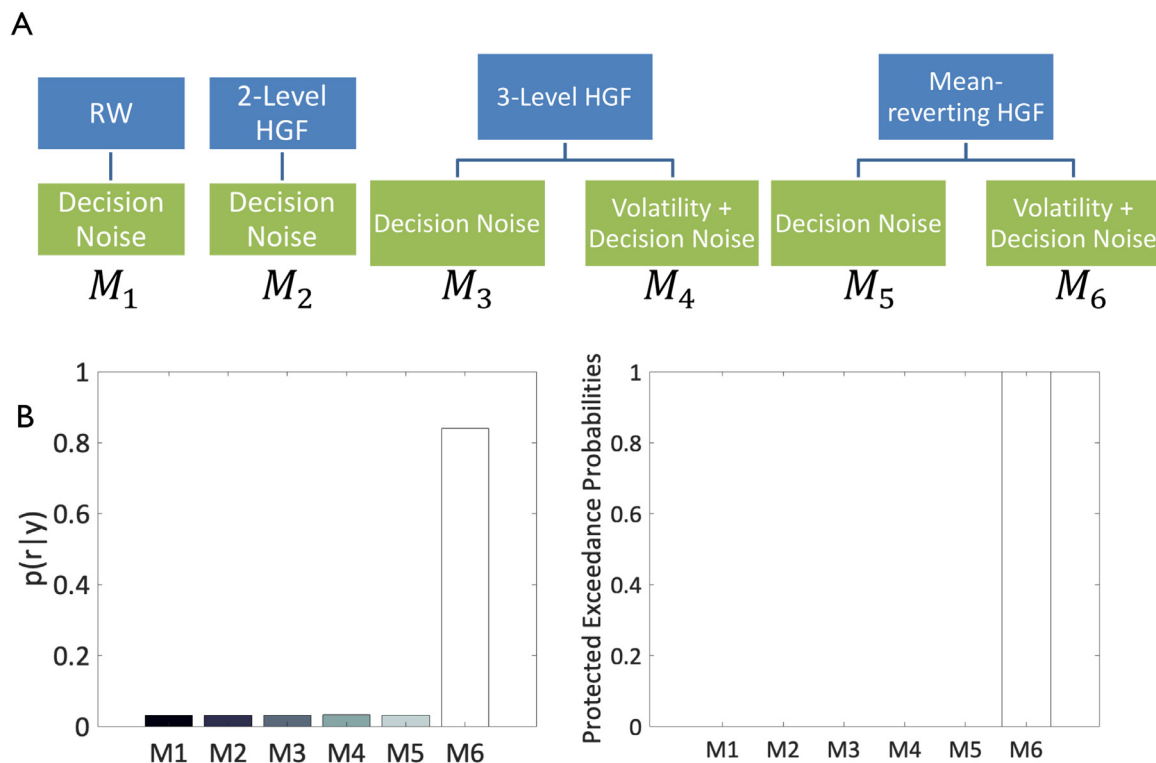


Fig. 2. Hierarchical structure of the model space: perceptual models, response models and Bayesian model selection. (A) The models considered in this study have a factorial structure that can be displayed as a tree: The nodes at the first level represent the perceptual model families (RW, 2-level non-volatility HGF, 3-level HGF, and 3-level mean-reverting). The nodes at the second level represent the individual models. Two response model families were formalized under the HGF models: the mapping of beliefs-to-decisions either (i) depended dynamically on the estimated volatility of the learning environment (“Volatility + decision noise” model) or (ii) was a fixed entity over trials (“Decision noise” model). (B) Bayesian model selection (BMS) reveals M_6 , the mean-reverting HGF perceptual model in combination with the “Volatility” decision model, to best explain the data.

(Diaconescu et al., 2014) and considered two possible mechanisms of how beliefs were translated into responses (Fig. 2A). Subjects' choices could either be affected by fixed decision noise ("Decision noise" model family; $M_{1,3,5}$) or the decision noise might vary trial-by-trial as a function of the estimated volatility of the stimulus-outcome probabilities ("Volatility" model family; $M_{4,6}$).

2.3.1. Perceptual model: the Hierarchical Gaussian Filter

The HGF is a hierarchical model of learning, which allows for inference on an agent's beliefs (and their uncertainty) about the state of the world from observed behaviour (Mathys et al., 2011) and has been used by several recent behavioural and neuroimaging studies on different forms of learning (Iglesias et al., 2013; Diaconescu et al., 2014, 2017; Hauser et al., 2014; Vossel et al., 2014; Schwartenbeck et al., 2015; Lawson et al., 2017; Powers et al., 2017). The model proposes that agents infer on the causes of the sensory inputs using hierarchically-coupled belief updates that evolve in time as Gaussian random walks where, at any given level, the variance (step size) is controlled by the state of the next higher level (Mathys et al., 2011, 2014). A standard formulation of the HGF for standard binary decision making tasks includes three levels, where the first (lowest) level encodes the probability of a trial outcome (here: whether a stimulus was rewarded or not), the 2nd level represents the tendency of a stimulus to be rewarded as a continuous quantity (in logit space), and the 3rd level represents the volatility of this probability (Fig. 3; see also Mathys et al., 2011).

The following subject-specific parameters determine how the above states evolve in time: (i) κ determines the degree of coupling between the second and third level in the hierarchy (x_2 and x_3) and the degree to which the volatility estimate influences the subject's uncertainty about the stimulus-reward probabilities; (ii) ω represents a constant (tonic) component of the log-volatility of x_2 , capturing the subject-specific magnitude of the belief update about the stimulus-outcome

probabilities that is independent of x_3 ; (iii) ϑ is a meta-volatility parameter and determines the evolution of x_3 , or how rapidly the volatility of the associations changes in time. Furthermore, we also estimated $\mu_3^{(k=0)}$, the subject's initial belief about volatility of the outcome probabilities.

A key notion of the HGF is that subjects update their beliefs about hierarchically coupled states in the external world by using a variational approximation to intractable full Bayesian inference (Mathys et al., 2011). The update rules that emerge from this approximation have a structural form similar to RW reinforcement learning, but with a dynamic (adaptive) learning rate determined by the next-higher level in the hierarchy. Formally, at each hierarchical level i , predictions (posterior means $\mu_i^{(k)}$) on each trial k are proportional to precision-weighted PEs, $\varepsilon_i^{(k)}$ (Eqs. (1) and (2)). The general form of this belief update (with subtle differences for categorical quantities at the lowest level) is the product of the PE from the level below $\delta_{i-1}^{(k)}$, weighted by a precision ratio $\psi_i^{(k)}$:

$$\Delta\mu_i^{(k)} \propto \psi_i^{(k)} \delta_{i-1}^{(k)} \quad (1)$$

$$\text{where } \psi_i^{(k)} = \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \quad (2)$$

Here, $\hat{\pi}_{i-1}^{(k)}$ and $\pi_i^{(k)}$ represent estimates of the precision of the prediction about input from the level below (e.g., precision of sensory data) and of the prediction at the current level, respectively (for details, see Mathys et al., 2011). This precision-weighting is critical for adaptive learning and emerges naturally from hierarchical Bayesian formulations (Friston, 2008; Corlett et al., 2010; Mathys et al., 2011; Adams et al., 2013; Iglesias et al., 2013). Simply speaking, PEs have a larger weight (and thus updates are more pronounced) when the precision of the data (input from the lower level) is high, relative to the precision of prior beliefs.

Level 3: Phasic Volatility

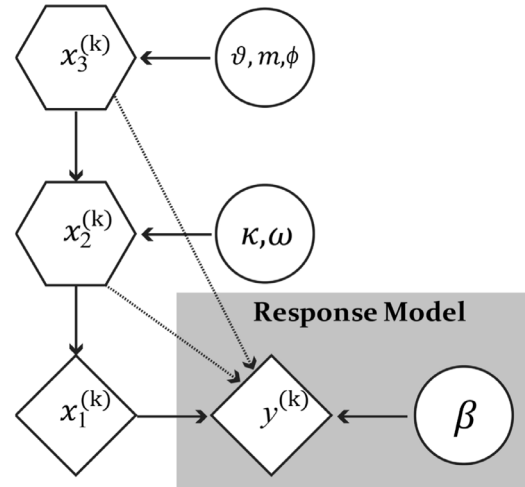
$$p(x_3^{(k)}) \sim \mathcal{N}(x_3^{(k-1)} + \phi(m_3 - x_3^{(k-1)}), \vartheta)$$

Level 2: Tendency of stimulus to be rewarding

$$p(x_2^{(k)}) \sim \mathcal{N}(x_2^{(k-1)}, e^{(\kappa x_3^{(k-1)} + \omega)})$$

Level 1: Observations (accurate or inaccurate)

$$p(x_1^{(k)} = 1) = \frac{1}{1 + e^{-x_2^{(k)}}}$$



Response Model:

$$p(y^{(k)} = 1 | \mu_1^{(k)}) = \frac{\mu_1^{(k)\beta}}{\mu_1^{(k)\beta} + (1 - \mu_1^{(k)})^\beta}$$

Fig. 3. Graphical representation of the winning model combination: “mean-reverting HGF” perceptual model and the “Volatility” response model. In this graphical notation, circles represent constants and diamonds represent quantities that change in time (i.e., that carry a time/trial index). Hexagons, like diamonds, represent quantities that change in time, but additionally depend on the previous state in time in a Markovian fashion. x_1 represents the cue probability, x_2 the cue-outcome contingency and x_3 the volatility of the cue-outcome contingency. Parameter κ determines how strongly x_2 and x_3 are coupled, ω determines the log-volatility or tonic component of x_2 , ϑ represents the volatility of x_3 , and m represents the mean of the drift towards which x_3 regresses to in time. The response model parameter β represents the inverse decision temperature and determines the belief-to-response mapping.

