Hierarchical Bayesian Inference in Psychosis

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Daniel Jonas Hauke

Mai 2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Auftrag von

Prof. Dr. Volker Roth, Erstbetreuer, Dissertationsleiter
Prof. Dr. Andreea Diaconescu, Erstbeutreuerin
Prof. Dr. Julia Vogt, Zweitbetreuerin
Prof. Dr. Philipp Sterzer, Externer Gutachter

Basel, 21. Juni 2022

Prof. Dr. Marcel Mayor (Dekan)

Thesis advisors: Roth / Diaconescu Daniel Jonas Hauke

Hierarchical Bayesian Inference in Psychosis

Abstract

Schizophrenia is a severe mental illness that affects millions of people worldwide and can have a drastic impact on a patient's life. The illness is characterised by symptoms such as hallucinations and delusions. In recent years, a powerful theoretical framework has been developed to understand better how such symptoms emerge, the predictive coding account of psychosis. In this thesis, I cast different symptoms of psychosis as instances of hierarchical Bayesian inference in a series of studies. The first study examined the question of how persecutory delusions emerge in early psychosis. We derived hypotheses based on previous literature and simulations and tested them empirically in a sample of 18 first-episode psychosis patients, 19 individuals at clinical high risk for psychosis (CHR) and 19 matched healthy controls (HC). Our results suggest that emerging psychosis may be accompanied by an altered perception of environmental volatility. In a second study, this modelling approach was applied to delusions more broadly in a large dataset including 261 patients with psychotic disorders and 56 HC to examine the relationship between delusions and reasoning biases that were previously reported in psychosis. The results of this study suggest that beliefs of patients with psychotic disorders were characterised by increased belief instability, which explained increased belief updating in light of disconfirmatory evidence. We also assessed the clinical utility of this approach by testing its ability to predict treatment response to a psychotherapeutic intervention and found that the parameters of the computational model were able to predict treatment outcome in individual patients. Lastly, in a final study, we modelled brain activity during an implicit sensory learning task in a third independent sample of 38 CHR, 18 early-illness schizophrenia patients, and 44 HC to assess the biological plausibility of this approach. Our results suggest that hierarchical precision-weighted prediction errors derived from the model modulate electroencephalography (EEG) amplitudes. Moreover, we found not only differences in the expression of precision-weighted prediction errors between schizophrenia patients and HC, but also between CHR, who later converted to a psychotic disorder, and non-converters. Jointly, this work demonstrates that this computational approach may not only be conceptually useful to understand the computational mechanisms underlying psychosis, but also clinically relevant and biologically plausible.

Contents

0 Introduction 1
- 0.1 The Dopamine Hypothesis of Schizophrenia . . . 2
- 0.2 The Glutamate Hypothesis of Schizophrenia . . . 3
- 0.3 The Dysconnectivity Hypothesis . . . 3
- 0.4 The Predictive Coding Account of Psychosis . . . 4
- 0.5 The Hierarchical Gaussian Filter . . . 6
- 0.6 Outlook . . . 8

1 Modelling Paranoid Delusions 11
- 1.1 Introduction . . . 11
- 1.2 Methods . . . 22
- 1.3 Results . . . 30
- 1.4 Discussion . . . 35

2 Modelling Reasoning Biases 39
- 2.1 Introduction . . . 39
- 2.2 Methods . . . 41
- 2.3 Results . . . 49
- 2.4 Discussion . . . 56

3 Modelling Sensory Learning 63
- 3.1 Introduction . . . 63
- 3.2 Methods . . . 65
- 3.3 Results . . . 69
- 3.4 Discussion . . . 71

4 Conclusions 77
- 4.1 Theoretical implications . . . 77
- 4.2 Clinical implications . . . 80
- 4.3 Limitations and future directions . . . 81
- 4.4 Final remarks . . . 83

Acronyms 85

References 106

List of Figures

1.1 Generative models of behaviour and neuroimaging data 13
1.2 Probing persecutory ideation: Inferring on others' intentions experimental paradigm and computational model 16
1.3 Functional anatomy of social inference 19
1.4 Model predictions: Beliefs and neural responses 21
1.5 Social learning task and volatility schedule 24
1.6 Simulating an altered perception of environmental volatility 27
1.7 Model space 28
1.8 Behavioural results and parameter group effects 32
1.9 Bayesian model selection results 33
1.10 Model and parameter recovery analyses 34

2.1 Fish task 43
2.2 Winning model 46
2.3 Behavioural effects versus model predictions for RQ1: Psychosis 50
2.4 Behavioural effects versus model predictions for RQ2: JTC 51
2.5 Behavioural effects versus model predictions for RQ3: Delusions 53
2.6 Behavioural effects versus model predictions for IQ-matched subsamples . . 55
2.7 Bayesian model selection and recovery analyses 57
2.8 Parameter group effects and treatment response prediction 59
2.9 Parameter effects for IQ-matched subsamples 60

3.1 Mismatch negativity 64
3.2 Computational analysis pipeline 69
3.3 Expression of low-level precision-weighted PEs in early psychosis 72
3.4 Expression of high-level precision-weighted PEs in early psychosis 75

List of Tables

1.1 Priors on free model parameters 27
1.2 Demographic and clinical characteristics 31

2.1 Priors on free model parameters 47
2.2 Sociodemographic and clinical characteristics 52
2.3 Overview of parameter effects 54

3.1 Summary of posterior parameter estimates 68
3.2 Demographic and clinical characteristics 70

I dedicate this work to patients with psychotic disorders and their families. During my time conducting this research, I have had the honour to meet many outstanding individuals. I am endlessly grateful for the wonderful conversations we shared and for their willingness to participate in our studies. I thank them for volunteering their energy and valuable time despite all the challenges they faced allowing us to pursue this research. I hope that this work will enable myself and others to provide better answers to their questions and ultimately make a difference in their lives.

Acknowledgments

First of all, I would like to thank my supervisors Andreea and Volker for their unending support during these past years. Andreea has not only been an academic mentor to me even before I started my PhD, but a mentor for life. I thank her for her trust in me, for letting me make mistakes without judging me for them, for making sure that I do not knock myself unconscious while doing backflips into a lake during a lab retreat, for all the wonderful and inspiring conversations, for always being available when I had academic or personal problems, for valuing my opinion and always being open to my ideas – no matter, how absurd they were – and for helping me to find my own path in the jungle of academia.

I would also like to thank Volker for being so open-minded when I suddenly came knocking at his door and asked him whether he would be interested in co-supervising me as a PhD student. I thank him for always taking the time to meet me and discuss my science when I needed him, for challenging me by asking detailed questions, for giving me a new perspective on my ideas, for helping me to develop mathematical skills that will be invaluable for my future career and for always supporting me throughout all the challenges I faced over the past years, including convincing my new landlord in Toronto that I am a clean tenant and that he should definitely rent his apartment to me.

Furthermore, I would like to thank Klaas for providing invaluable feedback on the work presented in this thesis and for answering countless questions I had over the years.

Moreover, I would like to thank my girlfriend Lea for always being there for me throughout this challenging period of my life. I thank her for her infinite love, for standing by me for all those years, despite many challenges including moving to three different countries during my PhD, and for listening to all my scattered thoughts about this paper or that experiment – which for some reason always seemed to come out of nowhere. I also thank her for her emotional support in difficult periods and for helping me throughout the roller-coaster ride of these past years.

Last but not least, I would like to thank my family. I thank them for always welcoming me with open arms, for their amazing food, and many loving conversations. My family always encouraged me to be curious, cared for me, let me go to pursue my dreams, and caught me when I was falling and I am forever grateful for this.

0 Introduction

Schizophrenia is one of the most debilitating illnesses that affects approximately 20 million people worldwide and accounts for more than 13 million years of life lived with disability of the global burden of diseases.[31] It is associated with severe consequences for those afflicted by it, for example diminished levels of functioning – both social and vocational[25] – and drastic reductions in life expectancy, with an average of over 14 years.[107] Characteristic symptoms of the illness are hallucinations and delusions, also referred to as positive symptoms as they go beyond or increase the usual range of experiences. Symptoms that involve a blurring of the borders of shared reality like delusions are also referred to as psychotic symptoms and can occur in other disorders as well, for example delusional disorder or bipolar disorder.[12]

In addition to positive symptoms, patients with schizophrenia also report negative symptoms, such as reduced motivation and diminished emotional expression, and cognitive symptoms, for example deficits in working memory and social cognition. However, how different symptoms of schizophrenia emerge and how they are maintained is still a subject of debate. In what follows, I will provide a brief overview of prominent theories of schizophrenia and psychosis. This overview will outline the dopamine hypothesis of schizophrenia,[50,110,218] the glutamate hypothesis of schizophrenia,[119,160,161,162] the dysconnectivity hypothesis,[87,90,91,221,222] and the predictive coding account of psychosis.[76,227] More accounts on schizophrenia have been proposed (e.g.,[117,172,256]), but these are most pertinent to the work presented in this thesis.

0.1 The Dopamine Hypothesis of Schizophrenia

Antipsychotic drugs were discovered in the middle of the past century.[52] Subsequent studies linked the efficacy of these drugs to their affinity for dopamine receptors.[43,209,210] Based on these findings and the observation that amphetamine, which increases synaptic monoamine levels, can induce psychotic symptoms[143], the first version of the dopamine hypothesis of schizophrenia[218] emerged, positing that schizophrenia was caused by hyperdopaminergia (i.e., dopamine excess).[110,218] Accounting for new results of animal research, post-mortem studies in humans and early neuroimaging work, Davis and colleagues later refined this hypothesis.[50] They proposed that, rather than by a general increase of dopamine availability, schizophrenia is better characterised by hyperdopaminergia in subcortical regions, but hypodopaminergia (i.e., reduced dopamine activity) in prefrontal cortex. Recently, the dopamine hypothesis was further modified based on several strands of new evidence.[110] The first line of research included a plethora of positron emission tomography (PET) and single photon emission computerised tomography (SPECT) studies, which allowed to infer on dopamine metabolism based on in-vivo measurements in humans. This lead to a more precise localisation of subcortical dopamine dysregulation to primarily presynaptic dopamine release in the striatum (see[110] for an overview).

Secondly, based on genetic findings,[5] research on environmental risk factors,[51] and work examining the interaction between the two,[159,247] Howes and colleagues[110] proposed that many different genetic and environmental pathways may converge onto a final common pathway, namely dopamine dysregulation. The authors[110] also suggest that dopamine dysregulation may not be specific to the diagnosis of schizophrenia, but rather to psychosis.

Lastly, this new iteration of the dopamine hypothesis attempted to establish a link between the previously predominantly neurophysiological description and the clinical expression of psychotic symptoms, bridging two different levels of analysis. This link was based on a body of research on physiological substrates of reward learning that suggested that dopamine may mediate incentive salience,[20] or signal the difference between actual and predicted reward (i.e., a reward *prediction error*),[204,205] which led to the aberrant salience framework.[125] Kapur et al.[125] proposed that the dysregulated release of dopamine leads to an assignment of unusually high importance or *aberrant salience* to inconspicuous stimuli, which patients may then experience for example as hallucinations. Delusions, on the other hand, are thought to be a cognitive mechanism that functions to make sense of these experiences of aberrant salience.[125]

In conclusion, the dopamine hypothesis has continuously evolved and provides an increasingly detailed account of psychosis. However, it primarily centres around positive symptoms such as hallucinations and delusions. Another theory – the glutamate hypothesis of schizophrenia[119,160,161,162] – has also evolved in the second half of the last century and focuses more strongly on other symptom domains of schizophrenia, namely cognitive and negative symptoms.

0.2 The Glutamate Hypothesis of Schizophrenia

In the late 1950's, two potent dissociative anaesthetics were synthesised, phencyclidine (PCP) and ketamine.[33] Luby et al.[145] and others observed that these drugs caused effects reminiscent of schizophrenia. For example, anesthetic doses produced positive symptoms such as hallucinations and delusions that sometimes persisted for several days (see[121] for review). Importantly, sub-anaesthetic doses also elicited schizophrenia-like symptoms in healthy controls including formal thought disorder, negative symptoms like emotional withdrawal and motor retardation, as well as cognitive symptoms.[121,145,162] It was later discovered that these drugs exert their effects by blocking *N*-methyl-D-aspartate (NMDA) receptors.[121]

Another line of evidence also suggests that an electrophysiological response to violations of statistical regularities in the environment – the so called *mismatch negativity* (MMN) – is consistently reduced in schizophrenia (see[67] for a recent meta-analysis). MMN reductions can be reproduced by PCP and ketamine administration, both in animal models and in humans.[238] Together, these results suggested that schizophrenia may not only be characterised by dopamine dysregulation, but also by alterations in the glutamatergic neurotransmitter system. In its current form, the glutamate hypothesis of schizophrenia postulates that schizophrenia may be characterised by glutamatergic hyperfunction, which is caused by NMDA receptor hypofunction.[162] This is based on the observation that PCP and ketamine act as NMDA receptor antagonists. Blocking NMDA receptors is thought to reduce firing rates of gamma-aminobutyric acid (GABA)ergic interneurons and – as a result – lead to disinhibition of post-synaptic targets or increased firing in excitatory neurons and excess activation of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors.[108,162]

0.3 The Dysconnectivity Hypothesis

The dysconnectivity hypothesis provides an integrative account of the two previous theories postulating that schizophrenia may be best characterised by abnormal modulation of synaptic plasticity.[87,90,91,221,222] This theory centres around the NMDA receptor and its role in regulating experience-dependent and activity-dependent synaptic plasticity.[221] It further argues that other neuromodulators such as dopamine, but also acetycholine and serotonin, influence NMDA receptor function, which could thus provide a final common pathway.[221] The dysconnectivity hypothesis therefore provides an integrative account that unifies the two preceding theories, but shifts the centre of attention to the interaction between these systems, i.e. abnormal synaptic modulation.[90] NMDA receptors play a critical role at mediating the influence of neuromodulators on glutamatergic neurotransmission as they dynamically regulate synaptic gain.[102]

This proposal is in line with recent genetic evidence suggesting that many genes associated with schizophrenia are either directly or indirectly involved in NMDA receptor function and its interaction with other neuromodulatory systems.[87,221] Moreover, the dysconnectivity hypothesis can also account for the heterogeneity observed in patients with schizophrenia,

as different neuromodulatory systems could be impaired to different extents in individual patients.[221] Importantly, disruptions in NMDA-receptor-neuromodulatory-system interactions can be closely linked to computational theories of brain function in general like the Bayesian brain hypothesis[87,221] and theories of psychosis specifically, for example the predictive coding account of psychosis.[76,227]

0.4 The Predictive Coding Account of Psychosis

Over the past decades, the perspective on basic functions of the brain, such as perception, has changed. Traditionally, perception has been viewed as a purely passive and one-directional process, in which the brain receives sensory information via sensory organs and processes this information in a bottom-up manner such that information flows through thalamic nuclei to early sensory and finally, higher order cortical regions. Today, many neuroscientists do not conceptualise perception as a purely passive and one-directional process anymore. Instead, the brain is thought to actively predict the underlying causes of the sensory inputs it receives, generating predictions about the world and updating its internal model based on incoming sensory information[86,87,88,141,183] – an idea that is also referred to as the Bayesian brain hypothesis[60] and can be traced back to observations by Hermann von Helmholtz.[249]

At the core of this proposal lies Bayes' theorem, which provides a mathematical description of how to optimally integrate new information with previous expectations to determine the causes of sensory signals:

$$\overbrace{p(c|u;m)}^{\text{Posterior}} \equiv \frac{\overbrace{p(u|c;m)}^{\text{Likelihood}}\overbrace{p(c;m)}^{\text{Prior}}}{\underbrace{p(u;m)}_{\text{Model Evidence}}}, \tag{1}$$

where c are the causes of sensations, u is the sensory input and m is the model that generated the data. Bayes theorem proposes a way of inferring the posterior probability $p(c|u;m)$ for the underlying causes under a model m, by multiplying the likelihood $p(u|c;m)$, which specifies the probability that a given cause c generated the data, with a prior distribution over the possible causes $p(c|;m)$. This product is normalised by the marginal likelihood $p(u;m)$ also referred to as model evidence. The Bayesian perspective provides a compelling account for many empirical findings, especially those relating to early sensory processing.[86,88,141,183] However, how could the brain perform Bayesian inference?

One popular proposal is that the brain performs predictive coding (see[220] for a recent review). Predictive coding can be viewed as a theory at Marr's computational level of analysis.[151] Marr[151] proposed three levels of analysis to understand complex systems like the brain: (1) the *computational level* specifying a problem that the systems needs to solve, (2) the *algorithmic level* describing how this problem can be solved algorithmically, and (3) the *implementational*

level characterising the physical substrate and its organisation which performs a specific computation (e.g., neuronal populations, neurotransmitters).

At its core, predictive coding posits that the causes of sensations can be inferred by minimising the error between the observed and predicted sensory inputs, referred to as *prediction errors* (PEs). Predictive coding can be motivated from many different perspectives ranging from signal processing to neurophysiology.[183,220] From a signal processing perspective transmitting these PEs is efficient, because they tend to have a smaller dynamic range and thus require less bandwidth or can be transmitted with higher accuracy using the same bandwidth.[220] Predictive coding is also in accordance with a range of physiological findings, including the counterstream architecture of the cortex, comprised of asymmetric forward and backward connections with distinct physiological profiles,[71,150] receptive fields of simple and complex cells in early visual cortex,[183] and other findings.[86,88,141] First and foremost, the predictive coding account is a theory of cortical responses in healthy individuals. However, it has been put forth as a theoretical framework to understand psychosis.[76,227]

As outlined in the previous paragraph, most empirical evidence in support of predictive coding centres around perception. While schizophrenia is associated with changes in perception such as reduced susceptibility to certain visual illusions[130] and perceiving stimuli in the absence of external stimulation (i.e., hallucinations), a core symptom of psychosis has been conceptualised in terms of beliefs rather than perceptions.[118] Delusions are often defined as false beliefs that are held with conviction, despite evidence that suggests otherwise and which are resistant to change.[118]

Traditionally, cognitive theories have distinguished between a belief and a perceptual system.[76] Both, abnormal perception as well as abnormal beliefs were proposed as explanations for symptoms of schizophrenia.[76] However, Fletcher and Frith[76] argued that perceiving is in fact the same as believing when conceptualising the brain as Bayesian inference machine that is hierarchically organised. Under this framework, symptoms of schizophrenia can be be understood as perturbations of the weighting of incoming sensory information and prior expectations at different levels of a processing hierarchy.[76,227] Perception can be thought of as the lower levels of this hierarchy, whereas beliefs may be localised at higher levels.[76] However, what type of information is hierarchically organised remains still unclear.[260] Higher levels could, for example, represent a hierarchy of causes[76] with an increasingly higher degree of abstraction or be structured according to time such that higher levels are associated with more slowly changing processes[128] or both.