2.3.2. Mean-reverting HGF

The standard HGF, described above, already allows for representing (and inferring) the precision of low-level PEs and prior beliefs and thus offers the two components required to test our hypothesis. We can refine this model further by using a variation of the classical HGF that allows for inferring on drifts in a subject's beliefs towards an equilibrium m (essentially the equivalent of an Ornstein-Uhlenbeck process in discrete time). Here, we used this approach to examine the possibility that CHR individuals might be characterised by a tendency to over-estimate the volatility of the environment, which would further enhance the weight (precision) of low-level PEs and lead to an inflation of uncertainty about probabilities over time. As described above, a scenario of this sort may lead to later compensation, for example by adopting high-order beliefs with inappropriately high certainty, and may thus represent a risk factor for delusion formation.

The equations describing the generative model are summarised in Fig. 3. Notably, the third level of this model includes subject-specific parameters that represent the agent's individual starting estimate of volatility, $\mu_3^{(k=0)}$, as well as an equilibrium value, m_3 , which the agent's estimate of volatility drifts towards and remains attracted to (Fig. 3). The prior on m_3 was equivalent to the prior on $\mu_3^{(k=0)}$ (see Supplementary Table 1), hence the model did not include any prior assumption about the direction in which the equilibrium of volatility estimates would be located, relative to the starting value of volatility. The higher the value of m_3 , the higher the agent's average level of estimated volatility and thus the greater the uncertainty of his/her estimates of environmental probabilities over time. The hypothesis described above – that CHR individuals are characterised by enhanced precision-weighting of low-level PEs – implies that either the estimated precision of sensory input is abnormally high or that the precision of beliefs is abnormally low (see Eq. (2)). In the context of our task, the latter option corresponds to increased uncertainty about cue-outcome contingencies. This increase in uncertainty would be greater, the higher the value of m_3 and thus the higher the average level of estimated volatility, $\hat{\mu}_3^{(k)}$, over time.

2.3.3. Response model

The response model embodies a (probabilistic) mapping from the agent's beliefs to their decisions (Daunizeau et al., 2010). The probability that the subject behaves according to his/her prediction of the outcome probabilities was described by a sigmoid function (Eq. (3)):

$$p(y^{(k)} = 1 | \mu_1^{(k)}) = \frac{\mu_1^{(k)\beta}}{\mu_1^{(k)\beta} + (1 - \mu_1^{(k)})^\beta} \quad (3)$$

Here, β represents the inverse of the decision temperature: as $\beta \rightarrow \infty$, the sigmoid function approaches a step function with a unit step at $b^{(k)} = 0.5$ (i.e., no decision noise). As described above, we considered two types of this belief-to-response mapping. The first “Decision noise” model family assumes constant decision noise; that is, β becomes a subject-specific free parameter. By contrast, the “Volatility” response model family proposes that the decision temperature parameter β might vary trial-by-trial with the estimated volatility, $e^{-\mu_3^{(k)}}$, such that the larger the volatility, the lower the (inverse) decision temperature and the higher the decision noise (see Diaconescu et al., 2014 for details). In other words, the more stable the stimulus probabilities, the more deterministic the participant's belief-to-response mapping.

Using priors over the model parameters based on previous studies with similar probabilistic learning paradigms (Iglesias et al., 2013; Diaconescu et al., 2014; Supplementary Table 1), maximum *a posteriori* (MAP) estimates of model parameters were obtained using the HGF toolbox version 3.0. This MATLAB-based toolbox is freely available as part of the open source software package TAPAS (<https://www.tnu.ethz.ch/en/software/tapas.html>).

2.3.4. Bayesian model selection and computational regressors

We compared the full set of resulting models M_{1-6} using Bayesian model selection (Stephan et al., 2009b), to determine which combination of perceptual and response models best explained the behavioural dataset and would thus optimally inform the subsequent analysis of fMRI data. Based on the model space outlined above (Fig. 2A), a total of six different models were compared.

From the winning model (Fig. 2B), we extracted the trajectories of several trial-wise computational quantities, estimated for each subject individually: (i) the prediction about the next outcome, (ii) uncertainty (Bernoulli variance) about the probability of the next outcome (“1st-level uncertainty”), (iii) their belief about the current volatility of the environment ($\hat{\mu}_3^{(k)}$), (iv) their absolute precision-weighted PE regarding the outcome on a given trial relative to their current beliefs about the probability of that outcome (ϵ_2), (v) their belief uncertainty (“2nd-level uncertainty”; σ_2) and (vi) their signed precision-weighted PE regarding the perceived volatility of the outcome on a given trial relative to their current belief about that volatility (ϵ_3). Each of these trajectories was then used as a regressor (parametric modulator) in the single-subject fMRI analyses described below (Section 2.7).

2.4. Behavioural analysis

We subjected the MAP estimates of the winning model to one-way analysis of variance (ANOVA) assessments, in order to test for differences in decision and learning parameters between CHR individuals and healthy controls. Our hypothesis described above implies that m_3 itself and/or the meta-volatility parameter ϑ would be significantly greater in the CHR group, suggesting that CHR participants in contrast to controls perceive an increased environmental volatility over time. To examine group differences in perceived (predicted) volatility induced by basic reversals of probabilistic contingency, we also performed a 2 (group: CHR, control) \times 3 (phase: stable, reversal, volatile) mixed-factor ANOVA to examine group-by-phase interaction effects on $\hat{\mu}_3^{(k)}$ (see the phases outlined in Fig. 1B). Additionally, we performed equivalent ANOVAs for ϵ_2 , σ_2 (informational uncertainty) and predicted environmental uncertainty (see Supplementary Materials for details; see also Eq. 27 in Mathys et al., 2011). The code for behavioural modelling and analysis is accessible via GitLab at <https://gitlab.ethz.ch/dandreea/apup>.

2.5. Image acquisition

Data were acquired on a 3 T Magnetom TIM Trio MRI scanner (Siemens, Erlangen, Germany) at the Max Planck Institute for Metabolism Research, Cologne. As the task was partially self-paced, a slightly different number of volumes was acquired for each subject (mean = 1217, approximately 40.5 min experiment duration). T2*-weighted echo-planar images (EPI) sensitive to blood-oxygenation level-dependant (BOLD) contrast were acquired during the task (TR = 2 s; TE = 30 ms; flip angle = 90°; 30 axial slices; in-plane resolution = 3.3 \times 3.3 mm; slice thickness = 2.7 mm; slice gap = 1.35 mm). Magnetic equilibration was accounted for via scanner automatic dummy removal. Images were acquired in parallel to the anterior-posterior commissural plane. Cardiac and breathing rates were recorded peripherally during scanning. Anatomical T1-weighted volumes were also acquired for each subject (TR = 1.9 s; TE = 3.51 ms; flip angle = 9°; field-of-view 256 \times 256 \times 128; voxel size 1.0 \times 1.0 \times 1.0 mm).

2.6. fMRI data preprocessing

Preprocessing of fMRI data was performed using FSL 5.0 (<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/>; Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, United Kingdom;

Smith et al., 2004). This incorporated standard steps of high-pass filtering (128 s cut-off), realignment of individual volumes to correct for head motion, removal of non-brain tissue from the images and Gaussian smoothing at 5 mm full-width half-maximum.

Following these initial steps, we performed single-subject data ‘denoising’: artefact classification and rejection based on single-session spatial independent component analysis (ICA) and machine learning techniques, using the FSL ‘MELODIC’ and ‘FIX’ tools (Beckmann and Smith, 2004; Salimi-Khorshidi et al., 2014). These steps were incorporated to remove any artefactual signal components that had survived conventional realignment and physiological noise correction (where available; see Section 2.7). Details are provided, along with a description of spatial normalisation procedures, in the Supplementary Materials.

2.7. fMRI analysis

Single-subject fMRI analyses were conducted using the general linear model (GLM) as implemented in SPM8 (<http://www.fil.ion.ucl.ac.uk/spm/software/spm8/>; Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom). Base regressors for the task were defined in terms of the onsets of the decision period, which had a variable duration (0–4 s) and the outcome period, which had a fixed duration (2 s). The decision period regressor was accompanied by three parametric modulator regressors encoding for the subject's trial-wise prediction of outcome, uncertainty at the 1st level of the HGF, and expected volatility at the 3rd level. The outcome period regressor was associated with three parametric modulators encoding for the absolute (unsigned) outcome-related precision-weighted PE (ε_2 ; (see Iglesias et al., 2013), uncertainty at the 2nd level (σ_2) and the volatility-related precision-weighted PE (ε_3). All parametric modulators were Z-normalised (zero mean, unit standard deviation) before entering into the GLM. Orthogonalisation was not performed. Temporal and dispersion derivatives of all regressors were added to the GLM of each subject in order to account for variability in the onset and width of hemodynamic responses. We performed physiological noise correction, based on RETROICOR (Glover et al., 2000) of cardiac and respiratory measurements (where available; 8 controls and 7 patients), using the PhysIO toolbox in TAPAS (Kasper et al., 2017) to compute 18 additional regressors that were included, along with six realignment parameters representing head motion, as confound regressors in the GLM for each subject.

Group analyses were conducted using second-level GLMs as implemented in SPM8. Outcome-related contrast estimates from the subject-level analysis were entered into a 2 (diagnostic groups: CHR vs. controls) \times 4 (outcome-related variables: base regressor, ε_2 , σ_2 (2nd-level uncertainty), ε_3) analysis of covariance (ANCOVA; unequal variance assumed). Age and sex were included as covariates of no interest. Decision-related contrast estimates were entered into a similar 2 (group: CHR vs. controls) \times 4 (decision-related variables: base regressor, prediction, 1st-level uncertainty, volatility estimate) analysis of covariance (ANCOVA).

Contrasts of interest at the group level examined, for each computational regressor, (i) the average activation across groups (CHR + controls) and (ii) significant differences between groups (CHR vs. controls). For the latter analyses, we conducted whole-brain comparisons but also used contrast-masking to restrict the analysis to regions showing significant average effects across groups (note that these are orthogonal contrasts, thus avoiding problems of non-independent inference). One exception to the latter approach was made in the analysis of group differences in ε_3 , where (iii) we restricted analyses to an anatomically defined masque of the anterior portion of the cingulate cortex (defined using the probabilistic volume in the Harvard-Oxford atlas provided with FSL and further masked by a study-specific group-level EPI template). This *a priori* masque, which included the basal forebrain, was motivated by previous observations of ε_3 -related

activation in the basal forebrain and anterior cingulate cortex (Iglesias et al., 2013; Diaconescu et al., 2017; see also Behrens et al., 2007). In line with the results of the same studies, we also employed (iv) a similar *a priori* region of interest analysis of ε_2 - and ε_3 -related activations using a masque of the dopaminergic midbrain. For each contrast, we corrected for multiple comparisons across the respective search volume – i.e., whole brain for (i), functional masks for (ii), and the anatomical masque for (iii) and (iv) – using family-wise error (FWE) correction at the cluster-level ($p < 0.05$, with a cluster-defining threshold of $p < 0.001$).

2.8. fMRI association with clinical variables

We hypothesised that key clinical features of the CHR state might be associated with the representation of low-level precision-weighted PEs in brain regions that were (i) activated by these learning signals and (ii) exhibiting aberrantly greater such activation in CHR relative to healthy control participants. To this end, and in line with previous convergent findings indicating a link between prefrontal cortical brain changes and CHR (Benetti et al., 2009; Allen et al., 2012; Cannon et al., 2015), we selected a prefrontal region comprising a cluster of differential fMRI activation to ε_2 (see Section 3. Results) in which to examine fMRI contrast beta-value associations (via Pearson correlation) with SIPS positive symptoms and SIPS-GAF measures, predicting that higher ε_2 -related activation in this region would be associated with greater symptom burden/severity.

3. Results

3.1. Behaviour

Initially, we applied a mixed-factor ANOVA with between-subject and within-subject factors to directly measurable behaviour (accuracy and reaction times, respectively) in order to test for significant main effects or interactions (2 \times 3 factorial design: group \times task phase). No main effects of group and no group \times phase interactions were found. For performance accuracy, we observed a main effect of phase ($df = (2, 48)$, $F = 5.59$, $p = 0.008$ with Greenhouse-Geisser (GG) nonsphericity correction, effect size $\eta^2_{\text{partial}} = 0.19$), with reduced performance in volatile compared to stable phases of the task. By contrast, the main effect of phase for RT closely failed to reach significance ($df = (2, 48)$, $F = 3.11$, $p = 0.07$ with GG correction, $\eta^2_{\text{partial}} = 0.12$). We then proceeded to computational analyses to test for group differences in latent variables underlying learning and decision-making.

3.2. Behavioural modelling

Bayesian model selection gave a clear result, showing that the mean-reverting HGF with a response model incorporating volatility mapping (M_6) was more likely to explain task behaviour than any other model type (Fig. 2B). A summary of parameters from the inversion of this winning model is provided for both groups in Table 1. Importantly, model selection results were equivalent in both groups, allowing for a direct comparison of parameter estimates across groups. Group average belief-updating trajectories computed from the winning model M_6 , are depicted in Figs. S1 and S2.

We investigated group differences in the parameters that (i) were recovered well from simulations and (ii) particularly impacted learning about volatility and thus the precision of high-order beliefs. While we did not observe any group differences for the meta-volatility ϑ parameter ($df = (1, 25)$, $F = 0.54$, $p = 0.47$, $\eta^2_{\text{partial}} = 0.02$), we found a significant difference in the mean-reversion equilibria values m_3 . Relative to controls, CHR individuals exhibited reversion to significantly higher equilibria levels (group: $df = (1, 25)$, $F = 4.29$, $p = 0.049$, $\eta^2_{\text{partial}} = 0.15$; Fig. 4A). This effect remained significant,

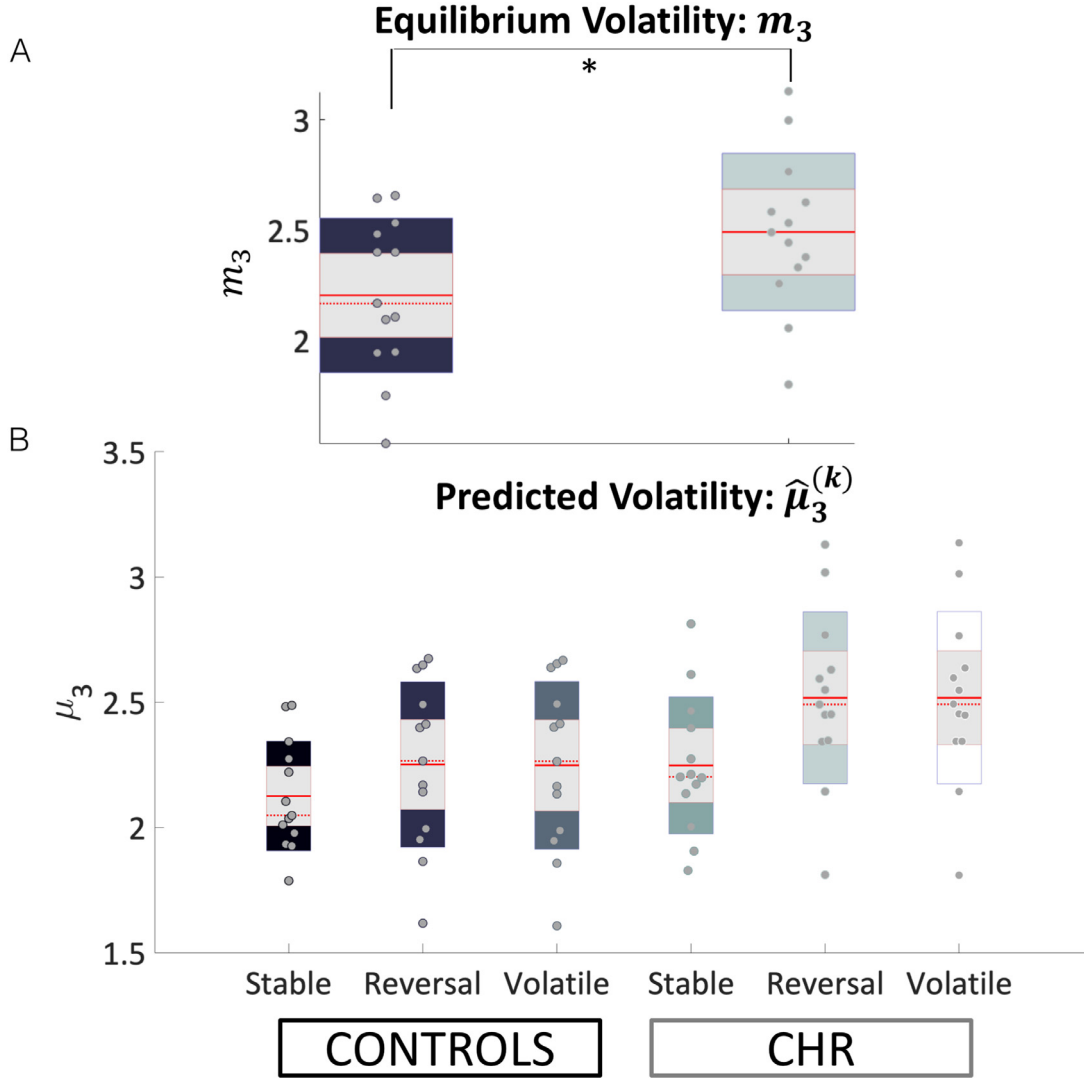


Fig. 4. Behavioural parameter group differences. (A) Group differences in reversion equilibria values (m_3): larger reversion equilibria values were detected in CHR compared to controls (group: $df = (1, 25)$, $F = 4.29$, $p = 0.049$, $\eta^2_{\text{partial}} = 0.15$); and (B) Group-by-phase interactions of perceived environmental volatility ($\hat{\mu}_3$): A mixed-factor ANOVA (with Greenhouse-Geisser nonsphericity correction), which included between-subject and within-subject factors, found a significant main effect of phase and a significant group \times phase interaction (phase: $df = (2, 48)$, $F = 41.68$, $p = 1.113e-06$, $\eta^2_{\text{partial}} = 0.63$; group \times phase: $df = (2, 48)$, $F = 5.71$, $p = 0.025$, $\eta^2_{\text{partial}} = 0.19$). See Section 3.2 in main text for details. No significant main effect of group was found (group: $df = (1, 24)$, $F = 3.42$, $p = 0.08$, $\eta^2_{\text{partial}} = 0.12$). Jittered raw data are plotted for each parameter. The solid red line refers to the mean, the dotted red line to the median, the grey background reflects 1 SD of the mean, and the coloured bars the 95% confidence intervals of the mean. “*” refers to group differences of significance level $p < 0.05$.

whether or not the single participant in the CHR group whose diagnosis failed to be upheld at follow-up (and who was thus excluded from the between-group fMRI analyses) was included. From Figs. S1 and S2, one can observe that the belief trajectories for the estimated volatility extend to higher levels in CHR, relative to controls, consistent with the group differences in the MAP estimates of m_3 .

Second, we investigated group \times task phase interaction effects on the phase-specific averages of predicted environmental volatility, or $\hat{\mu}_3$, following probabilistic contingency reversals, by performing a mixed-factor ANOVA with between-subject and within-subject factors (with GG nonsphericity correction). We found a significant group-by-phase interaction effect ($df = (2, 48)$, $F = 5.71$, $p = 0.025$, $\eta^2_{\text{partial}} = 0.19$; Fig. 4B), suggesting that, in comparison to controls, CHR individuals display a significantly larger increase in perceived environmental volatility following the first probability reversal. We also found a main effect of phase ($df = (2, 48)$, $F = 41.68$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.63$), reflecting significant increases in perceived volatility for both groups with increases in the volatility of the task schedule. Importantly, these

mixed-factor ANOVA results remained significant, whether or not the single participant in the CHR group whose diagnosis failed to be upheld at follow-up was included. In contrast, we found no significant group main effects or interactions for the trajectories of ε_2 and σ_2 (see Figs. S3 and S4). However, we did find a significant group \times task phase interaction for predicted environmental uncertainty (see Fig. S5).