Importantly, the predictive coding account of psychosis assigns a critical role to precision-weighting of both sensory evidence (i.e., the likelihood) and prior information,[76,87,227] where the precision that weights sensory evidence or prior information corresponds to the inverse of uncertainty. Uncertainty is usually expressed as the variance of a probability distribution, which quantifies a belief. Increased sensory precision will lead to larger updates of the internal model in light of new sensory evidence. This can naturally explain phenomena like the uncanny resistance to the hollow-mask illusion observed in patients with schizophrenia.[130]

This illusion has been explained through a strong expectation (or prior) that faces are convex in healthy individuals that may be overruled by more precise sensory PEs in psychosis.[227] Furthermore, this account can explain symptoms such as delusions of control – the belief that one's body parts are controlled by alien forces – by reduced precision of predictions about the consequences of one's own movements in line with an impaired efference copy of motor commands, which can lead to an altered attribution of agency.[70,94,227,235]

Importantly, PEs and associated precisions at different levels of hierarchical Bayesian inference – the key computational quantities hypothesised to be impacted in psychosis – may map onto different neuromodulatory systems, such as the dopaminergic, cholinergic, or serotonergic system,[87,221] although this mapping may be complex.[113,114,115] Thus, casting different symptoms of psychosis as instances of hierarchical Bayesian inference may not only be conceptually useful,[76,227] but may enable better understanding of the functional role of these neurotransmitter systems[113,115] and possibly even aid efforts to stratify schizophrenia spectrum patients into more homogeneous subgroups[227] and predict treatment response.[105] While predictive coding as a process theory defined at Marr's[151] computational level of analysis appears to be promising, multiple algorithmic approaches exist.[220] In this thesis, I employ the Hierarchical Gaussian Filter (HGF)[153,154] to cast different symptoms of psychosis as instances of hierarchical Bayesian inference and test the clinical utility of this approach.

0.5 The Hierarchical Gaussian Filter

Bayes theorem prescribes an optimal way of learning under uncertainty. While a plausible argument can be made that the human brain should have evolved to implement Bayesian inference,[100] and there is some evidence in support of this notion,[15,135,265] Mathys and colleagues[153] pointed out that there are a at least three serious issues with the notion that the brain can be understood as an ideal Bayesian learner: (1) Solving Bayes theorem often requires time-consuming computations of a complex integral. (2) It is unclear how this computation would be implemented in the brain, and (3) Bayesian inference constitutes a normative framework describing how information *should* be integrated, yet there are many cases in which individuals make sub-optimal decisions and even arrive at different conclusions, when provided with the same prior knowledge and sensory evidence, i.e., there is considerable inter-individual variability (examples of this will follow in Chapters 1 and 2).

To address these limitations, Mathys et al.[153,154] introduced a generic hierarchical Bayesian framework, the HGF. This model assumes a hierarchy of – in principle, infinitely many[154] – hidden states that are coupled via their variances. Each state evolves in time as a Gaussian random walk centered on the value of the state at the previous time point, where the step size of the random walk at each level is determined by the level above. Problem 1 was addressed by Mathys and colleagues[153,154] by deriving efficient closed-form updated rules using a variational Bayesian inversion scheme, maximising negative free energy, a lower bound to the log model evidence.[93,153] The authors assume Gaussian distributions for the hidden states and use

mean-field approximations and a quadratic approximation to the variational energies. This approximation is similar to the Laplace approximation, but differs from it in that the expansion point was chosen to be the value of the state at the previous time point (see [153] for more details). Under a set of minimal assumptions, assuming Gaussian distributions for the hidden states adheres to the principle of maximum entropy constituting the least arbitrary choice of distribution.[123,153] These assumptions are (1) that the brain performs some approximation to Bayesian inference, because exact inference is likely to be too costly and (2) that the posterior can be efficiently encoded by its first two moments.[153] The update equations derived under these assumptions are of the following form:

$$\underbrace{\Delta\mu_i^{(k)}}_{\text{Belief Update}} \propto \overbrace{\frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}}}^{\text{Precision Ratio}} \underbrace{\delta_{i-1}^{(k)}}_{\text{Prediction Error}}, \tag{2}$$

where $\mu_i^{(k)}$ is the expectation or belief at trial k and level i of the hierarchy, $\hat{\pi}_{i-1}^{(k)}$ is the precision (inverse of the variance) from the level below (the hat symbol denotes that this precision has not been updated yet and is associated with the prediction before observing a new input), $\pi_i^{(k)}$ is the updated precision at the current level, and $\delta_{i-1}^{(k)}$ is a PE expressing the discrepancy between the expected and the observed outcome. This update equation bares a striking resemblance to the update equations of other learning models, such as the Rescorla-Wagner model (see [185] for more details):

$$\underbrace{\Delta V}_{\text{Belief Update}} = \overbrace{\alpha}^{\text{Learning Rate}} \underbrace{(\overbrace{u^{(k)}}^{\text{Actual Outcome}} - \overbrace{V^{(k-1)}}^{\text{Predicted Outcome}})}_{\text{Prediction Error}}, \tag{3}$$

where V is the learned association strength between two stimuli, α is a learning rate, $u^{(k)}$ is the outcome at trial k, and $V^{(k-1)}$ is the prediction before observing the outcome.

While structurally similar, these models display an important difference with respect to how they define the learning rate. Whereas the learning rate is constant in the Rescorla-Wagner model, the learning rate of the HGF is dynamically regulated according to a precision ratio of the precisions at the level below and the precisions at the current level; both are updated on each trial and change over time. Similar comparisons can be made with the update equations for temporal-difference learning[232,233] and Q-learning[251,252]:

$$\underbrace{\Delta U^{\pi}(s)}_{\text{Belief Update}} = \overbrace{\alpha}^{\text{Learning Rate}} \underbrace{(\overbrace{R(s) + \gamma U^{\pi}(s')}^{\text{Actual Reward}} - \overbrace{U^{\pi}(s)}^{\text{Predicted Reward}})}_{\text{Prediction Error}}, \tag{4}$$

where $U^{\pi}(s)$ is the utility associated with state s under a policy π, α is a learning rate, $R(s)$ is the reward associated with state s, γ is a discount factor that down-weighs future rewards, and s' is the next state. Q-learning[251,252] is structurally similar, but reformulates reinforcement learning as learning the values of state-action pairs or so-called Q-values:

$$\underbrace{\Delta Q(s,a)}_{\text{Belief Update}} = \overbrace{\alpha}^{\text{Learning Rate}} \underbrace{(\overbrace{R(s) + \gamma \max_{a'} Q(s',a')}^{\text{Actual Reward}} - \overbrace{Q(s,a)}^{\text{Predicted Reward}})}_{\text{Prediction Error}}, \tag{5}$$

where $Q(s,a)$ is the Q-value associated with taking action a in state s, α is a learning rate, $R(s)$ is the reward associated with state s, γ is a discount factor, and $Q(s',a')$ is the Q-value associated with taking action a' in the next state s'. Again, these models assume a constant learning rate whereas the HGF has a dynamic learning rate, although other formulations, for example a decaying learning rate are used in practice as well. A second important difference between the HGF update equations and these ones is that they also consider actions and future outcomes in addition to the current outcome.

In short, Mathys and colleagues provide structurally interpretable and, more importantly, efficient update equations that allow performing approximate Bayesian inference in a time-efficient manner providing a solution for Problem 1 of Bayesian accounts of the brain, namely that computing exact Bayesian inference is computationally costly. However, can this belief updating process be implemented biologically by the brain (Problem 2)? As outlined in the predictive coding section, there is considerable evidence that the brain may compute quantities like PEs and precisions, which are at the core of the HGF's update equations.[153,154] Moreover, these quantities may map onto different neuromodulatory systems,[87,113,114,115,221] and are hypothesized to be affected in psychosis,[227] rendering the HGF an attractive model to investigate psychosis. Lastly, Mathys et al.[153,154] introduce a set of subject-specific parameters, which govern the dynamics of the hidden states and allow to model inter-individual variability (Problem 3). These parameters can be understood as encoding an individual's approximation to Bayesian inference[154] and enable deriving a computational fingerprint that provides a concise description of an individual's learning profile.

0.6 Outlook

In this thesis, I will employ the HGF[153,154] to cast different symptoms of psychosis as instances of hierarchical Bayesian inference in three different datasets. Chapter 1 will examine the question of how persecutory delusions emerge in early schizophrenia. First, I will derive hypotheses based on previous literature and simulations, which are then tested in a sample of individuals at clinical high risk for psychosis (CHR) and first-episode psychosis patients (FEP). Chapter 2 will apply this modelling approach to delusions more broadly in a large sample of patients with psychotic disorders, who have experienced delusions in the past or were expe-

riencing delusions at the time of the study. In this chapter, I will examine the relationship between delusions and reasoning biases and assess the clinical utility of this computational approach by testing its ability to predict treatment response to a psychotherapeutic intervention. In Chapter 3, I will model changes in implicit sensory learning in early schizophrenia and investigate, whether this approach is not only conceptually useful, but also biologically plausible. Lastly, I will discuss the implications and limitations of this work in Chapter 4.

I had to make sense, any sense, out of all these uncanny coincidences. I did it by radically changing my conception of reality.

Peter K. Chadwick (1993)[30]

1 Modelling Paranoid Delusions

Paranoid delusions are one of the key symptoms in early schizophrenia, but how do they emerge? This chapter will focus on addressing this question by modelling paranoid delusions in emerging psychosis. The introduction section for this chapter outlines our hypotheses that were derived from the current literature and simulations using the HGF. This work was published by Diaconescu, Hauke, and Borgwardt (2019) in *Molecular Psychiatry*[54] and adapted for this dissertation. Novel empirical results – not included in the published article – will follow in the methods and results section.

1.1 Introduction

Persecutory delusions, defined as unfounded beliefs that others are deliberately intending to cause harm, are core symptoms of psychosis and a burden for patients.[80] Persecutory ideation leads to increased incidence of violent behaviour,[36] suicidal ideation and relapse.[19] About half of FEP with persecutory delusions show psychological well-being levels lower than 2% of the general population.[85]

A recent approach to treatment of psychosis focuses on early detection and prevention. However, a fundamental problem for research on the early phases of psychosis is identifying robust markers for transition to psychosis from the clinical high risk state.[96] The clinical high risk state is defined by the presence of one or more of the following criteria: attenuated psychotic symptoms (APS), brief and limited intermittent psychotic symptoms (BLIP), trait vulnerability, as well as a marked decline in psychosocial functioning and unspecific prodromal symptoms. Whereas clinical variables have good prognostic accuracy for ruling out indi-

viduals who will not develop psychosis, there is a need to improve the prediction accuracy of future transition to psychosis.[96,198]

Previous studies have examined the predictive value provided by neuroimaging methods including structural[45,136,137,138] and functional magnetic resonance imaging (fMRI).[17,241] In contrast to clinical and environmental variables, whole-brain examinations of structural magnetic resonance imaging (sMRI) data using voxel-based morphometry delivered the largest prediction accuracy rates, reaching about 80% prediction accuracy in a cross-centre study.[138] A recent review of predictive models for psychosis transition indicated that using multiple variables (biological, environmental, and neurocognitive), and testing them sequentially in CHR individuals may substantially improve prediction rates.[198] This suggests that a multimodal, combinatorial approach is needed.

Although current methods link transition risk with particular differences in genetic polymorphisms or brain structures, they do not allow for quantifying the probability that a particular disease mechanism is present. This, however, is the basis for targeted treatment.

One solution for identifying disease mechanisms is to pursue a computational modelling strategy and employ generative models that focus on core symptoms, such as persecutory delusions. Generative models describe mechanisms that could have generated the observed behaviour or neuroimaging data. Individual differences in behaviour – potentially related to disease mechanisms – can be uncovered by estimating individual model parameters based on participants' behavior.[226] In addition to pure risk prediction, this approach, because it is mechanistic, may also prove useful for identifying pathophysiological mechanisms of emerging psychosis (see Figure 1.1).

One class of generative models, which can be fit to noninvasive measurements (electroencephalography (EEG) or fMRI), is models of effective connectivity such as dynamic causal modelling (DCM), describing causal (directed) influences between neurons or neuronal populations.[92] DCMs explain measured brain activity as arising from circuit dynamics that are a function of (1) intrinsic connectivity, (2) experimentally-induced perturbations, and (3) modulatory inputs that invoke contextual changes in synaptic strengths (i.e., short-term plasticity during learning or neuromodulatory influences). A complementary approach to neuroimaging-based models is afforded by generative models of behaviour. These can be fitted to trial-by-trial behavioural responses to capture (mal)adaptive aspects of learning and decision-making.[224]

Here, we introduce a computational framework that focuses on a central symptom of psychosis, namely persecutory ideation. This framework integrates computational models of behaviour with neural circuit models, which describe the neuronal causes of aberrant learning and can be fit to EEG and fMRI data. It makes specific predictions about pathophysiology in psychosis, which may be used to predict transition to psychosis in CHR individuals and treatment response in FEP.

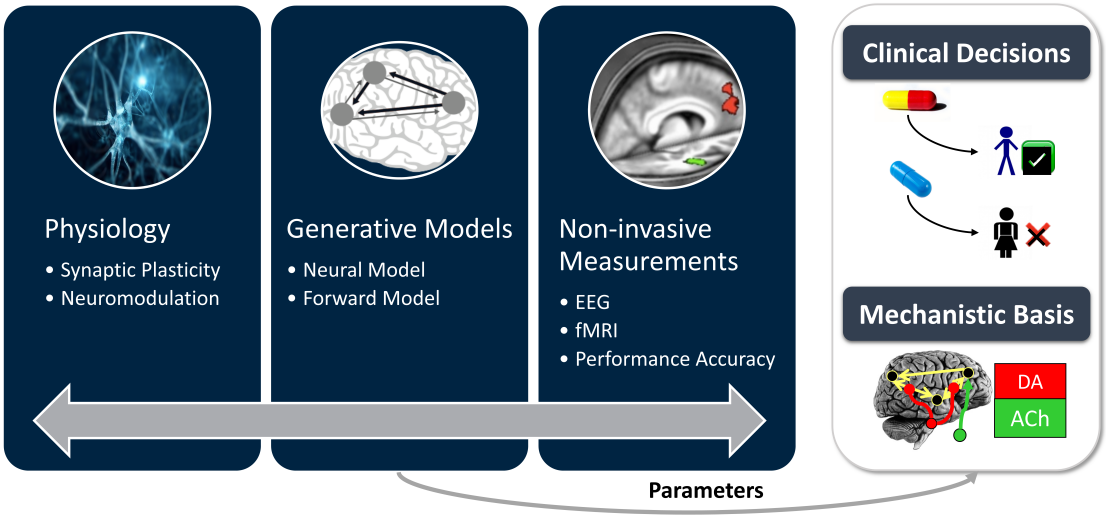

Figure 1.1: Generative models of behaviour and neuroimaging data. Generative models represent the joint probability of data and model parameters and allow us to infer on physiological mechanisms from noninvasive measurements. Inference refers to the application of mathematics to draw conclusions in the presence of uncertainty. This approach is mechanistic in the sense that it allows to identify causes of learning (e.g., behavioural – prediction error and precision - or neuronal - synaptic plasticity and neuromodulation). In the context of psychosis, parameters capturing key aspects of pathology (e.g., disruptions of DA-mediated NMDA-receptor plasticity) can be used to make clinical predictions such as treatment response, in order to inform clinical decisions, for example whether a given individual should be administered medication with a principally dopaminergic action. **EEG**: Electroencephalography. **fMRI**: Functional magnetic resonance imaging. **DA**: Dopamine. **NMDA**: *N*-methyl-D-aspartate. **ACh**: Acetylcholine. Adapted from Diaconescu, Hauke, and Borgwardt (2019), *Molecular Psychiatry*.[54]

1.1.1 Computational accounts of persecutory delusions

Delusions in general are conceptualised as false beliefs based on incorrect inference about the external world, which persist in the face of disconfirmatory evidence. Two major computational theories exist, which assume specific mechanisms of delusional belief genesis and persistence.

First, a popular notion is that patients with psychosis attribute inappropriately high aberrant salience to irrelevant events. This theory posits a key role of the dopamine system in mediating the misattribution of salience (for a review, see[261]). It is consistent with well-established theories of increased phasic dopamine release in psychosis[101,110,131,212] and supported by a host of fMRI studies in FEP.[171,197,217] Although compelling, this theory does not provide an explanation how aberrant salience attribution leads to the development of uncorrectable delusional beliefs.

A second and related theory of delusions focuses on the Bayesian brain hypothesis and the interplay between prior beliefs and "correction" signals or PEs.[76,95] The Bayesian account of perception proposes that the brain generates predictions about its sensory inputs and adjusts those predictions via incoming PEs.[86,183] Adopting a hierarchical Bayesian framework,

beliefs at multiple levels, from discrete sensory events to more abstract aspects of the environment (e.g., probabilistic associations and volatility), are updated based on precision-weighted PEs.[153,154]. Specifically, in hierarchical models, a ratio of precisions (assigned to sensory inputs relative to prior beliefs) serves to scale the amplitude of PE signals and thus their impact on belief updates.[153]

Recent theories of perceptual abnormalities in psychosis have built on hierarchical Bayesian frameworks extending the concept of aberrant salience by highlighting the role of uncertainty (or its inverse, precision).[3,38,42,76,221] One specific suggestion from these accounts is that aberrantly strong (or precise) incoming PEs indicate that prior predictions are inadequate and beliefs or actions must be changed to accurately predict states in the world. Thus, a plethora of incoming error signals leads to a brittle (or uncertain) model about states in the world, which ultimately sets the stage for the formation of delusions.[41,116] High-order beliefs of abnormally low precision lead to a lack of regularisation, which renders the environment seemingly unpredictable and volatile, enhancing the weight of incoming PEs.[3] A brittle model of the world may require adoption of extraordinary higher-order beliefs. Notably, these explanations are not exclusive but could co-exist; specifically, they relate to numerator and denominator of the precision ratio in Eq. 1 of Figure 1.2.

Fully developed delusions could be understood as implausible beliefs with overly high precision, which function to attenuate aberrant sensory evidence.[3] Recent studies have shown that strong prior beliefs govern the belief updating process in individuals who reported auditory hallucinations (hearing voices).[181] Prior beliefs were also more resistant to change in psychosis patients with acute delusions.[262] Furthermore, the utilisation of prior knowledge correlated with positive symptom severity in a perceptual discrimination task.[195] However, the study also reported decreased impact of experimentally-induced priors on the behaviour of psychosis patients[195] (also see[117]). On the other hand, a recent study found that delusion-prone individuals showed a reduced influence of experimental priors in perceptual but not cognitive discrimination tasks.[229] These somewhat ambiguous results may be reconciled by a developmental change in prior utilisation and/or distinct impact of belief precision at different levels of the processing hierarchy.[3,227]

In the context of psychosis, the most prominent delusional beliefs pertain to the social world and result from inference about the mental states of others, specifically that their intentions are of a persecutory nature.[81,189] A precise predictive model is particularly important for social contexts when interpreting others' intentions,[78,236] because human intentions are typically concealed or only expressed indirectly, requiring predictions from observations of ambiguous behaviour. Higher-level prior beliefs, which shape one's perception of others, may arise from one's own psychotic experiences including hearing voices, since individuals tend to regard their own predictions about states in the world as more reliable than second person accounts.[266]

Computational models of persecutory delusions must be based on existing cognitive models. Key cognitive predispositions for persecutory ideation are in line with the hypothesis of

an initially uncertain predictive model of others' intentions (for a review, see[6,18,21,81]): Individuals who later develop persecutory delusions report high levels of worry and rumination about how others perceive them.[84,188] These findings relate to the proposal of weak prior beliefs leading to causal misattribution.[116] The notion that persecutory ideation may be associated with abnormal inference and imprecise prior beliefs has been related to the jumping-to-conclusions (JTC) bias (e.g.,[98,178,219]; but see[170] and[1] for alternative interpretations). Individuals with persecutory delusions may adopt implausible explanations in social contexts[262] and overly negative attributions about others (e.g., negative events are attributed to active, malevolent intentions of another person).[182]

With regard to pathophysiology, psychosis represents a spectrum of disturbances in the interaction between NMDA-receptor dependent synaptic plasticity and neuromodulatory systems like dopamine and acetylcholine (see[222] for a review and[16,175] for recent empirical findings). However, the link between impaired social cognition, persecutory delusions, and disruptions in synaptic plasticity by neuromodulatory systems has not been established. This is because it requires ecologically valid and deception-free experimental paradigms that have also been studied neurobiologically.

Here, we propose such a paradigm to test the hypothesised link between social inference and persecutory ideation. This paradigm was adapted from a previous social learning task[15] and probes how one infers on the intentions of another agent (adviser) who provides iterative advice about the outcome of a probabilistic task based on additional information that the adviser obtains on every trial (Figure 1.2A). Importantly, this task maps onto existing pathophysiological mechanisms of psychosis.[57]

1.1.2 Inferring on others' intentions: A framework for probing persecutory delusions

To understand the genesis and persistence of persecutory delusions, the computational framework needs to be examined in an experimental context that is sensitive to the process of interest. Therefore, we propose a paradigm that has been developed to specifically address persecutory ideation, as it requires learning about the hidden and possibly changing intentions of another person. It requires hierarchical processing from non-social to social representations with increasing levels of abstraction, which can be mapped onto hypothesised pathophysiological mechanisms of psychosis, in particular precision-weighted PE belief-updating.[55,57]

Participants perform a binary lottery task and are additionally given advice from a more informed agent (the adviser) about which option to choose. In order to perform well, they not only have to predict the accuracy of current advice, but also the adviser's intention and how it might change over time (volatility; Figure 1.2a, upper panel). To examine the impact of precision on learning from advice, we manipulated volatility and thereby varied the association strength between the advice and the outcome. We assumed that the higher-level belief precision about the adviser's fidelity is low, when volatility is high and vice-versa.

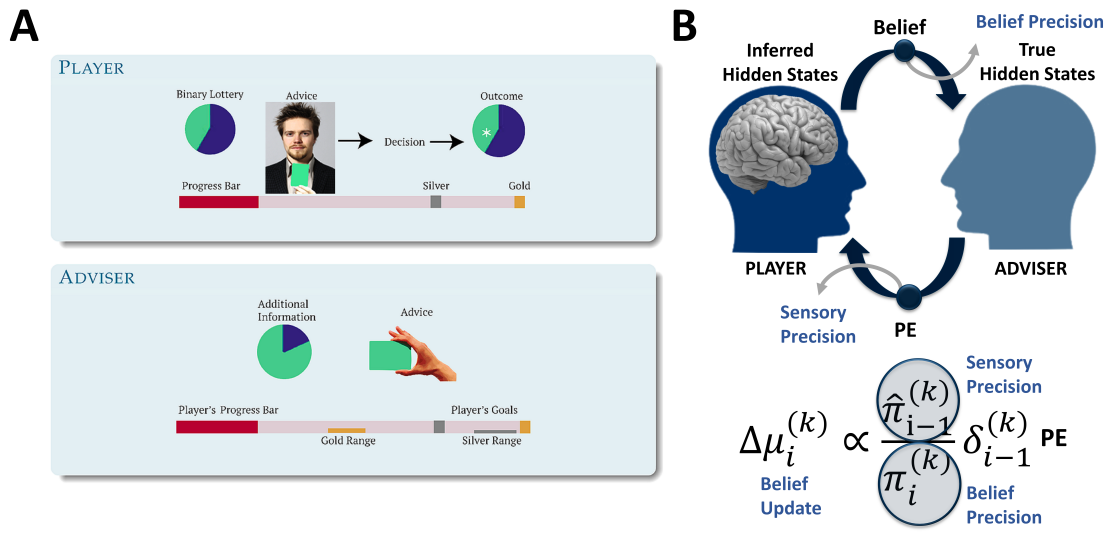

Figure 1.2: Probing persecutory ideation: Inferring on others' intentions experimental paradigm and computational model. **A** Participants took part in a face-to-face advice-taking task for monetary rewards and were randomly assigned to *player* and *adviser* roles. Players had to predict the outcome of a binary lottery draw, whereas advisers gave players suggestions on which option to choose. Both sets of participants received incentives and the pay-off structure differed to ensure the presence of both collaboration and competition between the two participants. Players profited from the adviser's recommendations as advisers always received more information about the outcome of the lottery (constant probability of 80%) whereas advisers gained from the players' compliance to take the advice into account. The advisers' motivation to provide valid or misleading information varied during the game as a function of their own incentive structure. Players were (truthfully) informed that the adviser had their own (undisclosed) incentive structure and because of it, intentions could change during the game (volatility). The social learning task was adapted for fMRI or EEG recordings by using 2 s video clips of the advisers recorded during the interactive sessions. **B** According to the model, agents infer on true hidden states in the world by continuously updating their predictions (or beliefs) via precision-weighted prediction errors (PEs). Assuming Gaussian distributions over beliefs, these can be described by their sufficient statistics, the mean (μ) and the variance/uncertainty (σ) or its inverse precision/certainty (π). Predictions about hidden states in the world (before observing an outcome), are denoted with a hat symbol (e.g., $\hat{\pi}$). At each hierarchical level i, belief updates (posterior means $\mu_i^{(k)}$) on each trial k are proportional to precision-weighted PEs. The belief update is the product of the PE from the level below $\delta_{i-1}^{(k)}$ weighted by a precision ratio: The ratio is composed of $\hat{\pi}_{i-1}^{(k)}$ and $\pi_i^{(k)}$, which represent estimates of the precision of the predicted input from the level below (sensory precision) and precision of the belief at the current level, respectively. **EEG**: Electroencephalography. **fMRI**: Functional magnetic resonance imaging. Adapted from Diaconescu, Hauke, and Borgwardt (2019), *Molecular Psychiatry*.[54]

The adviser's intentions and motivation to provide helpful advice change according to the incentive structure of the task (Figure 1.2a, lower panel). The task was adapted for testing along with either EEG or fMRI recordings by replacing face-to-face interactions with videos of the advisers, taken from trials when advisers truly intended to help or to mislead the players.[55,57] This ensured that all participants received the same input structure and therefore could be compared in terms of their learning parameters and how they inferred from advice. Although each participant received the same advice sequence throughout the task, the advisers displayed in the videos varied between participants to ensure that physical appearance and

sex did not impact on their decisions to take advice into account.

While there are other multiround trust games, which could potentially be used to examine persecutory ideation (see [132,182]), there are several features of the current paradigm that make it particularly useful for probing persecutory ideation. First of all, it is ecologically-valid: the videos of advice reflected instances when the adviser truly intended to help or truly intended to mislead the participant. Second, it is deception-free: Participants were fully informed that the adviser had a different incentive structure and thus was motivated to not always offer helpful advice (see [56] for details). Third, in contrast to other theory of mind tasks (e.g. the mind in the eye task, emotion recognition tasks, or variations of the Sally-Ann task) or decision-making tasks (a single-shot or short multiround dictator or trust game), this paradigm includes a prolonged, iterative interaction, which allows the examination of how beliefs are updated as a result of contradicting evidence or precision-weighted PEs. Fourth, it provides a context to test what we hypothesise to be impaired in persecutory ideation, namely the different contributions of sensory compared to belief precision. Finally, the paradigm includes volatility (due to the incentive structure offered to advisers), which can be used to manipulate the players' confidence about their estimates of adviser's fidelity.

1.1.3 Inferring on others' intentions as precision-weighted prediction error updates

In the context of learning about intentions, different hypotheses about how participants took decisions (i.e., going with or against the advice) were formalised in terms of a model space, which comprised different models of learning and belief-to-action mapping, including reinforcement learning models, that were formally compared. [225] The model, which best captured behaviour in this social learning task across multiple datasets, [55,56,57] was the HGF, [153,154] which emphasized the role of hierarchical precision-weighted PEs in belief updating (Figure 1.3B). Irrespective of participant-adviser assignment, but specific to the social task, we observed the same winning model, which assumed hierarchical learning about the advice and the volatility of the adviser's intentions as the mechanism for mapping beliefs to decisions. [56]

In previous studies, the inferred adviser fidelity and volatility of intentions estimated with the HGF reflected participants' overtly expressed beliefs about the adviser's intentions at different times during the task. Furthermore, the learning parameters describing each individual's belief updates predicted participants' ratings of their own perspective-taking tendencies, suggesting that the model captures key aspects of social cognition. [56,57]

According to this model, surprising advice outcomes have a greater impact on the agent's internal representation (and should have more influence on the belief update) when the sensory precision from the level below (i.e., $\hat{\pi}_{i-1}^{(k)}$) is high. For example, participants may have regarded unexpected misleading advice as evidence that the adviser has changed the strategy, thus adapting their beliefs about the adviser's intentions and decisions to follow the advice. However, if one has a strong prior belief that the adviser's intentions are to mislead, then the belief precision (i.e., $\pi_i^{(k)}$) is high and contrary evidence (i.e., surprising helpful advice) will be

ignored.

In summary, our proposal suggests that persecutory delusions can be understood as an imbalance between sensory and belief precision. Increased sensory precision augments the impact of social PEs on beliefs about fidelity, and likely marks the early stages of psychosis, whereas increasing belief precision has the opposite effect on belief updates and may reflect the consolidation of delusions. This is because belief precision refers to the confidence in one's model of intentions, which functions to "explain away" instances of incorrect advice.

One could appreciate the distinct impact of sensory compared to belief precision on the belief updating process with simulations (see Figure 1.1) and by considering the following intuitive example: Imagine that you buy bread from your local baker every morning. Every time, they offer you one of the two types of bread that is freshest that day. One day, you become sick after eating the loaf of bread recommended to you, implying an overly high precision at the sensory level. The next day, the baker recommends you confidently the same bread. You conclude they must have no clue about bread, and choose the other option (i.e., opposite of their advice). This reflects a process of "explaining away" PEs, by adopting a new prediction. It turns out that the other bread has an intense, pungent smell (referring to the aberrant salience of sensory inputs). This leads you to believe that the baker is purposely trying to poison you with bad bread, and even when they recommend a "good" bread that others in the store also buy, it further confirms your prediction that it is part of an elaborate plan to coax you to trust them again. This reflects the adoption of false and precise high-level beliefs, which can fully explain any instance of aberrant PEs. The aberrantly high precision on the higher-level beliefs is an adjustment in order to down-weigh the precision with respect to the sensory input (i.e., unexpected bad bread).

1.1.4 Functional anatomy of social inference

The computational quantities entering the belief updating process have been associated with neuromodulatory systems specifically implied in the pathophysiology of psychosis (for reviews, see [3,148,222]).

In the context of social learning, we demonstrated a dichotomy between low- and high-level precision-weighted PEs as they were related to dopaminergic and cholinergic systems. Whereas low-level precision-weighted PEs about advice were represented in the dopaminergic midbrain and dopaminoceptive regions such as the anterior cingulate cortex, medial, and dorsolateral prefrontal cortex, high-level precision-weighted PEs about the adviser's intentions were represented in the cholinergic septum and one specific targeted projection, the dorsal anterior cingulate cortex. Consistent results reproduced in two fMRI studies reflect fundamental neural computational architectures underlying social inference (Figure 1.3).

Not surprisingly, since social inference is particularly impaired in individuals at risk for psychosis, [28] the regions which encode these particular computational quantities include dopaminergic nuclei and dopaminoceptive areas, such as the striatum, shown to be affected in those at risk of developing psychosis [63,111] and in those who later transitioned to schizophrenia. [109]

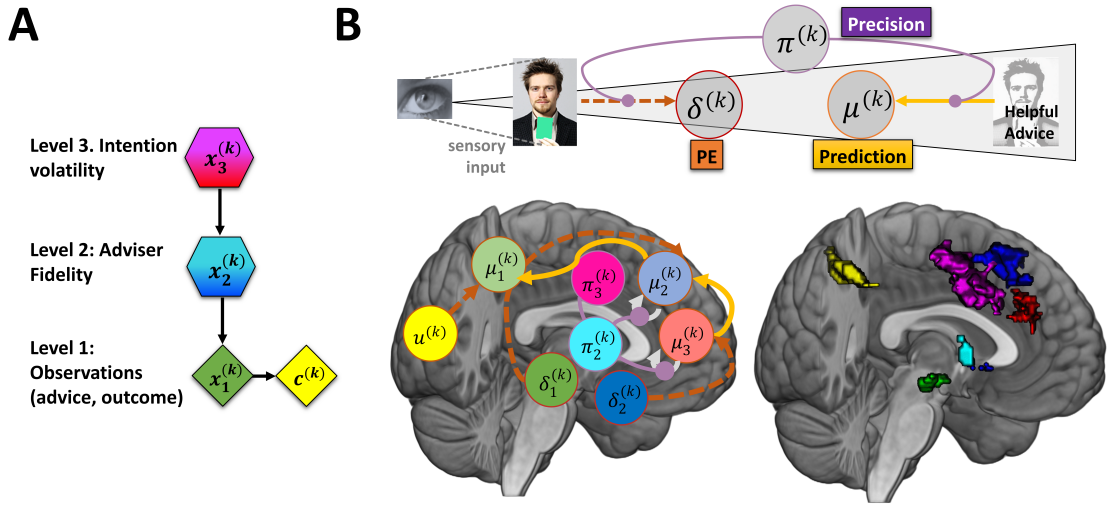

Figure 1.3: Functional anatomy of social inference. This schematic is an approximation of a neural process model of social inference. The neural signatures of the computational quantities are based on the previous, reproduced fMRI results.[57] **A** The hidden states that the agent infers on are arranged in a hierarchy as proposed by the HGF. In this graphical notation, diamonds represent quantities that change in time (i.e., that carry a time/trial index k). Hexagons, like diamonds, represent quantities which change in time, but additionally depend on the previous state in a Markovian fashion. From top to bottom, x_3 represents the volatility of the adviser's intentions, x_2 the adviser's fidelity or tendency to give helpful advice, and x_1 represents the accuracy of the current observation (advice or cue). **B** The inferred states are represented by circles. Thus, based on the empirical findings, we propose the following theoretical neural model of social inference: Cue-related PEs update predictions about the visual outcome and are conveyed via projections from lingual gyrus to posterior parietal cortex whereas advice PEs, which update the advice accuracy, are passed from low level regions (including the VTA) to higher level "theory of mind" regions, i.e., dorsomedial PFC. High-level volatility PEs are further transmitted via the cholinergic septum to cingulate regions. The precisions (advice and volatility) modulate the impact of PEs on medial PFC activity. **PE**: Prediction error. **PFC**: Prefrontal cortex. **VTA**: Ventral tegemental area. Adapted from Diaconescu, Hauke, and Borgwardt (2019), *Molecular Psychiatry*.[54]

1.1.5 Clinical predictions afforded by the computational model

As persecutory delusions predominate in major psychotic disorders and contribute to symptom severity, computational models that explain their formation and persistence may shed light onto the neural mechanisms that mark the different stages of psychosis.

In the context of social learning, we predict that the high risk state is defined by an imbalance between the precision of beliefs at low compared to high levels of the processing hierarchy, as suggested by recent studies of perceptual inference in relation to delusions.[194,196] Thus, the precision associated with advice PEs will likely be larger compared to the precision of the prediction about intentions, leading to a high learning rate and a reduced ability to form a cohesive model of the adviser's intentions, which could be predicted using simulations (Figure 1.4A).