A number of our model parameters were estimated based on participants’ choice behaviour. We also examined whether their actual values could be recovered from synthetic data. Thus, we simulated responses based on all participants’ MAP parameter estimates, and then fitted the model to those simulated responses. The results of the parameter recovery are included in Fig. S6. Several parameters could be recovered well from the data: The second-level parameters encoding tonic (ω) and phasic (κ) aspects of volatility and third-level parameters determining the dynamics of the volatility trajectory, including the equilibrium point m_3 and the meta-volatility parameter ϑ were recovered from the data well. By contrast, the initial value on the volatility $\mu_3^{(k=0)}$, the initial variance on the volatility $\sigma_3^{(k=0)}$, and the inverse

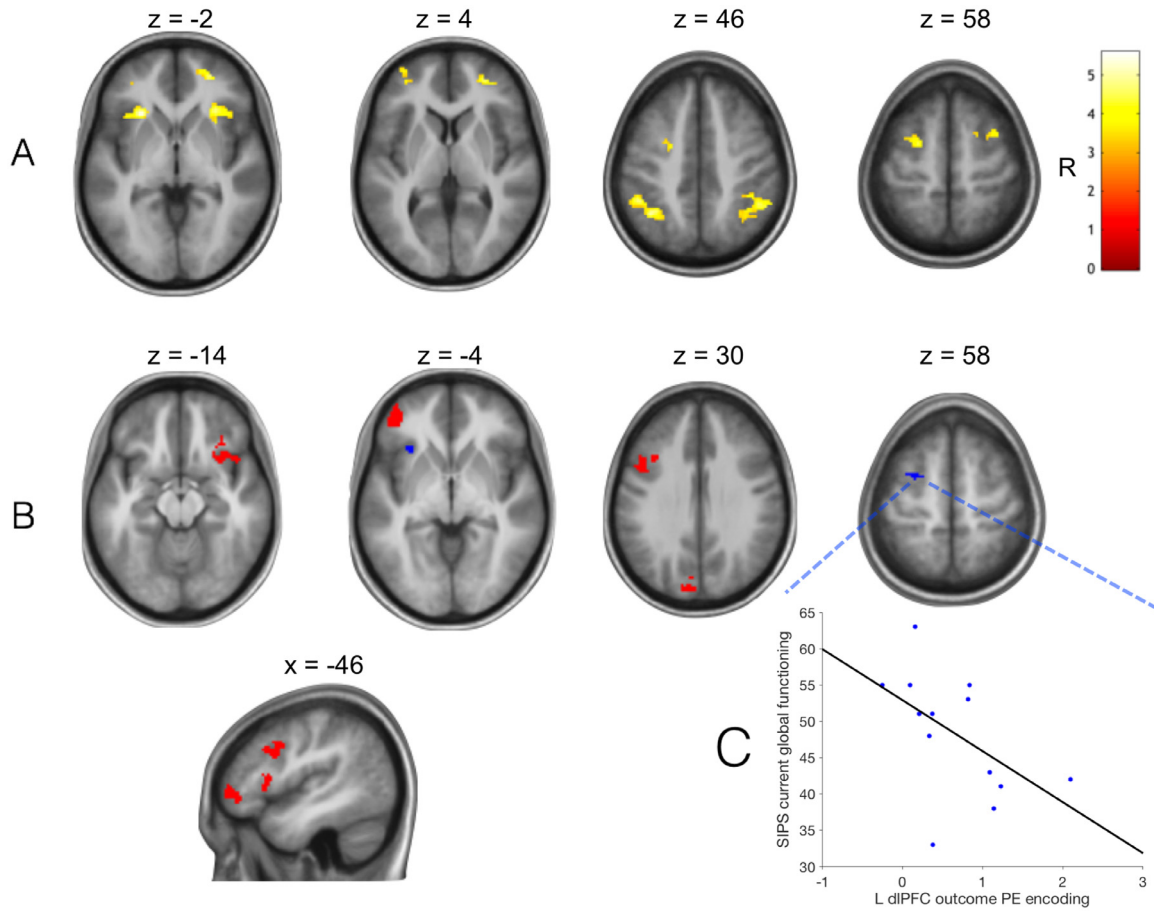


Fig. 5. The neural representation of low-level/outcome-related precision-weighted PEs (ϵ_2) in CHR patients and healthy controls. (A) A representative map of significant (cluster-level FWE-corrected $p < 0.05$) group-level (CHR + controls) outcome-related activations modulated parametrically by ϵ_2 , calculated via one-sample t -test ($n = 25$) and overlaid on an anatomical image calculated as the mean structural MRI of the whole group. (B) Significantly greater representation of ϵ_2 -related activation in a sub-set of these and other regions in CHR patients relative to controls. Solid colour maps of group differences are binarised and indicate spatial differences between whole-brain cluster-level corrected results (red) and results corrected using the group average map in ‘(A)’ for contrast-masking (dark blue). Colour bar represents t -statistics. Axial and coronal slices are orientated in line with neurological conventions ($R = \text{right}$). (C) Significant negative correlation in CHR patients ($n = 13$) between a clinical measure of current global functioning (SIPS-GAF) and ϵ_2 -related activation (beta-values, from the analysis as in panel ‘A’) in a region of left (L) superior dlPFC also showing significant group differences (CHR > controls, from the independent analysis as in panel ‘B’).

decision noise parameter β could not be recovered from the data well.

3.3. fMRI

When pooling across both groups, we found that the computational regressor associated with the precision-weighted outcome PE signal (ϵ_2) activated a set of bilateral cortical regions (Fig. 5A), similar to previous analyses of the same type of PE from a sensory learning task (Iglesias et al., 2013). More specifically, the nine clusters forming this network were located bilaterally in inferior parietal cortex, anterior insula, ventrolateral prefrontal cortex (vlPFC) extending also into right dorsolateral prefrontal cortex (dlPFC) and superior frontal cortex (spanning superior and middle frontal gyri), as well as in a region of left cerebellum ($p < 0.05$, whole-brain cluster-level FWE-corrected, see Section 2. Methods; Table 2). An additional region of interest analysis, using an anatomically defined *a priori* masque of the midbrain (Bunzeck and Düzel, 2006; Iglesias et al., 2013) revealed no significant activation in this region.

In an additional analysis step, the whole-brain corrected activation pattern was used as a functionally defined masque to restrict the subsequent analysis of group differences in ϵ_2 (group-by-PE interactions) to regions showing a main effect of ϵ_2 (note that the two contrasts are orthogonal and thus do not cause non-independence problems for inference). We found significantly greater activation (small volume FWE-

corrected $p < 0.05$) by ϵ_2 in CHR patients, relative to controls, in areas of left superior frontal cortex and left anterior insula (Fig. 5B; Table 2). A distinct set of regions also survived whole-brain correction for this contrast, revealing additional group differences in the left vlPFC (some overlap with the pooled whole group region encoding ϵ_2), precuneus, dlPFC, frontal insulo-opercular cortex and right anterior insula extending into orbitofrontal cortex (OFC; Fig. 5B; Table 2). An additional region of interest analysis focussing on the same *a priori* midbrain masque described above similarly revealed no significant group differences in activation. Although no significant outcome phase activation was found at the whole-group level for 2nd-level uncertainty (σ_2), this variable activated a region of the left fusiform gyrus significantly more in CHR than in controls (whole-brain cluster-level FWE-corrected; $p = 0.046$, peak $t = 5.38$; $x = -28$, $y = -58$, $z = -10$; 74 voxels).

Under whole-brain FWE-correction for multiple comparisons, a number of whole-group average and group difference (controls > CHR) effects were found for the precision-weighted probability PE (ϵ_3), which informs updates of beliefs about volatility (Table 2). At the whole-group level, ϵ_3 significantly activated regions of the bilateral posterior cingulate cortex (PCC) and the left parahippocampal gyrus (Fig. 6A; Table 2). The group contrast ‘ ϵ_3 : controls > CHR’ revealed a more widespread difference in the representation of volatility-related precision-weighted PEs in four additional regions, not overlapping with the whole group average, comprising regions of left vlPFC, left frontopolar

Table 2

Regions showing main effects (average effect across groups) and group differences of ε_2 - (upper panel) and ε_3 -related (lower panel) activations during the outcome phase. Upper panel: a contrast-masked (overlapping) subset of the ε_2 -related regions display significant group differences (CHR > controls; **bold** text). Additional regions displaying significant whole-brain corrected group differences are highlighted in *italics* (non-overlapping) or **bold italics** (some overlap). Secondary and tertiary values denote these various group difference effects. Lower panel: group differences in ε_3 -related activation (controls > CHR) do not overlap with average effects and thus are shown only in *italics* in distinct rows from the main average effects across groups; dlPFC = dorsolateral prefrontal cortex; OFC = orbitofrontal cortex; PCC = posterior cingulate cortex; TPJ = temporoparietal junction; vlPFC = ventrolateral prefrontal cortex; R = right and L = left hemisphere.

| PE-associated fMRI activations | Brain region | Cluster size (2 mm ³ voxels) | p (FWE-corrected) | t-value of peak voxel | x, y, z coordinates (MNI) |
|--|---|---|-----------------------------|--------------------------|-------------------------------------|
| fMRI activations associated with ε_2 | L inferior parietal | 382 | < 0.001 | 5.59 | -38, -54, 42 |
| | L anterior insula/opercular cortex | 140; 30; 90 | 0.002; 0.019 ; 0.019 | 5.30; 3.95 ; 4.47 | -28, 22, -2; -32, 22, 0; -44, 18, 4 |
| | R inferior parietal | 620 | < 0.001 | 5.23 | 40, -46, 38 |
| | L medial lateral cerebellum | 103 | 0.010 | 5.09 | -38, -60, -42 |
| | R vlPFC (and dlPFC) | 175 | < 0.001 | 4.86 | 26, 50, 2 |
| | R anterior insula/OFC | 156; 87 | 0.001; 0.022 | 4.71; 4.08 | 30, 24, -2; 32, 16, -14 |
| | L superior frontal | 156; 30 | 0.001; 0.019 | 4.45; 4.06 | -24, 0, 58; -30, 2, 58 |
| | R superior frontal | 77 | 0.039 | 4.27 | 34, 8, 60 |
| | R vlPFC | 92; 252 | 0.017; < 0.001 | 3.94; 5.10 | -36, 58, 6; -48, 44 , -4 |
| | L precuneus | 136 | 0.002 | 4.58 | -8, -86, 32 |
| | L dlPFC | 171 | < 0.001 | 4.59 | -50, 12, 32 |
| | PCC | 723 | < 0.001 | 5.17 | -2, -56, 14 |
| fMRI activations associated with ε_3 | L parahippocampal gyrus | 103 | 0.010 | 4.07 | -32, -40, -14 |
| | L vlPFC | 326 | < 0.001 | 5.35 | -48, 46, -4 |
| | R sup. temporal gyrus/TPJ | 73 | 0.049 | 5.04 | 54 -20, 4 |
| | L frontopolar cortex | 132 | 0.002 | 4.83 | -18, 50, 18 |
| | R OFC/anterior insula | 108 | 0.008 | 4.43 | 32, 20, -16 |
| | | | | | |

cortex, right superior temporal gyrus/temporoparietal junction and right OFC extending into anterior insula (Fig. 6B; Table 2). Furthermore, within our *a priori* anterior cingulate cortex masque, the subgenual cingulate gyrus (Fig. 6B) displayed significantly greater activation to ε_3 in controls than in CHR patients (small volume FWE-corrected, $p = 0.031$, peak $t = 5.08$; $x = -2$, $y = 16$, $z = -24$; 40 voxels). In addition, an exploratory region of interest analysis using the aforementioned midbrain masque revealed a trend-level cluster (small volume FWE-corrected) of group differences in encoding (ε_3 : controls > CHR) in the left substantia nigra/ventral tegmental area (SN/VTA) of the midbrain ($p = 0.087$, peak $t = 3.42$; $x = -2$, $y = -22$, $z = -20$;

4 voxels; Fig. 6B).

Finally, when pooling across both groups we found that the predicted volatility $\hat{\mu}_3^{(k)}$ during the decision phase of the probabilistic reversal learning task was associated with BOLD activation in a number of regions including bilateral cuneus, precuneus, superior frontal cortex, right superior temporal and right precentral gyrus/central sulcus (Fig. 7A). Contrast masking using this whole group result to examine for group differences revealed a cluster in posterior superior temporal cortex displaying significantly greater volatility-related activation in controls relative to CHR (Fig. 7B; Table 3). Two additional regions from a whole-brain corrected group difference contrast, comprising right

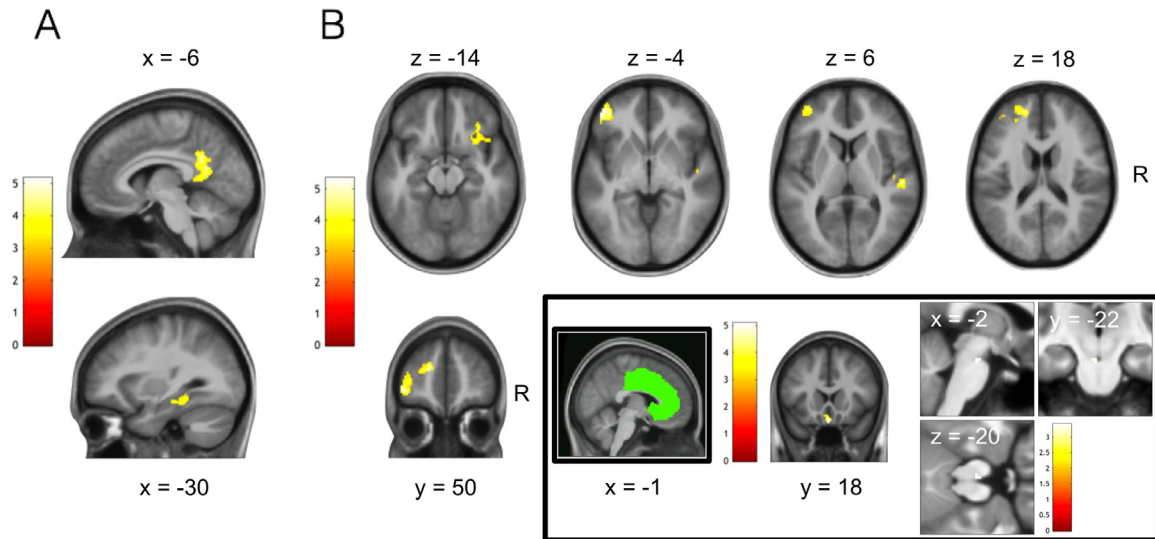


Fig. 6. Failures of monitoring and incorporating environmental uncertainty (volatility) in probabilistic learning by CHR relative to control individuals: prediction error response. (A) A representative map of significant (cluster-level FWE-corrected $p < 0.05$) group-level (CHR + controls) outcome-related activations modulated parametrically by precision-weighted volatility-related prediction error (ε_3), calculated via one-sample t -test ($n = 25$) and overlaid on an anatomical image calculated as the mean structural MRI of the whole group. (B) Greater neural representation of high-level/volatility-related precision-weighted PEs (ε_3) during decision feedback in healthy controls relative to CHR patients, (main) in a network of regions identified using whole-brain cluster-level correction, (inset) in a region of subgenual anterior cingulate identified under correction for an *a priori* anterior cingulate masque (bottom centre) and in a left midbrain region identified under correction for a dopaminergic midbrain masque (trend level $p = 0.087$, bottom right). Double inset, green: representative sagittal slice depicting anatomical anterior cingulate cortex masque used as search volume in statistical analysis and multiple comparison correction. Colour bars represent t -statistics.

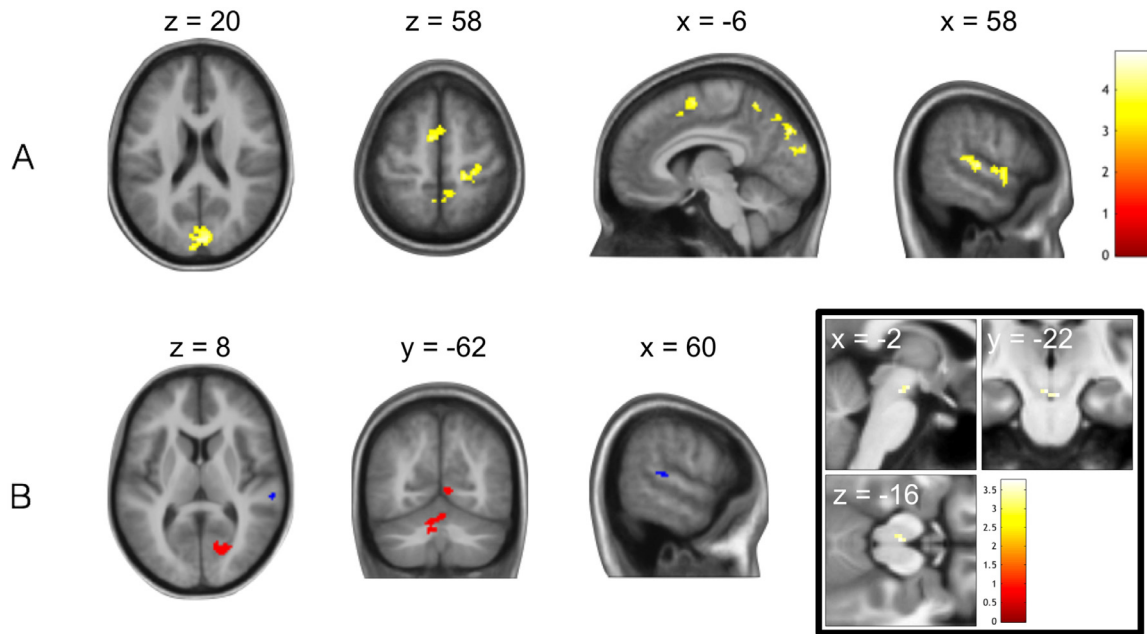


Fig. 7. Failures of monitoring and incorporating environmental uncertainty (volatility) in probabilistic learning by CHR relative to control individuals: decision tracking. (A) A representative map of significant (cluster-level FWE-corrected $p < 0.05$) group-level (CHR + controls) decision-related activations modulated parametrically by estimated volatility, calculated via one-sample t -test ($n = 25$) and overlaid on an anatomical image calculated as the mean structural MRI of the whole group. (B) Greater neural representation of estimated volatility during probabilistic decision-making in healthy controls relative to CHR patients, (main) in a network where solid colour maps of group differences are binarised and indicate spatial differences between whole-brain cluster-level corrected results (red) and results corrected using the group average map in 'A' for contrast-masking (dark blue), and (inset) in a midbrain region identified under correction for a dopaminergic midbrain masque ($p < 0.05$). Colour bars represent t -statistics.

visual cortex extending into precuneus and a left medial cerebellar region, also showed significantly greater volatility-related modulation of decision-phase activation in CHR than in control individuals (Fig. 7B, Table 3). Finally, within our *a priori* midbrain masque we found one region of SN/VTA (small volume FWE-corrected $p = 0.046$, peak $t = 3.97$; $x = 4$, $y = -22$, $z = -18$; 9 voxels; Fig. 7B) that showed significantly greater encoding of predicted volatility in control participants relative to CHR.