Based on neuroimaging results in the healthy population[55,57] and recent studies of aberrant salience in the at-risk population,[197,217,250] several hypotheses about pathophysiology can be

put forward, which could be falsified in future studies: First, the early prodromal stage of psychosis may be marked by an increased low-level (sensory) precision. Consistent with previous connectivity studies,[199,200,201] this would be translated into enhanced bottom-up connectivity from dopaminergic regions to key brain regions involved in the representation of social (advice) PEs, including the temporal-parietal junction and dorsomedial prefrontal cortices.[15,57] Thus, parameters which will likely predict transition to frank psychosis include learning parameters that determine the dynamics of precision-weighted PEs (see[56]) as well as the connectivity strengths of bottom-up connections from dopaminergic to parietal and prefrontal cortices (Figure 1.4A).

In the later stages of psychosis, the presence of delusions might reflect a compensatory response to the aforementioned deficiencies of hierarchical inference. Thus, in individuals who exhibit persecutory delusions, we predict an increased representation of high-level belief precision about the other's intentions (Figure 1.4B). This notion of rigid high-level priors leads to several experimentally testable predictions: At the behavioural level, this will likely be reflected as a reduced estimate of volatility. At the neural level, this will be expressed as either (1) a reduction in bottom-up connectivity from dopaminergic regions to parietal and medial prefrontal cortices reflecting the suppression of incoming PE signals, or (2) enhancement of top-down connectivity from cingulate to medial prefrontal and to parietal regions, reflecting an enhancement of the precision of predictions about intentions, or (3) a combination of both (Figure 1.4B). While reduction in functional connectivity has featured prominently in the literature, in particular between temporal and prefrontal regions,[147,243] enhanced connectivity was also reported.[11,77]

An alternative hypothesis is that the pathophysiology underlying persecutory delusions is unrelated to precision, but instead to social PEs. Accordingly, individuals with persecutory delusions regard the adviser as purposely misleading, and therefore place greater weight on negative advice PEs. At the neural level, this would be expressed as biased predictions and enhanced PE signals for misleading advice.

1.1.6 Testable Designs

We propose two experimental designs to test our hypotheses: (1) Individuals with high risk of developing psychosis and patients with persecutory delusions could be compared in a cross-sectional design. However, while generative modelling approaches may be useful for identifying inference and neurobiological processes leading to psychosis, validation studies are needed to determine their clinical utility. Regardless of how well a model may capture a putative pathophysiology, it needs to support differential diagnosis or prognosis, for example by predicting transition to psychosis or treatment response with sufficient accuracy and in individual patients. (2) This can only be tested in prospective studies where CHR individuals and FEP patients who receive first-line treatment are assessed at multiple time points and model parameters are used to predict transition to psychosis or treatment response, respectively.

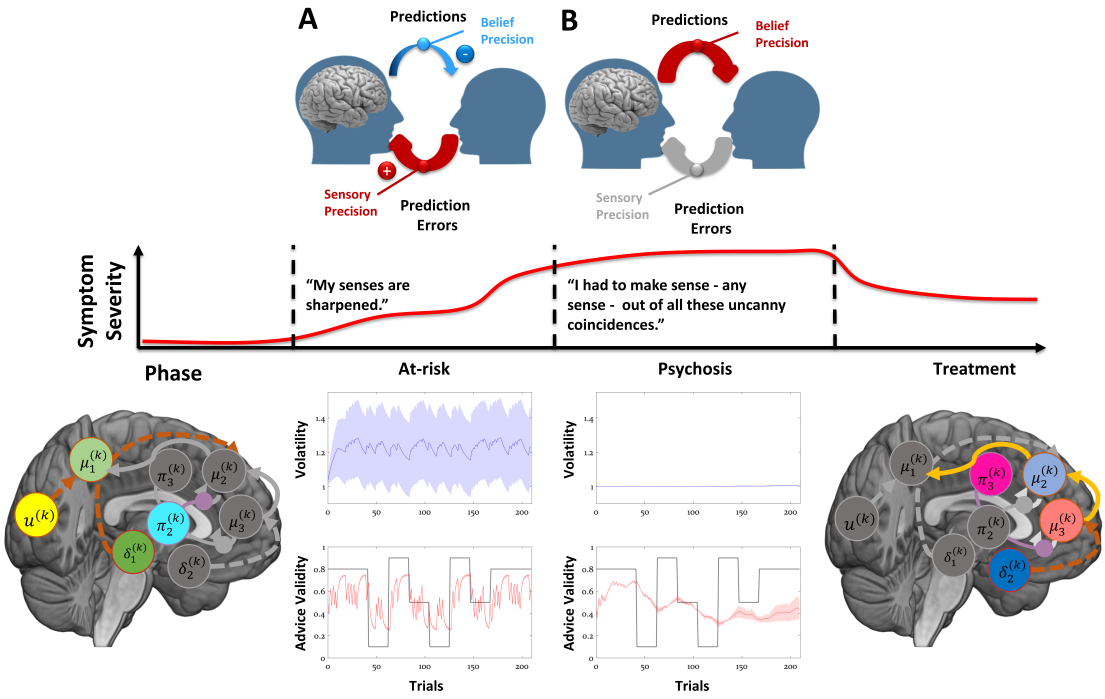

Figure 1.4: Model predictions: Beliefs and neural responses. Considering the psychosis spectrum timeline, one can make specific hypothesis about the parameters that could mark each stage by referring to Eq. 2 and the functional anatomy of social inference (Figures 1.2 and 1.3) using simulations. **A** In the early, prodromal stage of increased aberrant salience, we predict an increased representation of sensory precision ($\hat{\pi}_{i-1}^{(k)}$) with a decrease in the belief precision ($\pi_i^{(k)}$) during the social learning task. Behaviourally, the prodromal stage of increased aberrant salience is equivalent to an increased representation of sensory precision ($\hat{\pi}_{i-1}^{(k)}$) with a decrease in the belief precision ($\pi_i^{(k)}$) during the social learning task. Neurally, this may be expressed as enhanced low-level PEs and thus enhanced connectivity between dopaminergic and sensory to parietal and frontal regions. **B** In the later stages, when persecutory delusions are present, we expect an enhancement of the belief precision ($\pi_i^{(k)}$) during the social learning task; at the level of the hierarchical Bayesian model, this would be associated with reduced estimated volatility, tonic learning rate, and prior estimate about the adviser's fidelity. Neurally, this may be expressed as increased high-level precision and PEs and thus increased connectivity strength between cingulate and medial prefrontal regions. Adapted from Diaconescu, Hauke, and Borgwardt (2019), *Molecular Psychiatry*.[54]

From previous studies of aberrant learning in psychosis, it is unclear whether alterations in social inference are specifically required to explain persecutory delusions. In fact, alterations in higher-level inferential processes that are not necessarily specific to social contexts may affect processing of socially relevant information and produce delusions. To address this question, a control task which removes the aspect of intentionality may be needed. We have previously included such a control task[56] with blindfolded advisers who selected their advice from pre-defined card decks, thus eliminating the effect of intentionality, and demonstrated that the computational model proposed here, which assumes hierarchical learning about the advice and volatility of the adviser's intentions as the mechanism for mapping beliefs to decisions specifically captured the intentionality behind the advice.[56] In terms of more broadly distinguishing between mechanisms of abnormal plasticity linked to psychosis, additional

perceptual learning tasks that tap into different mechanisms, including intact NMDA receptor signalling, such as for example the auditory MMN task[238] may also be needed.

1.1.7 Conclusion and Future Directions

Mechanistically interpretable generative models like the ones outlined here allow for model comparison and testing of competing hypotheses as well as inference on disease mechanisms in individual patients at different stages of psychosis. Furthermore, the computational quantities derived from the model – such as the low- and high-level, precision-weighted PEs – could be associated with distinct neuromodulatory systems, dopaminergic and cholinergic,[57] respectively, which are ultimately the targets of pharmacological treatment in psychosis. Future studies in subclinical and clinical populations will examine the usefulness of this approach for predicting transition to psychosis or treatment response in individual patients.

1.1.8 Empirical evidence

In the remainder of this chapter, I will present empirical evidence from a new cross-sectional study, in which we investigated the computational mechanisms underlying emerging psychosis and tested the predictions afforded by the model outlined in Figure 1.4.

1.2 Methods

1.2.1 Participants

We included 19 CHR, 19 healthy controls (HC) that were matched to CHR with respect to age, gender, handedness, and cannabis consumption and 18 minimally-medicated ($\leq$ 10 days) FEP resulting in a total sample of $N = 56$ participants. FEP were recruited from both inpatient care and the outpatient departments of the University Psychiatric Hospital (UPK) Basel, CHR were recruited from the Basel Early Treatment Service (BEATS) and HC via online advertisements and advertisements in public places. All participants provided informed written consent. The study was approved by the local ethics committee (Ethikkommission Nordwest- und Zentralschweiz, no. 2017-01149) and conducted in accordance with the latest version of the Declaration of Helsinki.

1.2.2 In- and exclusion criteria

All participants were required to be at least 15 years old. Specific inclusion criteria for FEP were a first diagnosis of an acute psychotic disorder, which was assessed by the treating clinicians, and a treatment recommendation to begin neuroleptic medication issued independently of the study.

We included CHR who fulfilled either ultra-high risk for psychosis criteria, i.e. one or more of the following (1) APS, (2) BLIP, (3) a trait vulnerability in addition to a marked decline in psychosocial functioning also referred to as genetic risk and deterioration syndrome (GRD), assessed with the Structured Interview for Prodromal Symptoms (SIPS)[157], or basic symptom criteria,[133,206] i.e., cognitive-perceptive basic symptoms (COPER) or cognitive disturbances (COGIDS) assessed with the Schizophrenia Proneness Instrument, adult version (SPI-A)[207] or the Schizophrenia Proneness Instrument, child and youth version (SPI-CY)[208] by experienced clinical raters.

Important exclusion criteria were history of previous psychotic disorders, psychotic symptomatology secondary to an organic disorder, any neurological disorder (past or present), premorbid IQ < 70 (assessed with the Mehrfachwahl-Wortschatz-Test, Version A[142]), colour blindness, substance-abuse diagnoses according to ICD-10 criteria (except cannabis), alcohol or cannabis consumption within 24 hours prior to measurements, and regular drug consumption (except alcohol, nicotine, and cannabis), which was assessed during the admission interview and confirmed with a drug screening before the initial measurement (measurements were postponed following a positive test until a negative test result was obtained).

FEPs whose psychotic symptoms were associated with an affective psychosis or a borderline personality disorder at the time of the measurement were excluded. Since the data presented below was collected as part of a larger study that included neuroimaging assessments, additional exclusion criteria for CHR and HC were contraindications for fMRI and contraindications for EEG measurements for all three groups. However, I only present the behavioural results in this chapter.

1.2.3 Task

All participants were asked to perform a deception-free and ecologically valid social learning task (Figure 1.5A),[56,57] which required them to learn about the intentions of an adviser that changed over time. The task comprised two phases. In the first phase participants received *stable* helpful advice, whereas advisers intentions were changing more rapidly during a second phase, the *volatile* phase (see volatility schedule in Figure 1.5B). Participants were asked to predict the outcome of a binary lottery on each trial. To this end, they received information from two sources, a non-social cue displaying the true winning probabilities of the lottery, and a recommendation of an adviser (social cue) presented in form of prerecorded videos that were extracted from trials in which a human adviser either tried to help or deceive a player in a previous human-human interaction (see[56,57] for more details).

Participants were truthfully informed that the adviser received privileged – but not complete – information about the upcoming outcome and that inaccurate advice could be due to mistakes or that the adviser could pursuit a different agenda than the player and that the adviser's intentions could change during the course of the experiment.

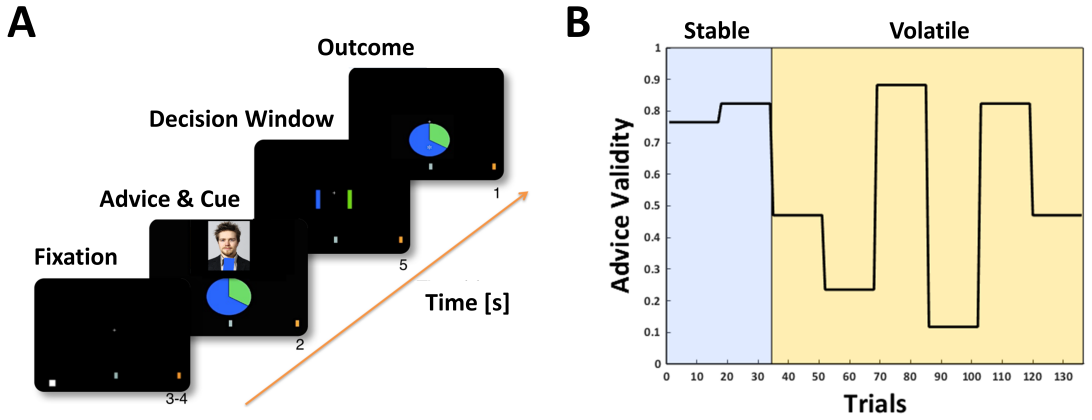

Figure 1.5: Social learning task and volatility schedule. **A** Social learning task. **B** Volatility schedule.

1.2.4 Computational modelling

Hierarchical Gaussian Filter

We modelled participants' behaviour during the social learning task with a 3-level HGF.[153,154] The model comprises a perceptual model and a response model, which will be detailed below.

Perceptual model The standard 3-level HGF assumes that participants infer on a hierarchy of hidden states in the world x_1, x_2, and x_3 that cause the sensory inputs that participants perceive.[153,154] Participants' inference on the true hidden states of the world $x_i^{(k)}$ at level i of the hierarchy on trial k are denoted $\mu_i^{(k)}$. In the context of this task, the states that participants' need to infer on based on the experimental inputs on each trial (non-social cue and advice) are structured as follows: The lowest level state corresponds to the *advice accuracy*. On each trial k an advice can either be accurate ($x_1^{(k)} = 1$) or inaccurate ($x_1^{(k)} = 0$). This state can be described by a Bernoulli distribution that is linked to the state at the second level $x_2^{(k)}$ through the unit sigmoid transformation:

$$p(x_1^{(k)}|x_2^{(k)}) = s(x_2^{(k)})^{x_1^{(k)}}(1 - s(x_2^{(k)}))^{1-x_1^{(k)}} \sim \text{Bernoulli}(x_1^{(k)}; s(x_2^{(k)})), \tag{1.1}$$

with

$$s(z) = \frac{1}{1 + e^{-z}}. \tag{1.2}$$

$x_2^{(k)}$ represents the unbounded tendency towards helpful advice $(-\infty, +\infty)$ or the *adviser's fidelity* and is specified by a normal distribution:

$$p(x_2^{(k)}|x_2^{(k-1)}, x_3^{(k)}, \kappa_2, \omega_2) \sim \mathcal{N}(x_2^{(k)}; x_2^{(k-1)}, \exp(\kappa_2 x_3^{(k)} + \omega_2)) \tag{1.3}$$

The state at the third level $x_3^{(k)}$ expresses the (log) volatility of the adviser's intentions over time and is also specified by a normal distribution:

$$p(x_x^{(k)}|x_3^{(k-1)}, \theta) \sim \mathcal{N}(x_3^{(k)}; x_3^{(k-1)}, \theta) \tag{1.4}$$

The dynamics of these states are governed by a number of subject-specific parameters, i.e., the *evolution rate* at the second level ω_2, the *coupling strength* between the second and third level κ_2, which determines the impact of the volatility of the adviser's intentions on the belief update at the level below, and the evolution rate at the third level or the *meta-volatility* θ, which we fixed to a value of 0.5 to reduce the number of free parameters. Additional subject-specific, free parameters were the *prior expectations* before seeing any input about the adviser's fidelity $\mu_2^{(0)}$ and the volatility of the adviser's intentions $\mu_3^{(0)}$ (see Table 1.1 for priors on all free parameters). These parameters can be understood as an individual's approximation to Bayesian inference and provide a concise summary of a participant's learning profile. Using a variational approximation, efficient one step update equations can be derived (see Section 0.5 The Hierarchical Gaussian Filter and[153,154] for more details), which take the following form:

$$\Delta\mu_i^{(k)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)}, \tag{1.5}$$

where $\mu_i^{(k)}$ is the expectation or belief at trial k and level i of the hierarchy, $\hat{\pi}_{i-1}^{(k)}$ is the precision (inverse of the variance) from the level below (the hat symbol denotes that this precision has not been updated yet and is associated with the prediction before observing a new input), $\pi_i^{(k)}$ is the updated precision at the current level, and $\delta_{i-1}^{(k)}$ is a PE expressing the discrepancy between the expected and the observed outcome.

We also employed a second, modified version of the HGF[37] that assumed that learning about an adviser's intentions was not only driven by hierarchical PE updates, but also included a mean-reverting process at the third level formalising the idea that an altered perception of volatility may underlie learning about others' intentions. In this mean-reverting HGF, the third level can again be described by a normal distribution:

$$p(x_x^{(k)}|x_3^{(k-1)}, \theta, \varphi_3, m_3) \sim \mathcal{N}(x_3^{(k)}; x_3^{(k-1)} + \varphi_3(m_3 - x_3^{(k-1)}), \theta), \tag{1.6}$$

where φ_3 represents a drift rate and m_3 the equilibrium point towards which the state moves over time.

In this model, we fixed the drift rate φ_3 to a value of 0.1 and estimated the equilibrium point m_3 as a subject-specific, free parameter. Note, that changing m_3 to values that are lower than the prior about the volatility of the adviser's intentions $\mu_3^{(0)}$ translates into reduced belief updates at all three levels of the hierarchy corresponding to perceiving the environment as increasingly stable over time (Figure 1.6). Conversely, if $m_3 > \mu_3^{(0)}$, the magnitude of belief updates increases in line with a perception that the environment is increasingly volatile over time and beliefs should thus be adjusted more rapidly. Lastly, if $m_3 = \mu_3^{(0)}$, agents would revert back to their prior beliefs about environmental volatility over time (i.e., "forget" about the observed inputs). For this reason, we refer to the model as *mean-reverting* HGF analogous to an Ornstein-Uhlenbeck process in discrete time.[240]

Response model The response model specifies how participants' inference on the hidden states translates into decisions, i.e., to go with our against the advice. In our case the response model assumes that participants' integrate the non-social cue $c^{(k)}$ (the outcome probability indicated by the pie chart) and their belief that the adviser is providing accurate advice $\hat{\mu}_1^{(k)}$ before seeing the outcome on the current trial k:

$$b^{(k)} = \zeta\hat{\mu}_1^{(k)} + (1-\zeta)c^{(k)}, \tag{1.7}$$

where ζ is a weight associated with the advice that expresses how much participants rely on the social information compared to the non-social cue.

The probability that a participant follows the advice ($y = 1$) can then be described by a sigmoid transformation of the integrated belief b:

$$p(y=1|b) = \frac{b^\beta}{b^\beta + (1-b)^\beta}, \tag{1.8}$$

with

$$\beta = \exp(-\hat{\mu}_3^{(k)} + \nu). \tag{1.9}$$

This relationship can be understood as a noisy mapping from the integrated beliefs to participants' decisions, where the noise level is determined by the current prediction of the volatility of the advisers' intentions $\hat{\mu}_3^{(k)}$, such that decisions become more deterministic (i.e., *exploitative*), if the environment is currently perceived as stable or more stochastic (i.e., *exploratory*), if the environment is perceived as volatile. Modelling the exploration-exploitation trade-off as a function of participants' perception of volatility was favoured in previous model selection results using the same task.[56,57] ν is another subject-specific parameters that captures decision noise that is independent of the perception of volatility (lower values indicate larger decision noise).

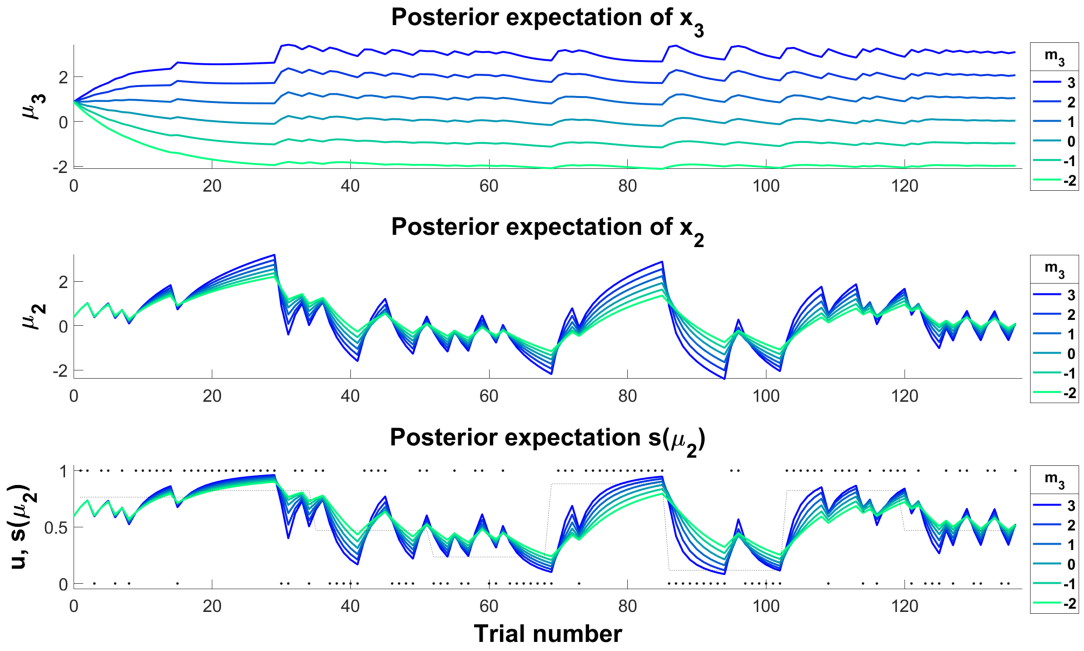

Figure 1.6: Simulating an altered perception of environmental volatility. Simulations showing the effect of changing the equilibrium point m_3. Increasing m_3 (colder colours) results in larger precision-weighted prediction errors leading to stronger belief updates across all levels of the hierarchy. Note, that high values of m_3 also increase susceptibility to noisy inputs (e.g., trials 120-136). For the simulations, all other parameter values were fixed to the values of an ideal observer given the input.

The models were implemented in Matlab (version: 2017a; `https://mathworks.com`) using the HGF toolbox (version: 3.0), which is made available as open-source code as part of the TAPAS[79] software collection (`https://github.com/translationalneuromodeling/tapas/releases/tag/v3.0.0`). The perceptual models were implemented using the 'tapas_hgf_binary' function for the standard 3-level HGF and the 'tapas_hgf_ar1_binary' function for the mean-reverting HGF.

	Equilibrium point	Coupling strength	Evolution rate	Prior expectations		Advice weight	Decision noise
Hypothesis I		$\kappa_2(logit(0.5),1),1$	$\omega_2(-2,4)$	$\mu_2^{(0)}(0,1)$	$\mu_3^{(0)}(1,1)$	$\zeta(logit(0.5),1),1$	$\nu(log(48),1)$
Hypothesis II	$m_3(1,1)$	$\kappa_2(logit(0.5),1),1$	$\omega_2(-2,4)$	$\mu_2^{(0)}(0,1)$	$\mu_3^{(0)}(1,1)$	$\zeta(logit(0.5),1),1$	$\nu(log(48),1)$

Table 1.1: Priors on free model parameters. Prior means and their respective variances are denoted in brackets, followed by upper bounds for parameters that were estimated in logit space: (Mean, Variance), upper bound.

Bayesian model selection

Based on our simulation analysis outlined in the introduction of this chapter[54] and previous findings,[37,56,58,184] we formulated competing hypotheses about the computational mecha-

nisms that could underlie emerging paranoid behaviour (Figure 1.7). A standard 3-level HGF (**Hypothesis I**) was compared to the mean-reverting HGF that assumed that learning about an adviser's intentions was not only driven by hierarchical PE updates, but also included a drift process at the third level formalising the idea, that an altered perception of volatility underlies learning about others' intentions in emerging psychosis (**Hypothesis II**; see also Figure 1.6). To arbitrate between the two hypotheses we performed random-effects Bayesian model selection.[186,225] Two additional control models were included, in which all parameters of the perceptual model were fixed to parameter values of an ideal Bayesian observer optimised based on the inputs alone using the 'tapas_bayes_optimal_binary' function to assess whether perceptual model parameters needed to be estimated for either of the two main models. We report protected exceedance probabilities φ, which measure the probability that a model is more likely than any other model in the model space,[225] protected against the risk that differences between models arise due to chance alone.[186] We also computed relative model frequencies f as a measure of effect size, which can be understood as the probability that a randomly sampled participant would be best explained by a given model. The model selection was implemented using the VBA toolbox[46] (`https://mbb-team.github.io/VBA-toolbox/`).

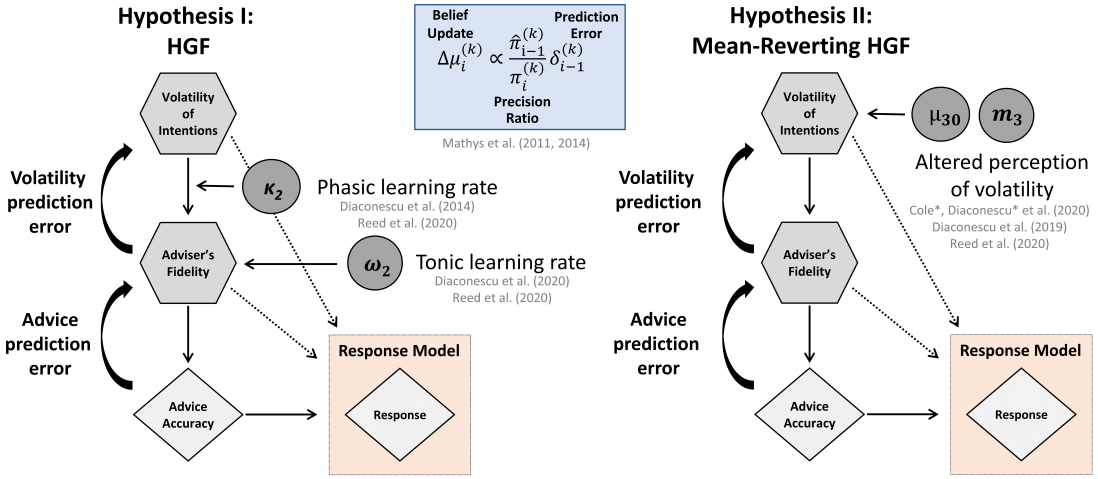

Figure 1.7: Model space. Left: Standard 3-level Hierarchical Gaussian Filter (HGF).[153,154] **Right**: Mean-reverting HGF with a drift at the third level, which captures learning about the volatility of the adviser's intentions. This model expresses the notion that early psychosis may be characterised by an altered perception of environmental volatility.

Model recovery

To assess whether models were recoverable, we conducted a series of simulations as done previously.[105] In brief, our model recovery analysis comprised simulating 20 synthetic datasets based on the empirical parameter estimates obtained from fitting all models to the empirical

data of every participant. The sample size of each synthetic dataset was chosen to be equivalent to the empirical sample size ($N = 56$). The noise level was set based on the empirically estimated decision noise ν_{est}. Each simulation was initialised using different random seeds to account for the stochasticity of the simulation. This led to a total of 4 (models) x 56 (participants) x 20 (simulation seeds) = 4,480 simulations. Subsequently, we re-inverted each of the proposed models on the synthetic data to determine, whether we could recover the true model under which synthetic data was generated. To assess model recovery, we then performed random-effects Bayesian model selection on each of the datasets with a sample size of $N = 56$ as in the empirical data and averaged the resulting protected exceedance probabilities across the 20 simulation seeds to obtain a model confusion matrix.

Parameter recovery

In line with our previous work,[105] we also performed a parameter recovery analysis to determine whether model parameter estimates were reliable. Using the simulation and model inversion results from the model recovery analysis (see preceding section), we assessed how accurately the parameters generating the data ('simulated') corresponded to the parameters that were estimated when re-inverting the same model on that data ('recovered'). We report Pearson correlations and their associated p-values to quantify our ability to recover the model parameters. Since, the significance of these correlations is influenced by sample size, we also computed Cohen's f^2, where an $f^2 \geq 0.35$ can be considered a large effect size[35] and was interpreted as evidence for good parameter recovery.

1.2.5 Statistical analysis

We tested for differences in behaviour using a linear mixed-effects model with advice taking as the dependent variable and fixed effects for group and task phase (*stable* vs *volatile*), as well as a group-by-task-phase interaction as predictors of interest and age, working memory performance, antipsychotic, and antidepressant medication as covariates of no interest. Additionally, the model included a random intercept per participant. Differences in model parameters were assessed using non-parametric Kruskal-Wallis tests. All statistical analyses were conducted in R (version: 4.04; `https://www.r-project.org/`) using R-Studio (version: 1.4.1106; `https://www.rstudio.com/`). We report both uncorrected p-values (p_{uncorr}) and Bonferroni-corrected p-values adjusted for the number of free parameters ($n = 7$). Based on previous literature and the simulations outlined in the introduction, we hypothesised that groups would differ with respect to coupling strength between the second and third level κ_2,[56,184] the evolution rate ω_2,[58,184] or parameters that are associated with the perception of volatility, i.e., the prior expectation about environmental volatility $\mu_3^{(k)}$ [184] or the equilibrium point of the drift at the third level m_3.[37,54]

1.3 Results

1.3.1 Sociodemographic and clinical characteristics

Sociodemographic and clinical characteristics are presented in Table 1.2.

1.3.2 Behavioural results

We identified a significant group-by-task-phase interaction ($F = 5.275$, $p = 0.008$; Figure 1.8A) suggesting that FEP showed reduced flexibility to take environmental volatility into account as the difference between stable and volatile phase was reduced. None of the covariates significantly impacted advice taking.

1.3.3 Modelling results

Bayesian model selection and model recovery

The model recovery analysis (Figure 1.10) indicated that the control models (CI and CII) could not be well-distinguished. This was likely due to the fact that the equilibrium point m_3 in CII was optimised based on the input alone, which resulted in a value for m_3 that was close to the prior, rendering the predictions of the two control models very similar. Most importantly, however, the two main models associated with Hypothesis I and II could be well-distinguished.

After confirming that the two hypotheses were distinguishable, we first performed Bayesian model selection including participants from all groups. The results were inconclusive ($\varphi = 74.37\%, f = 53.80\%$ in favour of Hypothesis II) possibly suggesting that different groups were best explained by different models (i.e., different computational mechanisms). To assess this possibility, we repeated the model selection for each group separately (Figure 1.9A). In HC, the winning model was the standard 3-level HGF (Hypothesis I; $\varphi = 96.63\%$, $f = 95.93\%$). Conversely, in FEP the mean-reverting HGF that included a drift at the third level was selected (Hypothesis II; $\varphi = 99.95\%, f = 95.92\%$). For CHR, we observed a more heterogeneous results: While the mean-reverting model was favoured (Hypothesis II; $\varphi = 84.57\%, f = 60.24\%$), there was also evidence for the standard HGF, albeit to a much lesser extent (Hypothesis I; $\varphi = 14.35\%, f = 37.19\%$). Further inspection of the model attributions for all individual participants revealed an interesting pattern (Figure 1.9B). All HC were attributed to the standard HGF with over 97% probability, whereas FEP were attributed to the mean-reverting model with over 99%. Interestingly, model attributions for CHR were more heterogeneous ranging from 0 to 100% probability, suggesting that some individuals were better explained by the standard HGF, but others by the mean-reverting model.

	HC $n = 19$	CHR $n = 19$	FEP $n = 18$	Test statistic	Post hoc contrasts
Age	21.37	21.05	33.44	$F = 18.182$	FEP > HC
mean [SD]	[2.52]	[3.52]	[11.70]	$\mathbf{p < 0.001}$	FEP > CHR
IQ	108.11	105.95	112.29	$F = 1.015$	
mean [SD]	[9.85]	[12.28]	[16.25]	$p = 0.370$	
Working memory[a]	6.42	6.74	5.83	$F = 1.011$	
mean [SD]	[1.71]	[2.16]	[1.98]	$p = 0.371$	
Sex	11/8	11/8	7/11	$\chi^2 = 1.767$	
f/m				$p = 0.413$	
Cannabis	7/12	8/11	5/13	$\chi^2 = 0.842$	
y/n				$p = 0.656$	
High risk type[b]					
APS		15			
BLIP		1			
GRD		0			
COGDIS		4			
COOPER		2			
Psychotic disorder diagnosis					
F20 Schizophrenia			3		
F22 Delusional disorder			6		
F23 Brief psychotic disorder			9		
Antipsychotics	19/0	3/16	16/2	$\chi^2 = 36.800$	FEP > CHR
y/n				$\mathbf{p < 0.001}$	FEP > HC
Antidepressants	19/0	9/10	1/17	$\chi^2 = 17.268$	CHR > FEP
y/n				$\mathbf{p < 0.001}$	CHR > HC
PANSS Positive	$8_{n=19}$	$11_{n=19}$	$16_{n=16}$	$\eta^2 = 0.514$	FEP > CHR > HC
median $[25^{th}, 75^{th}]$	[7, 8]	[10, 14]	[11, 23]	$\mathbf{p < 0.001}$	
PANSS Negative	$7_{n=19}$	$9_{n=19}$	$12_{n=16}$	$\eta^2 = 0.364$	FEP > CHR > HC
median $[25^{th}, 75^{th}]$	[7, 8]	[8, 10]	[9, 15]	$\mathbf{p < 0.001}$	
PANSS General	$18_{n=19}$	$29_{n=19}$	$34_{n=16}$	$\eta^2 = 0.674$	FEP > CHR > HC
median $[25^{th}, 75^{th}]$	[16, 19]	[22, 32]	[32, 40]	$\mathbf{p < 0.001}$	
PCL Frequency	$23_{n=19}$	$30_{n=19}$	$36_{n=17}$	$\eta^2 = 0.202$	FEP > HC
median $[25^{th}, 75^{th}]$	[19, 25]	[24, 33]	[23, 44]	$\mathbf{p = 0.004}$	CHR > HC
PCL Conviction	$26_{n=19}$	$33_{n=19}$	$30_{n=17}$	$\eta^2 = 0.086$	
median $[25^{th}, 75^{th}]$	[22, 31]	[28, 39]	[22, 55]	$p = 0.099$	
PCL Distress	$26_{n=19}$	$29_{n=19}$	$30_{n=17}$	$\eta^2 = 0.008$	
median $[25^{th}, 75^{th}]$	[20, 37]	[23, 38]	[21, 46]	$p = 0.799$	

Table 1.2: Demographic and clinical characteristics. All p-values are uncorrected. **HC**: Healthy controls. **CHR**: Individuals at clinical high risk for psychosis. **FEP**: First-episode psychosis patients. **APS**: Attenuated psychotic symptoms. **BLIP**: Brief and limited intermittent psychotic symptoms. **GRD**: Genetic risk and deterioration syndrome. **COGDIS**: Cognitive disturbances. **COPER**: Cognitive-perceptive basic symptoms. **PANSS**: Positive and Negative Syndrome Scale.[126] **PCL**: Paranoia Checklist.[83] Bold print highlights p-values significant at: $p < 0.05$, uncorrected. [a] Assessed with the digit span backwards task from the Wechsler Adult Intelligence Scale–Revised.[255] [b]High risk types are not mutually exclusive.