3.4. fMRI association with clinical variables

Following our hypothesis that key clinical features of the CHR state might be associated with the representation of low-level precision-weighted PEs in prefrontal regions, we extracted subject-level fMRI analysis beta-values of ε_2 -related activation from the cluster of left superior dlPFC that also displayed a significant group difference in this ε_2 encoding representation (CHR > controls). We then performed Pearson correlation analysis of these beta-values with sub-scales of the SIPS. We found a significant negative correlation between these fMRI data and the SIPS-GAF measure of current global functioning

($r = -0.53$, one-tailed $p < 0.05$; Fig. 5C), as well as a significant *positive* correlation between the fMRI data and the degree of decline in these SIPS-GAF scores within the past year ($r = 0.60$, one-tailed $p < 0.05$; see Fig. S7). In other words, the higher the low-level outcome-related PE encoding in dlPFC, the lower the current functioning (and the greater the recent decline in functioning) and thus the greater the symptom burden or severity at the time of testing.

4. Discussion

The CHR state is characterised by attenuated or brief, self-limiting psychotic symptoms, including delusional beliefs. While delusions are thought to reflect the endpoints of aberrant learning and inference processes, with evidence for links to dopaminergic mechanisms in full-blown schizophrenia (Murray et al., 2008; Romaniuk et al., 2010; Gradin et al., 2011), it is not clear whether cognitive disturbances of this sort are already evident in the CHR state.

The specific hypothesis tested in this paper derives from concepts of hierarchical Bayesian inference that play a prominent role in theories of schizophrenia (Stephan et al., 2006, 2009a; Corlett et al., 2009a, 2010;

Table 3

Upper panel: regions showing main effects of estimated volatility-related activation during the decision phase, a subset of which display contrast-masked group differences (**controls > CHR; bold text**); Lower panel, in *italics*: additional regions showing controls > CHR differences, identified using whole-brain correction; R = right and L = left hemisphere.

| Activations assoc. with $\hat{\mu}_3^{(k)}$ | Brain region | Cluster size (voxels) | p (FWE) | t-value | x, y, z |
|---|---|-----------------------|---------------------|-------------------|--------------------------------|
| Decision-phase volatility estimate controls + CHR | Bilateral cuneus | 512 | < 0.001 | 4.90 | 8, -84, 22 |
| | R posterior superior temporal | 131; 18 | 0.002; 0.037 | 4.78; 4.49 | 62, -18, 6; 58, -34, 10 |
| | Bilateral precuneus | 181 | < 0.001 | 4.36 | -2, -54, 52 |
| | R precentral gyrus/central sulcus | 105 | 0.009 | 4.28 | 28, -30, 60 |
| | Bilateral superior frontal | 172 | < 0.001 | 4.16 | -8, 0, 58 |
| | R anterior superior temporal | 181 | < 0.001 | 4.03 | 58, 2, -2 |
| Decision-phase volatility estimate controls > CHR | <i>Right visual/intracalcarine/precuneal cortex</i> | 158 | <i>0.001</i> | 4.76 | 4, -60, 2 |
| | <i>L anterior medial cerebellum</i> | 104 | <i>0.009</i> | 4.75 | -10, -64, -32 |

Fletcher and Frith, 2009; Adams et al., 2013). The central idea here is that the brain instantiates a generative model of its sensory inputs, i.e., a model that makes predictions about the environment and how its (hidden) state gives rise to sensations. Perceptual inference rests on inverting the model to determine the most likely cause of sensory input; learning serves to update beliefs such that future sensory inputs can be better predicted. Importantly, under very general assumptions, these belief updates have a generic form: they are proportional to prediction error, weighted by a precision ratio that serves as a dynamic learning rate and balances the expected precision of low-level (e.g., sensory) input against the precision of prior beliefs (see Eqs. (1) and (2); see also Mathys et al., 2014). The corollary from this general update rule is that unusually pronounced belief updates can arise from two sources: by assigning too much precision to sensory inputs, or by overly uncertain prior beliefs.

This perspective allows for formalising the long-standing concept of “aberrant salience” (Kapur, 2003; Heinz and Schlagenhauf, 2010) and predicts that the initial phase of delusion formation could be equally characterised by abnormally high low-level precision (‘sensory precision’) or by abnormally low precision of higher-level, prior beliefs (‘belief precision’). As indicated by Eqs. (1) and 2, these two factors both work in the same direction by elevating the weights of PEs and thus enhancing belief updates. A chronic persistence of either factor could eventually lead to highly unusual beliefs, or necessitate a compensatory adoption of beliefs. For example, constant surprise about events in the world might eventually only be resolved by the adoption of unusual and rigid higher-order beliefs (see, e.g., Corlett et al., 2009b). Conversely, high-order beliefs of abnormally low precision (Adams et al., 2013) render the environment seemingly unpredictable (e.g., more volatile) and boost the weight of lower-level PEs.

Using a combination of computational modelling of behaviour and fMRI, this study examined these hypothetical abnormalities of learning and inference in non-medicated CHR individuals, compared to control subjects. Firstly, computational modelling of behaviour indicated that, across the duration of the task, CHR individuals converged to significantly higher levels in their beliefs about volatility. Secondly, concerning the fMRI results, we found evidence for both aberrantly enhanced neural encoding of low-level precision-weighted PEs and aberrantly attenuated encoding of high-level precision-weighted PEs in CHR individuals.

When testing for the average effect of low-level outcome-related PEs (ϵ_2) across groups, we found activation in a set of fronto-parieto-insular cortical regions, in line with previous findings (Iglesias et al., 2013; Diaconescu et al., 2017). We found a significant group difference in the degree of this activation within lateral and superior frontal, insular and precuneal cortex, with CHR individuals displaying increased ϵ_2 -related activation compared with controls. The possibility that this increase in ϵ_2 -related activation might be a prominent pathophysiological characteristic of the CHR state is supported by our observation that beta magnitudes of low-level PE encoding in left superior dlPFC were significantly correlated with more burdensome clinical scores of prodromal symptoms (SIPS-GAF). One mechanistic explanation for the observed overweighting (in terms of neural encoding) of low-level PEs could be a reduced precision of higher-level beliefs in CHR individuals; in other words, CHR may be characterised by abnormal estimates of environmental uncertainty.

Accordingly, with respect to activations by higher-level precision-weighted PEs (ϵ_3) that inform updates of volatility estimates, we indeed found opposing effects, in that many of the same regions – dlPFC, vlPFC and insula, as well as additional effects in temporoparietal junction, the subgenual cingulate and, marginally, the dopaminergic midbrain – displayed greater ϵ_3 -related activations in controls, relative to CHR individuals. Taken together, our behavioural and fMRI findings related to higher-order beliefs therefore suggest that CHR individuals may perceive the volatility of the environment differently from controls. The mechanism underlying the observed constellation of an increased ‘set

point’ of estimated volatility (according to the behavioural data) and reduced activation in response to high-level precision-weighted PEs (ϵ_3) that inform updates of volatility estimates, remains to be clarified but suggests a neurobiological difference in the processing of volatility.

Collectively, our results indicate that associative learning under volatility in the CHR state is characterised by higher estimates of environmental volatility (as expressed at the behavioural level) and overly high low-level precision-weighted PE activations (at the neural level). These effects may reflect an enhanced tendency towards belief updating and might explain the empirically observed “jumping to conclusions” bias in CHR individuals (Broome et al., 2007; Winton-Brown et al., 2015). More generally, this cognitive style may represent a risk factor for delusion proneness.

A final finding is more difficult to interpret: in contrast to the behavioural results, which suggest that CHR individuals’ updates of volatility were converging to significantly higher levels than those of the controls, fMRI activations by volatility estimates were generally lower in CHR individuals during decision-making. These relative deactivations by volatility were found in several regions, including temporal and occipital cortices, as well as the midbrain. It is presently not clear how this reduced neural representation of volatility during decision-making is related to the behavioural evidence for increased volatility updating. Generally, however, this atypical cortical representation of volatility does support the general notion that processing of high-level uncertainty is abnormal in CHR individuals.

Our study has some strengths but also several important limitations that deserve mentioning. Regarding strengths, we implemented strict inclusion criteria that limited recruitment to those CHR individuals who had never been exposed to antipsychotic medication. Given the potential relation of low-level prediction errors and precision to dopamine function, this step avoided a potentially critical confound. Additionally, we took several steps to maximise the sensitivity of our analyses, including careful correction for physiological noise based on ICA and RETROICOR, orthogonal contrast masking and *a priori* hypotheses about the encoding of specific computational signals in specific anatomical regions of interest. Concerning limitations, while not markedly lower than previous fMRI studies on the CHR state, which have typically examined up to 18 participants (for example, see Allen et al., 2010; Roiser et al., 2013; Schmidt et al., 2013; Falkenberg et al., 2015b; Modinos et al., 2015; Ermakova et al., 2018), our sample size would nonetheless have to be regarded as rather small. We thus emphasise that our results should be treated with caution until replicated in additional samples. A second issue is that the two groups were not perfectly matched with regard to age and sex. We controlled for these differences by including these variables as covariates of no interest in all group-level fMRI analyses. In addition, although none of the sample reported substance abuse or dependence according to DSM-IV criteria, in future studies it would be prudent to acquire more detailed information regarding participants’ histories of illicit drug use, which may be a determining factor in their development of psychosis symptoms (e.g., Arseneault et al., 2002). Finally, it is also notable that the SIPS clinical questionnaire we incorporated into our analyses is comprised of a number of different sub-scales, and the reported significant association between one of these (SIPS-GAF) and our fMRI measures of low-level precision-weighted PE encoding would not survive Bonferroni correction for testing all scales. It should thus be treated as a preliminary result that requires confirmation by future studies.

In conclusion, our findings contribute to advancing a mechanistic understanding of cognitive abnormalities during the CHR state. Using computational modelling, functional neuroimaging and clinical measures, we found that behaviour and brain activity of individuals at risk for psychosis, relative to healthy control subjects, shows preliminary evidence for two potential mechanisms – increased low-level precision (of outcome PEs) and greater updates in high-level uncertainty (i.e., volatility) – that converge in their impact and render an individual more prone to adjusting high-order beliefs. In addition, we provide

empirical evidence that individual neural representations of outcome-related learning signals (low-level precision-weighted PEs) correlate with individual differences in symptom severity. These findings support previous proposals focusing on the importance of aberrant salience (Kapur, 2003; Roiser et al., 2013) and imprecise higher-order beliefs (Adams et al., 2013) for delusion proneness. The present results may usefully inform future investigations that employ computational and biophysical models to study prediction errors and uncertainty in larger CHR samples and examine the relevance of these quantities for predicting conversion to psychosis in prospective designs.