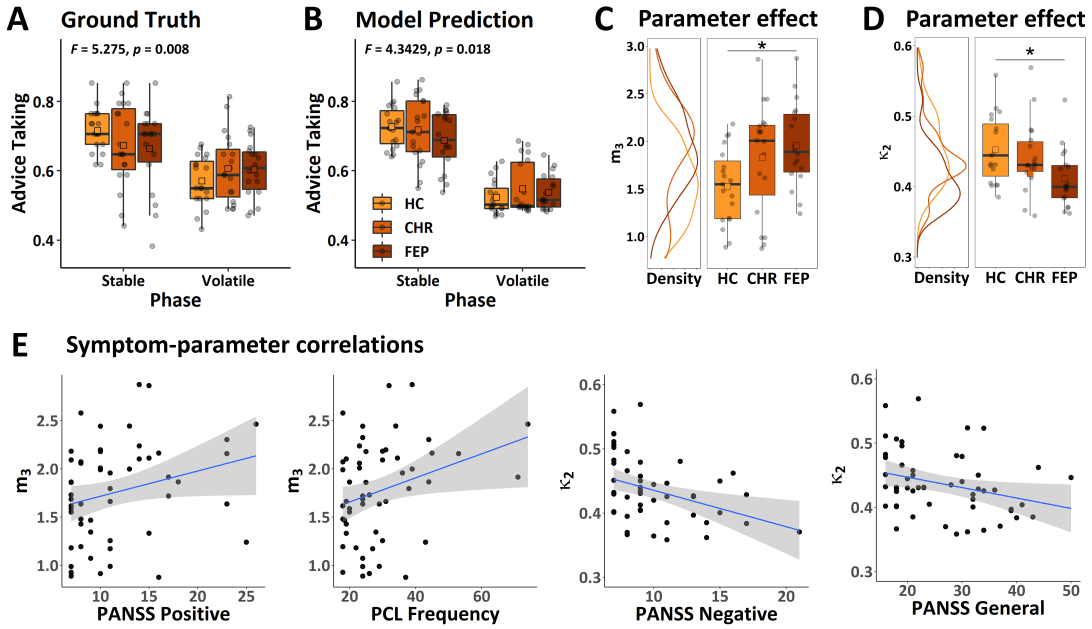

Figure 1.8: Behavioural results and parameter group effects. **A** Behavioural results (ground truth). **B** Model prediction. **C** Parameter effect for drift equilibrium point m_3. **D** Parameter effect for coupling strength κ_2. **E** Correlation between model parameters and either Positive and Negative Syndrome Scale [126] (**PANSS**) or Paranoia Checklist [83] (**PCL**). Note, that raw scores are displayed for illustration purposes only. Statistical analyses were conducted using nonparametric Kendall rank correlations. Displayed regression lines were computed using a linear model based on the raw scores. Note, that one outlier ($\kappa_2 = 0.006$) was removed for displaying the effect on κ_2 in **D** and **E**. This outlier was outside of $7\times$ the interquartile range. Excluding this participant did not affect the significance of the results. **P**: Positive symptoms. **N**: Negative symptoms. **G**: General symptoms. *F*- and *p*-values indicate results of ANCOVAs corrected for working memory performance, antipsychotic medication, antidepressant medication, and age. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: $^*p < 0.05$, using Bonferroni correction.

Posterior predictive checks and parameter recovery

To assess whether the mean-reverting model (Hypothesis II) captured the behavioural effects of interest, we conducted posterior predictive checks by repeating the behavioural analysis on this model's predictions. This analysis confirmed that the mean-reverting model recapitulated the group-by-task-phase interaction effect (Figure 1.8B). Our parameter analysis indicated good recovery (i.e., Cohen's $f^2 \geq 0.35$) for four out of the seven model parameters including the drift equilibrium point m_3 (Figure 1.10). However, recovery for $\mu_3^{(0)}$, $\mu_2^{(0)}$, and κ_2 fulfilled this criterion only in 55%, 65%, and 55% of the simulations respectively.

Parameter group effects

Since the model selection indicated that the mean-reverting model was a better explanation for behaviour of FEP, we were interested in assessing whether the perception of volatility in FEP

increased or decreased over time (see also simulations illustrating these two possibilities in Figure 1.6). To distinguish between these possibilities, we compared the drift equilibrium point m_3 across the three groups and found that m_3 was significantly different across the groups ($\eta^2 = 0.142, p_{uncorr} = 0.020$). Post hoc tests revealed that m_3 was significantly increased in FEP compared to HC suggesting that FEP perceived the intentions of the adviser as increasingly more volatile over time ($\eta^2 = 0.139, p = 0.017$, Bonferroni-corrected for the number of comparisons across groups, i.e., $n = 3$; Figure 1.8C). We also performed an exploratory analysis including all other free model parameters. This analysis revealed an additional effect on coupling strength κ_2 ($\eta^2 = 0.138, p_{uncorr} = 0.022$), which was driven by reduced coupling strength in FEP compared to HC ($\eta^2 = 0.142, p = 0.016$, Bonferroni-corrected for the number of comparisons across groups, i.e., $n = 3$; Figure 1.8D). However, neither the effect on m_3 nor κ_2 survived Bonferroni correction for the number of parameters, i.e. $n = 7$ ($p = 0.140$ and $p = 0.157$, respectively), possibly due to a lack of power.

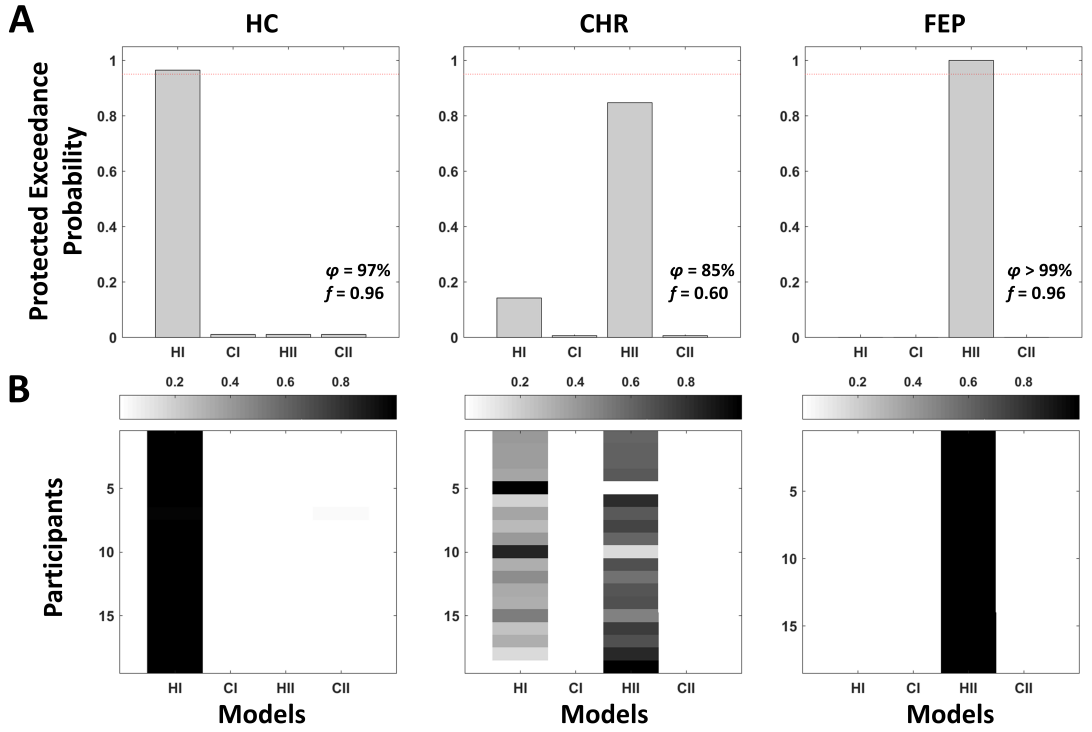

Figure 1.9: Bayesian model selection results. **A** Protected exceedance probabilities for within-group random-effects Bayesian model selection [225,186] to arbitrate between Hypothesis I (**HI**; standard 3-level HGF) and Hypothesis II (**HII**; mean-reverting HGF with drift at 3^{rd} level in line with an altered perception of volatility). Two corresponding control models were included (**CI** and **CII**), for which the perceptual model parameters were fixed. Model selection was performed separately in healthy controls (**HC**), individuals at clinical high risk for psychosis (**CHR**), or first-episode psychosis patients (**FEP**). The dashed line indicates 95% exceedance probability. **B** Model attributions for each participant.

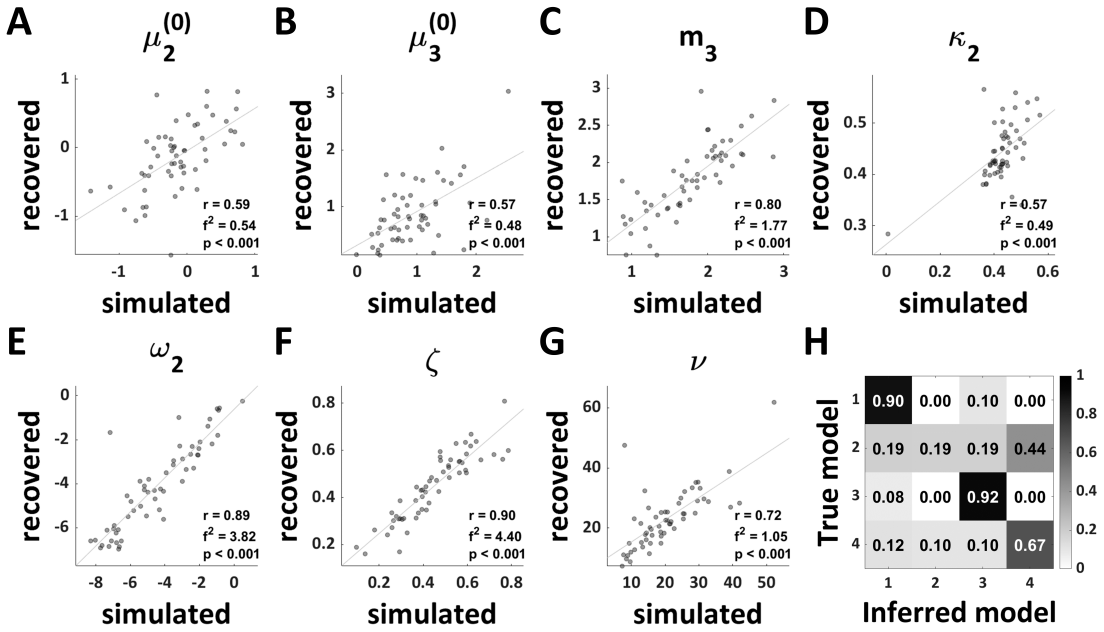

Figure 1.10: Model and parameter recovery analyses. A-G Parameter recovery result for one random seed for the mean-reverting HGF with drift at the 3^{rd} level (Hypothesis II; Figure 1.7). **H** Model recovery analysis. The grey scale indicates protected exceedance probability averaged across all 20 random seeds.

Symptom-parameter correlations

Some authors[68] have argued that psychosis may be better conceptualised as a continuum rather than categorically based on evidence that a significant percentage of the general populations reports some psychosis symptoms.[127,237] In line with this proposal, we assumed a continuum perspective and investigated whether the equilibrium point m_3 and coupling strength κ_2 were correlated with specific symptom subscales of the Positive and Negative Syndrome Scale (PANSS)[126] across all three groups with non-parametric Kendall rank correlations (see Figure 1.8E). We found a positive correlation between m_3 and PANSS positive symptoms ($\tau = 0.203$, $p_{uncorr} = 0.038$) and negative correlations between κ_2 and PANSS negative and general symptoms ($\tau = -0.253$, $p_{uncorr} = 0.011$ and $\tau = -0.219$, $p_{uncorr} = 0.022$ respectively). These correlations, however, did not survive Bonferroni correction, possibly due to a lack of power ($p = 0.228$, $p = 0.068$, and $p = 0.132$ respectively, adjusted for 2 (#parameters) x 3 (#PANSS subscales) = 6 comparisons). Since the PANSS[126] was specifically designed to assess symptom expression in clinical populations, we also calculated correlations with the Paranoia Checklist (PCL),[83] an instrument more sensitive to expressions of paranoia in healthy or subclinical populations. We found a correlation between m_3 and the PCL frequency subscale ($\tau = 0.201$, $p_{uncorr} = 0.034$). Again, this correlation did not survive Bonferroni correction ($p = 0.204$, adjusted for 2 (#parameters) x 3 (#PCL subscales) = 6

comparisons).

1.4 Discussion

In this study, we investigated the computational mechanisms underlying emerging psychosis. Our model selection results suggest that FEP may operate under a different computational mechanism compared to HC that is characterised by perceiving the environment as *increasingly volatile*. A strength of our study is that this effect is unlikely due to medication effects as FEP were only minimally medicated. Furthermore, we observed more heterogeneity in CHR, possibly indicating that this modelling approach may be useful to stratify the CHR population and identify individuals that are more likely to transition to psychosis. Assuming a psychosis continuum perspective, we also found tentative evidence suggesting that the drift equilibrium point m_3 and the coupling strength between hierarchical levels κ_2 may be affected in emerging psychosis and that these parameters provide a clinically relevant description of individuals' learning profiles. However, due to the small sample size, these results should be interpreted with caution.

1.4.1 Related modelling work

The predictive coding account of psychosis[227] and the aberrant salience hypothesis[125] propose that psychosis may be characterised by aberrant PEs that provide the breeding ground for delusions to form. Our results are in line with these proposals and the prediction for early psychosis derived in the introduction of this chapter through simulations[54] (Figure 1.4A). Moreover, our results enable a more nuanced characterisation and point towards an altered perception of environmental volatility as a possible computational mechanisms underlying aberrant PEs. Specifically, perceiving the intentions of another person as increasingly volatile over time leads to reduced precision of beliefs about environmental volatility. This, in turn, results in larger precision-weighted PEs through decreasing the denominator of the precision ratio that weighs PEs (see Equation 2). However, we note that this model was only conclusively selected in the FEP group and not already in the CHR group, although the mean-reverting model was favoured in the model attributions for some CHR individuals (Figure 1.9B). In contrast to our a priori hypothesis (Figure 1.4B), we did not find evidence for a compensatory increase in the precision of high-level priors. This was proposed as a cognitive mechanism to make sense of aberrant PEs by Kapur and colleagues[125] and observed empirically by others,[14,58,257] although Baker et al.[14] used a non-social probabilistic reasoning task.

Reed and colleagues[184] employed the HGF to investigate the computational mechanisms underlying paranoia in a subclinical population and schizophrenia patients using a non-social reversal learning task. They found increased expected volatility ($\mu_3^{(0)}$) in participants with higher levels of paranoia using the standard 3-level HGF. Our model selection suggested that this model explains behaviour better in HC, whereas FEP were better characterised by a mean-reverting HGF that included a drift at the third level. It should be noted that increasing $\mu_3^{(0)}$

and including a drift at the third level, which increases over time, can both be interpreted as expecting the environment to be more volatile, but the drift provides a more nuanced description of changes that occur *during* the learning session. Our results are thus in line with previous results, but possibly provide a perspective that takes within-task dynamics more explicitly into account. Moreover and in contrast to our results, Reed et al.[184] found increased and not reduced coupling strength κ_2. This discrepancy may be related to differences in the tasks employed (non-social three-option reversal learning task vs our social learning task), but we also note that κ_2 was not always well-recoverable in our simulation analysis. Therefore, we do not wish to draw strong conclusions based on the κ_2 effect our study, although we found trend-level effects suggesting that κ_2 may be related to negative and general symptoms.

1.4.2 Is the perception of environmental volatility altered specifically in social contexts?

Here, we employed an ecologically valid social learning task[56,57] to study changes in learning about other's intentions. Some authors[184,231] have raised the question of whether changes in learning like the ones observed in this study are reflective of a specifically social or rather a domain-general learning deficit. Here, we did not assess whether differences with respect to the perception of environmental volatility were specific to a social context since we did not include a non-social control task. However, this will be an interesting question to address in future studies.

Interestingly, a recent study by Cole and colleagues[37] also identified an HGF with a drift at the third level as the winning model in a sample of CHR participants who were asked to perform a non-social, two-option reversal learning task. Others[184,231] found changes in model parameters related to the perception of environmental volatility in healthy, subclinical, and schizophrenia patient populations. Suthaharan et al.[184] also included a social control task, which did not affect the parameter effects. Therefore, this mechanism may not be specifically tied to social contexts, but instead may be related to a more general deficit in learning under uncertainty.[184,231] However, we do note that the social control task employed by Suthaharan and colleagues[231] was not as ecologically valid as other tasks that were used to study paranoia such as the dictator game[182] or our task which was adapted from empirically-observed human-human interactions in a previous study.[56] Finally, it is also possible that there are both domain-general and domain-specific changes, but that these can only be studied at the neuronal level and converge on the same behavioural model parameters.

1.4.3 What causes an altered perception of volatility?

Interestingly, there may be at least two different pathways that can lead to an altered perception of environmental volatility. First, abnormalities in monoamine systems may lead to aberrant PEs that are unpredictable and lead to the expectation that the environment is very volatile.[54,125] In line with this pathway, Reed et al.[184] found that methamphetamine admin-

istration induced changes in model parameters that impacted learning about environmental volatility in rats. Moreover, Diaconescu et al.[57] found activation in dopaminoceptive regions such as the dopaminergic midbrain during the same social learning task, which was also used in this study. Secondly, external shifts in the volatility of the environment, for example a global health crisis like the COVID-19 pandemic, may also result in an altered perception of volatility and emergence of paranoid thoughts or endorsement of conspiracy theories.[231] This second (environmental) pathway may also be relevant for understanding increased incidence of schizophrenia in individuals that experience migration[211] and those living in urban environments[248] as individuals exposed to both of these risk factors may be confronted with – in some cases drastically – changing environments. In summary, there may be at least two (possibly interacting) pathways that could give rise to an altered perception of environmental volatility.

1.4.4 Clinical implications

We identified trend-correlations between the drift equilibrium point m_3 and PANSS positive symptoms and the frequency of paranoid thoughts and between the coupling strength κ_2 and PANSS negative and general symptoms. While the evidence was not conclusive in this study since these correlations were not significant after multiple testing correction, we note that the effects were in the expected direction, such that perceiving the environment as increasingly volatile (higher m_3) was associated with higher frequency of paranoid thoughts and more severe positive symptoms in general. Future well-powered studies are needed to assess whether these effects can be confirmed in larger samples. Interestingly, we observed heterogeneous model attributions specifically in CHR, whereas the model selection clearly favoured the standard 3-level HGF in HC and the mean-reverting model in FEP. This finding suggests that this model may be helpful to identify CHR patients that will more likely transition to a psychotic disorder.

1.4.5 Limitations

Several limitations of this study merit attention. First, the sample size of this study was small due to very selective inclusion criteria with respect to medication, which, however, enabled us to minimise the impact of medication effects. Larger studies are needed to replicate our results and increase statistical power to identify correlations between model parameters and symptoms. Secondly, we cannot assess the specificity of our results with respect to the social domain since we did not include a non-social control task. Lastly, we also cannot speak to the specificity with respect to other diagnoses, because we did not include a clinical control group, which is an important avenue for future research.