CRedit authorship contribution statement

David M. Cole: Formal analysis, Writing - original draft, Visualization. **Andreea O. Diaconescu:** Validation, Formal analysis, Writing - original draft, Visualization. **Ulrich J. Pfeiffer:** Methodology, Investigation, Writing - review & editing. **Kay H. Brodersen:** Methodology, Software, Data curation, Writing - review & editing. **Christoph D. Mathys:** Methodology, Software, Writing - review & editing. **Dominika Julkowski:** Investigation, Resources, Writing - review & editing. **Stephan Ruhrmann:** Investigation, Resources, Writing - review & editing, Supervision. **Leonhard Schilbach:** Resources, Writing - review & editing. **Marc Tittgemeyer:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration. **Kai Vogeley:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration. **Klaas E. Stephan:** Conceptualization, Methodology, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge support by the University of Zurich, the UZH Clinical Research Priority Program (CRPP) “Molecular Imaging” (DMC, KES), the René and Susanne Braginsky Foundation (KES) and SNF Ambizione PZ00P3_167952 and Krembil Foundation (AOD).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.nicl.2020.102239](https://doi.org/10.1016/j.nicl.2020.102239).

References

- Adams, R.A., Huys, Q.J.M., Roiser, J.P., 2016. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* 87, 53–63.
- Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The computational anatomy of psychosis. *Front psychiatry* 4, 47.
- Allen, P., Luigjes, J., Howes, O.D., Egerton, A., Hirao, K., Valli, I., Kambeitz, J., Fusar-Poli, P., Broome, M., McGuire, P., 2012. Transition to psychosis associated with prefrontal and subcortical dysfunction in ultra high-risk individuals. *Schizophr. Bull.* 38, 1268–1276.
- Allen, P., Stephan, K.E., Mechelli, A., Day, F., Ward, N., Dalton, J., Williams, S.C., McGuire, P., 2010. Cingulate activity and fronto-temporal connectivity in people with prodromal signs of psychosis. *Neuroimage* 49, 947–955.
- Arseneault, L., Cannon, M., Poulton, R., Murray, R., Caspi, A., Moffitt, T.E., 2002. Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. *BMJ* 325, 1212–1213.
- Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imag.* 23, 137–152.
- Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221.
- Benetti, S., Mechelli, A., Picchioni, M., Broome, M., Williams, S., McGuire, P., 2009. Functional integration between the posterior hippocampus and prefrontal cortex is impaired in both first episode schizophrenia and the at risk mental state. *Brain* 132,

- 2426–2436.
- Broome, M.R., Johns, L.C., Valli, I., Woolley, J.B., Tabraham, P., Brett, C., Valmaggia, L., Peters, E., Garety, P.A., McGuire, P.K., 2007. Delusion formation and reasoning biases in those at clinical high risk for psychosis. *Br. J. Psychiatry Suppl.* 51, s38–s42.
- Bunzeck, N., Düzel, E., 2006. Absolute coding of stimulus novelty in the human substantia nigra/ventral striatum. *Neuron* 51, 369–379.
- Cannon, T.D., Chung, Y., He, G., Sun, D., Jacobson, A., van Erp, T.G.M., McEwen, S., Addington, J., Bearden, C.E., Cadenhead, K., Cornblatt, B., Mathalon, D.H., McGlashan, T., Perkins, D., Jeffries, C., Seidman, L.J., Tsuang, M., Walker, E., Woods, S.W., Heinssen, R., 2015. Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk. *Biol Psychiatry* 77, 147–157.
- Corlett, P.R., Frith, C.D., Fletcher, P.C., 2009a. From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology (Berl)* 206, 515–530.
- Corlett, P.R., Krystal, J.H., Taylor, J.R., Fletcher, P.C., 2009b. Why do delusions persist? *Front. Hum. Neurosci.* 3, 12.
- Corlett, P.R., Murray, G.K., Honey, G.D., Aitken, M.R.F., Shanks, D.R., Robbins, T.W., Bullmore, E.T., Dickinson, A., Fletcher, P.C., 2007. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain* 130, 2387–2400.
- Corlett, P.R., Taylor, J.R., Wang, X.-J., Fletcher, P.C., Krystal, J.H., 2010. Toward a neurobiology of delusions. *Prog. Neurobiol.* 92, 345–369.
- Cuthbert, B.N., Insel, T.R., 2013. Toward the future of psychiatric diagnosis: the seven pillars of rdcc. *BMC Med* 11, 126.
- Dauvizeau, J., den Ouden, H.E.M., Pessiglione, M., Kiebel, S.J., Stephan, K.E., Friston, K.J., 2010. Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS ONE* 5, e15554.
- Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S., 1995. The Helmholtz machine. *Neural Comput.* 7, 889–904.
- Diaconescu, A.O., Hauke, D.J., Borgwardt, S., 2019. Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Mol. Psychiatry* 24, 1258–1267.
- Diaconescu, A.O., Jensen, J., Wang, H., Willeit, M., Menon, M., Kapur, S., McIntosh, A.R., 2011. Aberrant effective connectivity in schizophrenia patients during appetitive conditioning. *Front. Hum. Neurosci.* 4, 239.
- Diaconescu, A.O., Mathys, C., Weber, L.A.E., Dauvizeau, J., Kasper, L., Lomakina, E.I., Fehr, E., Stephan, K.E., 2014. Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput. Biol.* 10, e1003810.
- Diaconescu, A.O., Mathys, C., Weber, L.A.E., Kasper, L., Mauer, J., Stephan, K.E., 2017. Hierarchical prediction errors in midbrain and septum during social learning. *Soc. Cogn. Affect. Neurosci.* 12, 618–634.
- Ermakova, A.O., Knolle, F., Justicia, A., Bullmore, E.T., Jones, P.B., Robbins, T.W., Fletcher, P.C., Murray, G.K., 2018. Abnormal reward prediction-error signalling in antipsychotic naive individuals with first-episode psychosis or clinical risk for psychosis. *Neuropsychopharmacology* 43, 1691–1699.
- Falkenberg, I., Valmaggia, L., Byrnes, M., Frascarelli, M., Jones, C., Rocchetti, M., Straube, B., Badger, S., McGuire, P., Fusar-Poli, P., 2015a. Why are help-seeking subjects at ultra-high risk for psychosis help-seeking? *Psychiatry Res.* 228, 808–815.
- Falkenberg, I., Chaddock, C., Murray, R.M., McDonald, C., Modinos, G., Bramon, E., Walshe, M., Broome, M., McGuire, P., Allen, P., 2015b. Failure to deactivate medial prefrontal cortex in people at high risk for psychosis. *Eur. Psychiatry* 30, 633–640.
- Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 10, 48–58.
- Frässle, S., Yao, Y., Schöbi, D., Aponte, E.A., Heinze, J., Stephan, K.E., 2018. Generative models for clinical applications in computational psychiatry. *Wiley Interdiscip. Rev. Cogn. Sci.* 9, e1460.
- Friston, K., 2005. A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836.
- Friston, K., 2008. Hierarchical models in the brain. *PLoS Comput. Biol.* 4, e1000211.
- Fusar-Poli, P., et al., 2013a. The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* 70, 107–120.
- Fusar-Poli, P., et al., 2015. Heterogeneity of psychosis risk within individuals at clinical high risk: a meta-analytical stratification. *JAMA Psychiatry* 73, 1–8.
- Fusar-Poli, P., Bechdolf, A., Taylor, M.J., Bonoldi, I., Carpenter, W.T., Yung, A.R., McGuire, P., 2013b. At risk for schizophrenic or affective psychoses? A meta-analysis of DSM/ICD diagnostic outcomes in individuals at high clinical risk. *Schizophr. Bull.* 39, 923–932.
- Gardner, M.P.H., Schoenbaum, G., Gershman, S.J., 2018. Rethinking dopamine as generalized prediction error. *Proc. R. Soc. B Biol. Sci.* 285, 20181645.
- Glover, G.H., Li, T.Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn. Reson. Med.* 44, 162–167.
- Grace, A.A., 1993. Cortical regulation of subcortical dopamine systems and its possible relevance to schizophrenia. *J. Neural. Transm.* 91, 111–134.
- Gradin, V.B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., Steele, J.D., 2011. Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* 134, 1751–1764.
- Gu, Q., 2002. Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience* 111, 815–835.
- Hauser, T.U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., Brem, S., 2014. Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry* 71, 1165–1173.
- Heinz, A., Schlagenhauf, F., 2010. Dopaminergic dysfunction in schizophrenia: salience attribution revisited. *Schizophr. Bull.* 36, 472–485.
- Howes, O.D., Kapur, S., 2009. The dopamine hypothesis of schizophrenia: version III—the first common pathway. *Schizophr. Bull.* 35, 549–562.
- Iglesias, S., Mathys, C., Brodersen, K.H., Kasper, L., Piccirelli, M., den Ouden, H.E.M.,