1.4.6 Future directions

While we found evidence for increased uncertainty associated with higher-level beliefs about the volatility of others' intentions, future studies will have to examine whether a compensatory increase in the precision of higher-level beliefs occurs during later stages of schizophrenia, possibly also fluctuating with the severity of psychosis, or whether other models are better suited to capture the conviction associated with delusory beliefs during acute psychotic states (see for example [14,66]). Furthermore, the neural correlates of belief updating in emerging psychosis during social learning should be examined to identify neural pathways that may underlie the changes in perception that were suggested by the model. Lastly, longitudinal studies are needed to assess, whether model parameters can be leveraged as predictors for transition to psychosis or treatment response in individual patients with psychosis.

1.4.7 Conclusions

In conclusion, our results suggest that emerging psychosis is characterised by an altered perception of environmental volatility. Furthermore, we observed heterogeneity in model attributions in individuals at high risk for psychosis suggesting that this computational approach may be useful to stratify the high risk state and for predicting transition to psychosis in clinical high risk populations.

All models are wrong, but some are useful.
Georg E. P. Box (1976)[26]

2 Modelling Reasoning Biases

This chapter will extend the approach outlined in Chapter 1 and focuses on modelling reasoning biases associated with psychotic disorders and delusions. It also assesses the clinical utility of using this computational approach by testing its ability to predict treatment response to a psychotherapeutic intervention. This work was published by Hauke et al. (2022) in *Schizophrenia Bulletin*[54] and adapted for this dissertation.

2.1 Introduction

Delusions occur in various forms: Prominently featuring among others are persecutory delusions – the belief that others deliberately intend to cause harm[82] – as outlined in the previous chapter. However, a plethora of other delusional themes exist, for example grandiose delusions, believing that one has superior power, knowledge or a special identity.[134] While delusions are key symptoms of schizophrenia, they also occur in other disorders with psychotic symptoms, such as delusional disorder, and psychoaffective disorders, including bipolar disorder.[12] It is therefore important to assume a transdiagnostic perspective and understand the mechanisms underlying delusion formation and persistence across psychotic disorders.

A substantial body of work[61,156,187] has examined the relationship between reasoning biases and delusions. For example, patients with psychotic disorders change their beliefs more than HC, when faced with evidence that contradicts their current beliefs (i.e., *disconfirmatory evidence*).[69,97,178,267] Another extensively studied bias is JTC, the tendency to draw conclusions based on limited evidence. JTC was found to be more prevalent in patients with psychotic

disorders, especially in those with delusions.[61,156] Traditionally, JTC was assessed with the beads task, a probabilistic learning task, in which participants are asked to decide from which of two urns an experimenter is drawing a sequence of coloured beads.[112,179] In another version of the task, the fish task,[166,219,263] participants are shown fish that a fisherman caught from one of two lakes with different ratios of coloured fish and are asked to determine from which lake the fisherman was fishing.

While relationships between reasoning biases such as JTC, delusions, and psychotic disorders have been found across different tasks,[61,156,187] as of yet, it is unclear, whether JTC is contributing to delusion formation by increasing premature acceptance of implausible ideas,[61,81] or, whether it is merely an epiphenomenon of psychotic disorders[246] (see[156] for discussion of other biases). A third possibility is that JTC – and increased updating to disconfirmatory evidence – both reflect a noisy and unstable cognitive system that is more vulnerable to affective or habitual biases, thus enabling delusions without directly causing them.[4] Although answering this question may ultimately require longitudinal data, computational modelling allows us to study the computational mechanisms underlying behavioural differences across individuals. Computational models describe how changes in information processing give rise to observable differences in behaviour. This approach is useful, because there is often a many-to-many mapping between computational parameters and behavioural effects. For example, JTC could be caused by greater initial uncertainty, faster belief-updating, or noisier responding. Modelling allows the investigator to distinguish between these possibilities in each individual. This is potentially important, because specific computational mechanisms may relate to specific treatment effects (e.g., blocking dopamine D2 receptors might reduce noisy responding, but not affect belief-updating).

Here, we employed a computational modelling approach to understand the relationship between JTC, psychotic disorders, and delusions. The research objective of this study was to dissect this relationship based on the computational mechanisms underlying belief updating in the fish task. To this end, we formulated three research questions (RQ):

- **RQ1**: *What are the computational mechanisms underlying differences in probabilistic reasoning between HC and patients with psychotic disorders?*
- **RQ2**: *What are the computational mechanisms underlying differences in probabilistic reasoning between individuals with and without JTC?*
- **RQ3**: *What are the computational mechanisms underlying differences in probabilistic reasoning between patients with low and high current delusions?*

While computational analyses may provide relevant theoretical insight, the ultimate goal of understanding computational mechanisms is to improve patients' well-being. To examine the clinical utility of this approach, we investigated whether computational parameters predicted treatment response to Metacognitive Training (MCT),[167] an intervention that specifically targets reasoning biases. Based on previous results,[1] we expected that belief instability

and decision noise would predict treatment response. This hypothesis was also based on the observation that several modules of Metacognitive Training are designed to make cognition more robust. In which case, those with greatest belief instability or decision noise may stand to benefit most from a cognition-focused intervention.

2.2 Methods

2.2.1 Participants

Our sample consisted of $N = 333$ participants of three different studies.[9,10,165] All studies were approved by the local ethics committee and conducted in accordance with the most recent version of the Declaration of Helsinki. Participants provided written informed consent and were reimbursed for clinical assessments. We excluded 13 participants for which raters had indicated miscomprehension of task instructions and three participants due to incomplete probability ratings. The final sample ($N = 317$) consisted of 56 HC and 261 patients with psychotic disorders. We will briefly highlight important differences and similarities across the parent studies below.

Aims

Two of the three parent studies were designed as clinical trials to assess the efficacy of MCT.[10,165] The third study's original aim was to identify neuroimaging correlates that underlie JTC and the effects of MCT.[9]

Recruitment

Clinical trials[10,165] included patients and the third study[9] compared patients and healthy controls. Patients were recruited through the Department of Psychiatry, University Medical Center Hamburg-Eppendorf (UKE) and postings in online psychosis forums. Healthy controls were recruited through postings on university recruitment sites and local media.

In- and exclusion criteria

All studies included patients between 18 and 65 years old, who met criteria for a diagnosis of schizophrenia spectrum disorder confirmed with the Mini-International Neuropsychiatric Interview (M.I.N.I.),[213] and experienced delusions currently or in the past. Exclusion criteria were a history of cranio-cerebral trauma or serious medical or neurological conditions that might affect cognitive performance, and premorbid IQ $\leq$ 70. In group therapy trials, participants with scores on PANSS[126] item P6: *Suspiciousness/Persecution* $>$ 6 or PANSS item P7: *Hostility* $>$ 5 were excluded. Healthy participants with any past or current psychiatric disorder (including substance use disorders), and history of schizophrenia or bipolar disorder in a first degree relative were excluded (for further details, consult:[9,10,165]).

Metacognitive Training intervention

In clinical trials[10,165], MCT was administered as a group training to groups of 4-10 patients at a time. It consisted of eight modules that targeted relevant cognitive biases, such as dysfunctional attributions, JTC, belief inflexibility, deficits in social cognition, overconfidence in errors, and deficits in emotional processing. The goal of this psychotherapeutic intervention is to convey knowledge about cognitive distortions that can occur in psychotic disorders and to raise awareness of the dysfunctionality of such biases through exercises. The exercises are meant to provide corrective experiences and to teach alternative coping and information-processing strategies. MCT can be considered a hybrid therapy with elements from two established cognitive therapy approaches, i.e., cognitive-behavioural therapy (CBT) and cognitive remediation therapy (CRT). As CBT, MCT employs a "back-door approach" since it focuses on cognitive processes first and only then proceeds to core psychiatric symptomatology. Conversely, the presentation of the material is structured analogous to CRT, centering around exercises and error feedback (see[165,168,169] for more details).

Protocols

Both clinical trials[10,165] compared patients that were randomly allocated to either MCT or a control intervention that consisted of a computerised cognitive treatment program (CogPack®; `http://www.markersoftware.com/`) using a fixed, pseudo-randomisation schedule. Additionally, all participants continued treatment as usual. Clinical assessments were conducted by raters that were blind to the treatment allocation. In contrast to Moritz et al. (2013)[165], Andreou et al. (2017)[10] expanded the conventional MCT program (4 weeks) to not only include predominantly non-delusional scenarios, but also additional applications of the learned material to challenge the content of individual delusional beliefs referred to as individualised MCT (6 weeks). In all studies, probabilistic reasoning was assessed at baseline using the fish task,[166,219,263] and clinical symptoms were assessed at baseline and follow-up. Analyses presented here were restricted to the baseline and post-intervention assessment (after 4[165] or 6[10] weeks).

2.2.2 Task

To assess probabilistic reasoning at baseline, we employed a graded estimates version[267] of the fish task.[166,219,263] Participants were instructed that a fisherman was fishing from one – and only one – of two lakes with different ratios of coloured fish (80:20 in lake A and reversed in lake B; Figure 2.1). They were also instructed that these ratios did not change as the fisherman always threw the fish back into the water (sampling with replacement). Participants were presented with a sequence of ten fish. After each fish in the sequence they were asked (1) to estimate the probability that the fish were drawn from lake A (0-100%) and (2) if they were certain enough to decide from which lake the fisherman was fishing and if so, what their conclusion was (i.e., lake A or B).

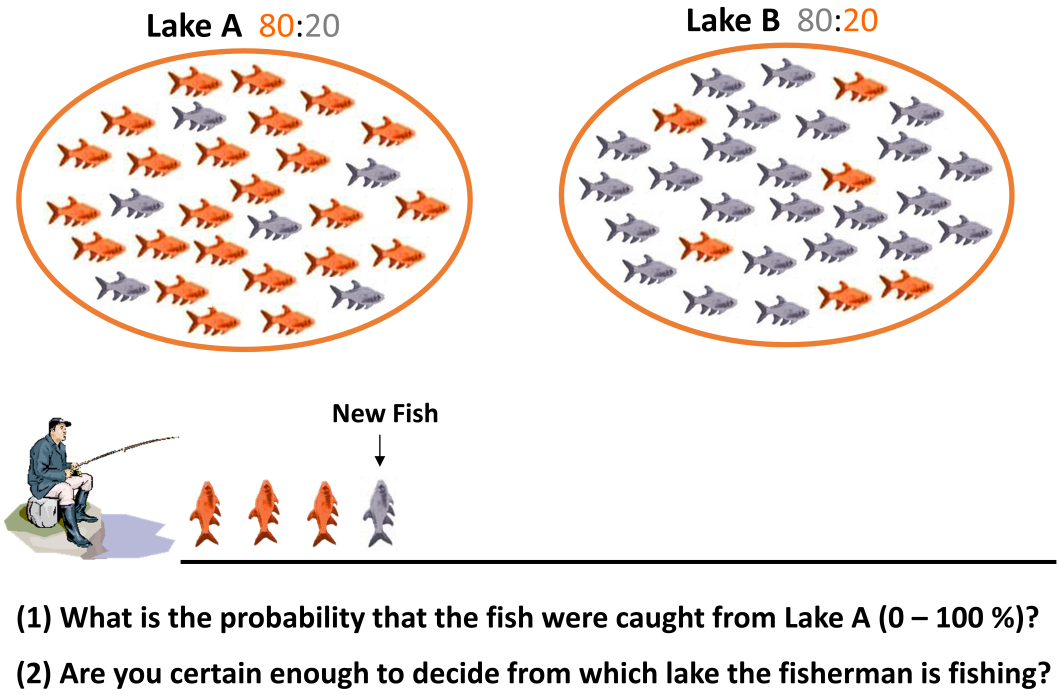

Figure 2.1: Fish task. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

2.2.3 Computational modelling

Hierarchical Gaussian Filter

We modelled behaviour with the HGF.[153,154] This model was employed previously to understand probabilistic reasoning during the beads task in schizophrenia patients[1] and other symptoms of schizophrenia (e.g., hallucinations[181] or paranoid delusions[54,184,231]). The HGF assumes that learning is driven by precision-weighted PE updates and that learners integrate prior and new information in a Bayes-optimal manner given their individual learning parameters, which are estimated from participants' behaviour. These parameters can be understood as encoding an individual's approximation to Bayesian inference[154] and provide a concise summary of individual learning profiles. Differences in model parameters or architectures across participants can then be leveraged to understand the computational mechanisms underlying different populations. We closely followed the approach of Adams and colleagues,[1] briefly summarise below.

We modelled participants' behaviour with a 2-level, nonvolatile version of the HGF (see Figure 2.2A). The third level of the HGF expresses learning about the volatility of the environment. As Adams et al.,[1] who employed this model to model probabilistic reasoning during the beads task in a more homogeneous sample of schizophrenia patients, we decided to employ a 2-level version of this model, because the environment was stable throughout

the experiment (the fisherman was always fishing from the same lake). This also helped to reduce the number of model parameters to be estimated, which was important given the small number of experimental trials that were available per participant.

The generative model assumes that a fish $u^{(k)}$ is drawn from the probability $x_1^{(k)}$ that the fisherman is fishing in one of the lakes (e.g., lake A) on trial k (see Figure 2.1). The state at the level above is the unbounded tendency $(-\infty, +\infty)$ that the fisherman was fishing in lake A $x_2^{(k)}$, which can be transformed to the probability $x_1^{(k)}$ using the sigmoid transformation $s(x_2^{(k)})$, where $s(x_2^{(k)})$ is the unit square sigmoid transformation: $s(z) = \frac{1}{1+e^{-z}}$.

However, $x_2^{(k)}$ is not known to the participant and needs to be inferred when observing a sequence of fish. The participant's posterior estimate of $x_2^{(k)}$ on trial k is denoted $\mu_2^{(k)}$. Again, this unbounded tendency can be transformed into the participant's posterior estimate of the probability of the sequence of fish being fished from that lake with range $[0, 1]$ using a sigmoid transformation as before $\hat{\mu}_1^{(k)} = s(\kappa_1 \mu_2^{(k)})$, which is equivalent to the participant's prediction for the next trial (denoted with ^). However, here, κ_1 represents a subject-specific parameter that captures the degree of belief instability.

Using this model, we assume that participants start with a prior mean for $\mu_2^{(0)} = 0$. After the sigmoid transformation, this corresponds to the expectation that both lakes are equally probable or that the probability for each lake is 50%, i.e. $\hat{\mu}_1^{(0)} = 0.5$. Before participants observe a new input on each trial k, their prediction $\hat{\mu}_2^{(k)}$ and $\hat{\mu}_1^{(k)}$ and precisions thereof (inverse of the variances) $\hat{\pi}_2^{(k)}$ and $\hat{\pi}_1^{(k)}$ are given by the following equations:

$$\hat{\mu}_2^{(k)} \equiv \mu_2^{(k-1)} \tag{2.1}$$

$$\hat{\mu}_1^{(k)} \equiv s(\kappa_1 \hat{\mu}_2^{(k)}) \tag{2.2}$$

$$\hat{\pi}_1^{(k)} \equiv \frac{1}{\hat{\mu}_1^{(k)}(1 - \hat{\mu}_1^{(k)})} \tag{2.3}$$

$$\hat{\pi}_2^{(k)} \equiv \frac{1}{\sigma_2^{(k-1)} + \exp(\omega)}. \tag{2.4}$$

When participants observe a new fish, a PE δ_1, expressing the discrepancy between what participants observe $u^{(k)}$ and what they expected $\hat{\mu}_1^{(k)}$, is generated and used to adjust the participant's expectations:

$$\delta_1^{(k)} \equiv u^{(k)} - \hat{\mu}_1^{(k)}, \tag{2.5}$$

where $u^{(k)}$ is a new input observed at trial k, which can be either 0 (grey fish) or 1 (orange fish). The posterior expectations after observing the new evidence is computed as follows:

$$\pi_2^{(k)} = \hat{\pi}_2^{(k)} + \frac{\kappa_1^2}{\hat{\pi}_1^{(k)}} \tag{2.6}$$

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \frac{\kappa_1}{\pi_2^{(k)}} \delta_1^{(k)}. \tag{2.7}$$

Note, that κ_1 affects the size of the updates, such that increased belief instability κ_1 leads to stronger updates and thus more unstable beliefs over time, when observing new evidence.

The posterior expectation translates into a new prediction for the next trial again using the sigmoid transformation:

$$\hat{\mu}_1^{(k+1)} \equiv s(\kappa_1 \mu_2^{(k)}). \tag{2.8}$$

The response model that maps participants' expectation onto their responses takes the form of a beta distribution described by its mean μ_{resp} and variance ν. These statistics are related to the conventional shape parameters a and b of a beta distribution as follows:

$$\mu_{resp} := \frac{a}{a+b} \tag{2.9}$$

$$\nu := a + b. \tag{2.10}$$

We implemented this model, using the tapas toolbox (version: 4.0.0; `https://github.com/translationalneuromodeling/tapas/releases/tag/v4.0.0`)[79] in Matlab (version: 2017a; `https://mathworks.com`) and the functions 'tapas_hgf_ar1_binary' and 'beta_obs' for the perceptual and response model, respectively.

Bayesian model selection

We formulated two competing hypotheses describing different learning mechanisms, including **Hypothesis I**: *Standard Bayesian belief updating* and **Hypothesis II**: *Bayesian belief updating subject to belief instability* (controlled by parameter κ_1). Note, that equation (2) is reduced to a simple sigmoid transformation of $\mu_2^{(k)}$, if $\kappa_1 = 1$. In this case **Model 2** is reduced to **Model 1**. However, if $\kappa_1 \geq 1$, a simulated participant will show increased belief instability that leads them to quickly change their mind when confronted with disconfirmatory evidence (see Figure 2.2B), but also leads to smaller updates when presented with consistent evidence (e.g., fishes 5-8). Model 2, estimates κ_1 from participants' behaviour and therefore tests the hy-

pothesis that participants' learning can be better described by Bayesian belief updating subject to belief instability.

For both models, we estimated participants' *prior uncertainty* or $\sigma_2^{(0)}$ at the beginning of the experiment, participants' *evolution rate* or ω_2, and the response stochasticity or *decision noise* ν. In Model 2, we additionally estimated *belief instability* κ_1. All other parameters of the HGF were fixed (see Table 2.1 for overview over model priors).