- Stephan, K.E., 2013. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* 80, 519–530.
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* 167, 748–751.
- Insel, T.R., 2010. Rethinking schizophrenia. *Nature* 468, 187–193.
- Jensen, J., Willeit, M., Zipursky, R.B., Savina, I., Smith, A.J., Menon, M., Crawley, A.P., Kapur, S., 2008. The formation of abnormal associations in schizophrenia: neural and behavioral evidence. *Neuropsychopharmacology* 33, 473–479.
- Kahn, R.S., Sommer, I.E., 2015. The neurobiology and treatment of first-episode schizophrenia. *Mol. Psychiatry* 20, 84–97.
- Kapur, S., 2003. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am. J. Psychiatry* 160, 13–23.
- Kapur, S., Phillips, A.G., Insel, T.R., 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* 17, 1174–1179.
- Kasper, L., Bollmann, S., Diaconescu, A.O., Hutton, C., Heinze, J., Iglesias, S., Hauser, T.U., Sebold, M., Manjaly, Z.-M., Pruessmann, K.P., Stephan, K.E., 2017. The physio toolbox for modeling physiological noise in fMRI data. *J. Neurosci. Method.* 276, 56–72.
- King, R., Barchas, J.D., Huberman, B.A., 1984. Chaotic behavior in dopamine neurodynamics. *Proc. Natl. Acad. Sci. USA* 81, 1244–1247.
- Klosterkötter, J., Schultze-Lutter, F., Bechdolf, A., Ruhrmann, S., 2011. Prediction and prevention of schizophrenia: what has been achieved and where to go next? *World Psychiatry* 10, 165.
- Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E.M., Smieskova, R., Studerus, E., Kambeitz-Illankovic, L., von Saldern, S., Cabral, C., Reiser, M., Falkai, P., Borgwardt, S., 2015. Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophr. Bull.* 41, 471–482.
- Laruelle, M., Abi-Dargham, A., van Dyck, C.H., Gil, R., D'Souza, C.D., Erdoz, J., McCance, E., Rosenblatt, W., Fingado, C., Zoghbi, S.S., Baldwin, R.M., Seibyl, J.P., Krystal, J.H., Charney, D.S., Innis, R.B., 1996. Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proc. Natl. Acad. Sci.* 93, 9235–9240.
- Lawson, R.P., Mathys, C., Rees, G., 2017. Adults with autism overestimate the volatility of the sensory environment. *Nat. Neurosci.* 20, 1293–1299.
- Mathys, C., Daunizeau, J., Friston, K.J., Stephan, K.E., 2011. A bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* 5, 39.
- Mathys, C.D., 2013. **Hierarchical Gaussian filtering (unpublished doctoral thesis). Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.** <https://doi.org/10.3929/ethz-a-007595146>.
- Mathys, C.D., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., Stephan, K.E., 2014. Uncertainty in perception and the Hierarchical Gaussian Filter. *Front. Hum. Neurosci.* 8, 825.
- McGuire, P., Sato, J.R., Mechelli, A., Jackowski, A., Bressan, R.A., Zugman, A., 2015. Can neuroimaging be used to predict the onset of psychosis? *Lancet Psychiatry* 2, 1117–1122.
- Miller, T.J., McGlashan, T.H., Rosen, J.L., Cadenhead, K., Ventura, J., McFarlane, W., Perkins, D.O., Pearson, G.D., Woods, S.W., 2003. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophr. Bull.* 29, 703–715.
- Miller, T.J., McGlashan, T.H., Rosen, J.L., Somjee, L., Markovich, P.J., Stein, K., Woods, S.W., 2002. Prospective diagnosis of the initial prodrome for schizophrenia based on the structured interview for prodromal syndromes: preliminary evidence of interrater reliability and predictive validity. *Am. J. Psychiatry* 159, 863–865.
- Modinos, G., Tseng, H.-H., Falkenberg, I., Samson, C., McGuire, P., Allen, P., 2015. Neural correlates of aberrant emotional salience predict psychotic symptoms and global functioning in high-risk and first-episode psychosis. *Soc. Cogn. Affect. Neurosci.* 10, 1429–1436.
- Montague, P.R., Dolan, R.J., Friston, K.J., Dayan, P., 2012. Computational psychiatry. *Trend. Cogn. Sci.* 16, 72–80.
- Montague, P.R., Hyman, S.E., Cohen, J.D., 2004. Computational roles for dopamine in behavioural control. *Nature* 431, 760–767.
- Moorhead, T.W.J., Stanfield, A.C., McKechnie, A.G., Dauvermann, M.R., Johnstone, E.C., Lawrie, S.M., Cunningham Owens, D.G., 2013. Longitudinal gray matter change in young people who are at enhanced risk of schizophrenia due to intellectual impairment. *Biol. Psychiatry* 73, 985–992.
- Murray, G.K., Corlett, P.R., Clark, L., Pessiglione, M., Blackwell, A.D., Honey, G., Jones, P.B., Bullmore, E.T., Robbins, T.W., Fletcher, P.C., 2008. Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Mol. Psychiatry* 13 (239), 267–276.
- Nieman, D.H., McGorry, P.D., 2015. Detection and treatment of at-risk mental state for developing a first psychosis: making up the balance. *Lancet Psychiatry* 2, 825–834.
- Owen, M.J., 2014. New approaches to psychiatric diagnostic classification. *Neuron* 84, 564–571.
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., Frith, C.D., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442, 1042–1045.
- Powers, A.R., Mathys, C., Corlett, P.R., 2017. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science* 357, 596–600.
- Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87.
- Rescorla, R., Wagner, A., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Class. Cond. Curr. Res. Theory.*
- Roiser, J.P., Howes, O.D., Chaddock, C.A., Joyce, E.M., McGuire, P., 2013. Neural and behavioral correlates of aberrant salience in individuals at risk for psychosis. *Schizophr. Bull.* 39, 1328–1336.
- Roiser, J.P., Stephan, K.E., den Ouden, H.E.M., Barnes, T.R.E., Friston, K.J., Joyce, E.M., 2009. Do patients with schizophrenia exhibit aberrant salience? *Psychol. Med.* 39, 199–209.
- Romaniuk, L., Honey, G.D., King, J.R.L., Whalley, H.C., McIntosh, A.M., Levita, L., Hughes, M., Johnstone, E.C., Day, M., Lawrie, S.M., Hall, J., 2010. Midbrain activation during Pavlovian conditioning and delusional symptoms in schizophrenia. *Arch. Gen. Psychiatry* 67, 1246–1254.
- Ruhrmann, S., Schultze-Lutter, F., Salokangas, R.K.R., Heinimaa, M., Linszen, D., Dingemans, P., Birchwood, M., Patterson, P., Juckel, G., Heinz, A., Morrison, A., Lewis, S., von Reventlow, H.G., Klosterkötter, J., 2010. Prediction of psychosis in adolescents and young adults at high risk: results from the prospective European prediction of psychosis study. *Arch. Gen. Psychiatry* 67, 241–251.
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468.
- Schmidt, A., Smieskova, R., Aston, J., Simon, A., Allen, P., Fusar-Poli, P., McGuire, P.K., Riecher-Rössler, A., Stephan, K.E., Borgwardt, S., 2013. Brain connectivity abnormalities predating the onset of psychosis: correlation with the effect of medication. *JAMA Psychiatry* 70, 903–912.
- Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. *Science* 275, 1593–1599.
- Schwartenbeck, P., FitzGerald, T.H.B., Mathys, C., Dolan, R., Friston, K., 2015. The dopamineergic midbrain encodes the expected certainty about desired outcomes. *Cereb. Cortex* 25, 3434–3445.
- Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niaz, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23 (Suppl 1), S208–S219.
- Stephan, K.E., Bach, D.R., Fletcher, P.C., Flint, J., Frank, M.J., Friston, K.J., Heinz, A., Huys, Q.J.M., Owen, M.J., Binder, E.B., Dayan, P., Johnstone, E.C., Meyer-Lindenberg, A., Montague, P.R., Snyder, U., Wang, X.-J., Breakspear, M., 2016. Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry* 3, 77–83.
- Stephan, K.E., Baldeweg, T., Friston, K.J., 2006. Synaptic plasticity and dysconnection in schizophrenia. *Biol. Psychiatry* 59, 929–939.
- Stephan, K.E., Friston, K.J., Frith, C.D., 2009a. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr. Bull.* 35, 509–527.
- Stephan, K.E., Iglesias, S., Heinze, J., Diaconescu, A.O., 2015. Translational perspectives for computational neuroimaging. *Neuron* 87, 716–732.
- Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009b. Bayesian model selection for group studies. *Neuroimage* 46, 1004–1017.
- Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., Corlett, P.R., 2018. The predictive coding account of psychosis. *Biol. Psychiatry* 84, 634–643.
- Suarez, J.A., Howard, J.D., Schoenbaum, G., Kahnt, T., 2019. Sensory prediction errors in the human midbrain signal identity violations independent of perceptual distance. *Elife* 8.
- Tuang, M.T., Van Os, J., Tandon, R., Barch, D.M., Bustillo, J., Gaebel, W., Gur, R.E., Heckers, S., Malaspina, D., Owen, M.J., Schultz, S., Carpenter, W., 2013. Attenuated psychosis syndrome in DSM-5. *Schizophr. Res.* 150, 31–35.
- van Os, J., Tamminga, C., 2007. Deconstructing psychosis. *Schizophr. Bull.* 33, 861–862.
- Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K.J., Stephan, K.E., 2014. Spatial attention, precision, and bayesian inference: a study of saccadic response speed. *Cereb. Cortex* 24, 1436–1450.
- Wang, X.-J., Krystal, J.H., 2014. Computational psychiatry. *Neuron* 84, 638–654.
- Winton-Brown, T.T., Broome, M.R., Allen, P., Valli, I., Howes, O., Garety, P.A., Johns, L.C., McGuire, P., 2015. Misattributing speech and jumping to conclusions: a longitudinal study in people at high risk of psychosis. *Eur. Psychiatry* 30, 32–37.
- Winton-Brown, T.T., Fusar-Poli, P., Ungless, M.A., Howes, O.D., 2014. Dopaminergic basis of salience dysregulation in psychosis. *Trend. Neurosci.* 37, 85–94.
- Wood, S.J., Yung, A.R., McGorry, P.D., Pantelis, C., 2011. Neuroimaging and treatment evidence for clinical staging in psychotic disorders: from the at-risk mental state to chronic schizophrenia. *Biol. Psychiatry* 70, 619–625.
- Yung, A.R., Yuen, H.P., McGorry, P.D., Phillips, L.J., Kelly, D., Dell'Olio, M., Francey, S.M., Cosgrave, E.M., Killackey, E., Stanford, C., Godfrey, K., Buckley, J., 2005. Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. *Aust. N. Z. J. Psychiatry* 39, 964–971.