We compared these competing hypotheses with random-effects Bayesian model selection [186,225] and computed protected exceedance probabilities for the two models. Protected exceedance probabilities measure the probability that a model is more likely than any other model in the model space, [222] protected against the risk that differences between models arise due to chance alone. [186] We also computed relative model frequencies as a measure of effect size, which can be understood as the probability that a randomly sampled participant would be best explained by this model. The model selection was implemented using the VBA toolbox [46] (`https://mbb-team.github.io/VBA-toolbox/`).

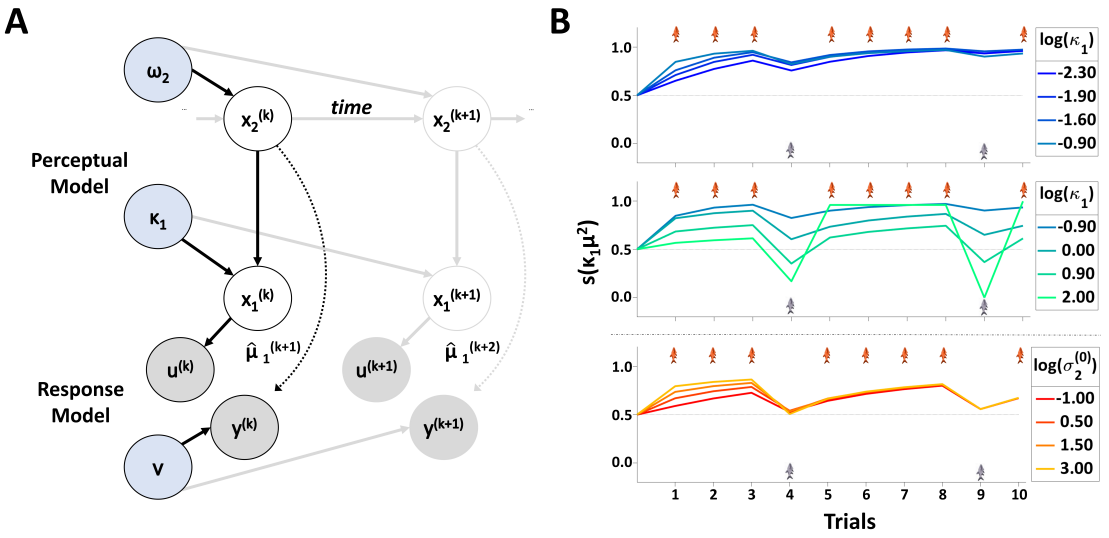

Figure 2.2: Winning model. **A** Graphical representation of the generative model adapted from Adams et al. [1] Observed quantities are denoted with grey circles. White circles represent hidden states and blue circles subject-specific parameters. Black lines indicate probabilistic network at trial k and grey lines at trial $k+1$. Solid lines indicate generative model in the world, which participants infer on, [48,47] whereas dotted lines represent participants' inference on these states. **B** Simulation showing the impact of changing belief instability κ_1 and prior uncertainty $\sigma_2^{(0)}$. Displayed is the inferred probability that the fisherman is fishing from lake A $s(\kappa_1\mu_2^{(k)})$ for very low (**upper panel**) or low to high levels of belief instability (**middle panel**), and changing prior uncertainty (**lower panel**). All other parameters were fixed to the posterior medians. Increasing $\log(\kappa_1)$ above approximately -0.9 leads to higher belief instability, as simulated agents are changing their beliefs more rapidly when faced with disconfirmatory evidence. Increasing κ_1 in the very low range leads to larger belief updates early in the experiment. Note, however, that the exact value of κ_1 at which the model's behaviour undergoes this qualitative change depends on the other parameter values. Increasing $\sigma_2^{(0)}$ consistently leads to larger belief updates early in the experiment. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*. [105]

Model recovery

To assess whether models were recoverable, we conducted a series of simulations. We simulated 20 synthetic datasets based on the empirical parameter estimates obtained from fitting all models to the empirical data of every participant. The sample size of each synthetic dataset was chosen to be equivalent to the empirical sample size ($N = 317$). The noise level was set based on the empirically estimated decision noise ν_{est}. Each simulation was initialised using different random seeds to account for the stochasticity of the simulation. This led to a total of 2 (models) x 317 (participants) x 20 (simulation seeds) = 12,680 simulations. Subsequently, we re-inverted each of the proposed models on the synthetic data to determine, whether we could recover the true model under which synthetic data was generated. To assess model recovery, we then performed random-effects Bayesian model selection[186,225] on each of the datasets with a sample size of $N = 317$ as in the empirical data and averaged the resulting protected exceedance probabilities across the 20 simulation seeds to obtain a model confusion matrix.

Parameter recovery

To determine whether model parameter estimates were reliable, we also performed a parameter recovery analysis. Using the simulation and model inversion results from the model recovery analysis (see preceding section), we assessed how accurately the parameters generating the data ('simulated') corresponded to the parameters that were estimated when re-inverting the same model on that data ('recovered'). We report Pearson correlations and their associated p-values to quantify our ability to recover the model parameters. Since, the significance of these correlations is influenced by sample size, we also computed Cohen's f^2, where an $f^2 \geq 0.35$ can be considered a large effect size[35] and was interpreted as evidence for good parameter recovery.

2.2.4 Statistical analyses

We tested the three research questions with linear mixed effects models with individual probability estimates as dependent variable. Each model was comprised of a random intercept per participant, a fixed effect of trial, a fixed group effect for either diagnosis (RQ1), presence of

	Belief instability	**Learning rate**	**Prior uncertainty**	**Decision noise**
Hypothesis I		$\omega_2(-2, 16)$	$\sigma_2^{(0)}(log(0.8), 0.5)$	$\nu(log(128), 1)$
Hypothesis II	$\kappa_1(log(1), 1)$	$\omega_2(-2, 16)$	$\sigma_2^{(0)}(log(0.8), 0.5)$	$\nu(log(128), 1)$

Table 2.1: Priors on free model parameters. Prior means and their respective variances are denoted in brackets: (Mean, Variance). Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

JTC (RQ2), or delusions (RQ3) as well as a trial-by-group interaction as predictors of interest and education, medication, sex, and study as covariates of no interest. JTC was defined as reaching a decision after $\leq$ 2 fish as done previously.[29] Low and high current delusions were defined based on a median split of the Psychotic Symptom Rating Scales (PSYRATS): Delusion subscale.[103] Four patients were excluded from the delusion analysis (RQ3), due to incomplete PSYRATS data. We also performed a supplementary analysis for RQ3 to examine the effect of absence or presence of current clinically-relevant delusions, which was defined as PANSS[126] item P1: *Delusions* $\geq$ 3. This analyses was motivated by comparing patients that either experienced or did not experience *any* clinically relevant delusions at baseline, as ratings of 1 indicate the absence of symptoms, and ratings of 2 indicate extremes in the healthy population or only suspected symptoms, while ratings of 3 or higher are indicative of presence of clinically relevant delusions (regardless of the delusional theme). In all analyses, missing education values ($n = 3$) were imputed using group-wise median imputation. As the distributions of probabilistic judgements across participants were heavily skewed, we also conducted pairwise comparisons of trial-by-trial behaviour across groups with non-parametric Kruskal-Wallis tests. Similarly, model parameters were compared using Kruskal-Wallis tests. We corrected for multiple testing using Bonferroni correction for the number of trials ($n = 10$) or the number of parameters ($n = 4$) in the behavioural and parameter analyses, respectively. Please, note, that Bonferroni-correction is likely to be too conservative as responses were correlated across trials. Statistical analyses were conducted in R (version: 4.04; `https://www.r-project.org/`) using R-Studio (version: 1.4.1106; `https://www.rstudio.com/`).

2.2.5 IQ-matched analysis

We repeated the behavioural and parameter analyses on an IQ-matched subsample of patients to assess whether group differences could be explained by differences in IQ. Premorbid IQ was assessed in $N = 229$ of the participants using either the 'Mehrfachwahlwortschatztest'[142] (used in[165]) or the 'Wortschatztest'[202] (used in[9,10]). A non-parametric Kruskal-Wallis tests indicated that patients with psychotic disorders (mean: 104.19, median: 104) showed significantly lower premorbid IQ than HC (mean: 109.12, median: 107; $\eta^2 = 0.024$, $p_{uncorr} = 0.005$). To obtain IQ-matched subsamples, we excluded patients whose IQ fell below the range observed in HC (i.e., $\leq$ 85). Additionally, we only included a random subset of patients with IQs ranging from 90 to 95 to remove skewness of the IQ distribution in patients, that was not observed in HC resulting in a subsample of $N = 202$. After this matching, patients with psychotic disorders ($n = 146$, mean: 107.58, median: 107) and HC ($n = 56$, mean: 109.12, median: 107) did no longer significantly differ with respect to IQ ($\eta^2 = 0.005$, $p_{uncorr} = 0.321$). Similarly, IQ between individuals with (mean: 103.28, median: 101) and without JTC (mean: 106.98, median: 107) significantly differed before the matching ($\eta^2 = 0.017$, $p_{uncorr} = 0.022$), but not afterwards ($n = 82$ with JTC: mean: 107.13, median: 107; $n = 120$ without JTC: mean: 108.60, median: 107; $\eta^2 = 0.003$, $p_{uncorr} = 0.301$). Since the delusion subgroups included only patients and did not differ in terms of IQ ($n = 83$

patients with low current delusions: mean: 104.81, median: 104; $n = 96$ patients with high current delusions: mean: 103.48, median: 101; $\eta^2 = 0.001$, $p_{uncorr} = 0.628$), we only report IQ-matched analysis for RQ1 and RQ2.

2.2.6 Treatment response prediction

We compared three prognostic models that were trained on three different feature sets. The first model was trained to predict treatment response from the model-derived computational fingerprint of the participants, i.e., the four model parameters of the winning model (κ_1, $\sigma_2^{(0)}$, ω, and ν). The second (behavioural) model was trained on participants raw behavioural data (i.e., probability estimates and decisions) as well as a binary variable indicating presence of JTC (reaching a decision after seeing $\leq$ 2 fish). The third (clinical) model was trained on clinical baseline data (all individual PANSS items measured at baseline). We trained random forest classifiers[27] to predict treatment response from the three feature sets. Our preprocessing pipeline consisted of (1) imputing missing values using median imputation, (2) covariate correction for sex, medication and education (see for example[136]). Preprocessing steps were embedded in a stratified, repeated k-fold cross-validation with 10 folds and 10 permutations to prevent information leakage into the test data. We used default parameter settings for the random forest algorithm without further parameter optimization (i.e., 500 trees, for more details, please, consult the package documentation). This classification pipeline was implemented in R (version: 4.04; `https://www.r-project.org/`) using the mlr package (version: 2.19.0; `https://mlr.mlr-org.com/`) and the randomForest package (version: 4.6-14; `https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest`). To assess model performances, we computed balanced accuracy (BAC), area under the curve (AUC), sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV), where responders were defined as the positive class. To test whether classifiers' performances were significantly greater than chance, permutation tests with 1000 label permutations were employed. Lastly, we report feature importance using the decrease in accuracy averaged across cross-validation folds. Higher values indicate greater performance degradation of the algorithm when removing a given feature and thus imply that this feature is more important.

2.3 Results

Clinical and demographic characteristics are reported in Table 2.2. Since, there was no conclusive evidence for increased JTC in patients with psychotic disorders ($\chi^2 = 3.435$, $p_{uncorr} = 0.064$), we analysed HC and patients together in all subsequent analyses investigating JTC.

2.3.1 Behavioural results

RQ1: Group differences between HC and patients with psychotic disorders

We found a significant group-by-trial interaction when comparing HC and patients with psychotic disorders ($F = 4.420, p < 0.001$; Figure 2.3), which held in IQ-matched subsamples (Figure 2.6A). This effect was driven by a stronger decrease in probability for the more likely lake A in patients with psychotic disorders in trial 9 with one of the two rare (i.e., disconfirmatory) fish ($\eta^2 = 0.028, p = 0.031$). We also observed trend-effects in trials 4 and 8, which did not survive Bonferroni correction, however ($\eta^2 = 0.015, p_{uncorr} = 0.028, p = 0.281$ and $\eta^2 = 0.020, p_{uncorr} = 0.012, p = 0.123$, respectively). None of the covariates was significant.

RQ2: Group differences between individuals with and without JTC

Comparing individuals with and without JTC, we found a significant JTC-by-trial interaction ($F = 11.598, p < 0.001$; Figure 2.4), which held when comparing IQ-matched subsamples (Figure 2.6B). This effect was driven by increased probability estimates for lake A in individuals with JTC within the first three trials of the fish sequence (trial 1: $\eta^2 = 0.043$, $p = 0.002$, trial 2: $\eta^2 = 0.077, p < 0.001$, and trial 3: $\eta^2 = 0.056, p < 0.001$). Furthermore, we observed a significant main effect of medication as medicated individuals estimated lower probabilities overall ($F = 7.138, p = 0.008$), but none of the other covariates.

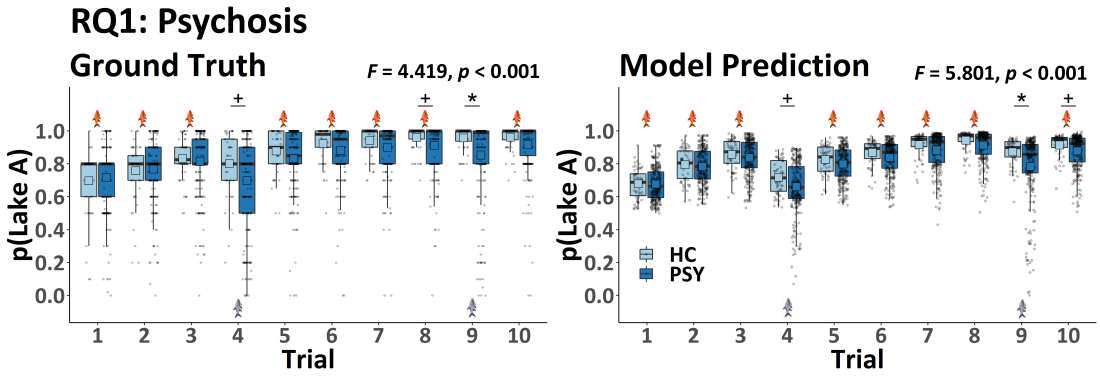

Figure 2.3: Behavioural effects versus model predictions for RQ1: Psychosis. Comparing behaviour in the fish-task between healthy controls (**HC**) and patients with psychotic disorder (**PSY**). F- and p-values indicate results of ANCOVAs corrected for education, medication, sex, and study. **Y-axis**: Participants' estimates of the probability that the fisherman was fishing from lake A (see question (1) in Figure 2.1). **Left panels**: Behavioural effects. **Right panels**: Model prediction of the winning model. **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: $^{*}p < 0.05$, using Bonferroni correction, or at $^{+}p < 0.05$ uncorrected. Note, that Bonferroni correction is likely to be too conservative as responses were correlated across trials. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

RQ3: Group differences between patients with low and high current delusions

There was no evidence for differences in probabilistic reasoning between patients with low or high current delusions (delusion-by-trial interaction: $F = 0.503, p = 0.873$; Figure 2.5A). We observed a trend-effect of delusions in trial 10 that did not survive Bonferroni correction ($\eta^2 = 0.017, p_{uncorr} = 0.019, p = 0.194$). Among the covariates, we only found a significant main effect of education suggesting that longer education was associated with higher probability estimates overall ($F = 4.016, p = 0.046$).

This result remained unchanged when investigating the alternative definition of delusions (i.e., PANSS P1: *Delusions* $\geq$ 3). There was no significant difference between patients with and without *any* current delusions (delusion-by-trial interaction: $F = 0.712, p = 0.698$; Figure 2.5B). However, there was a significant main effect of education as before ($F = 4.463$, $p = 0.036$) and a trend-level effect of study ($F = 2.429, p = 0.090$).

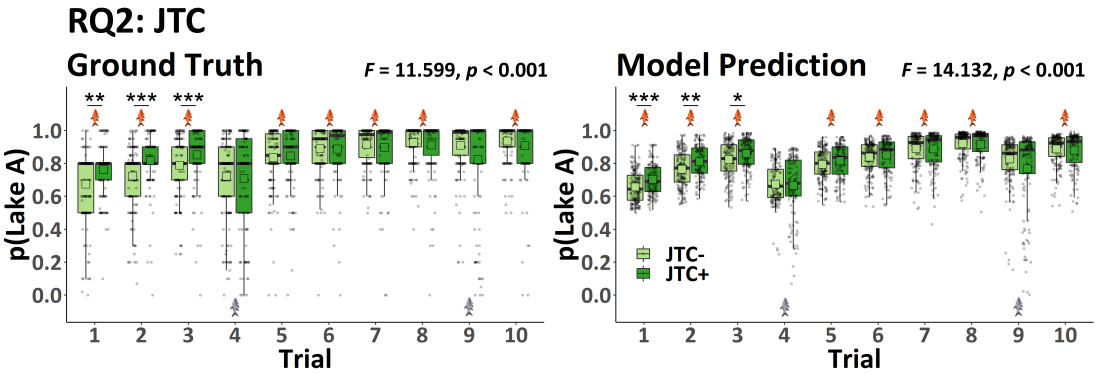

Figure 2.4: Behavioural effects versus model predictions for RQ2: JTC. Comparing behaviour between individuals without (**JTC-**) and with (**JTC+**) jumping-to-conclusion bias (decision after $\leq$ 2 fish). F- and p-values indicate results of ANCOVAs corrected for education, medication, sex, and study. **Y-axis**: Participants' estimates of the probability that the fisherman was fishing from lake A (see question (1) in Figure 2.1). **Left panels**: Behavioural effects. **Right panels**: Model prediction of the winning model. **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within 1.5$\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$, using Bonferroni correction. Note, that Bonferroni correction is likely to be too conservative as responses were correlated across trials. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

2.3.2 Modelling results

Bayesian model selection and model recovery

Model selection strongly suggested that **Hypothesis II**: *Bayesian belief updating subject to belief instability* was the most likely mechanism explaining behaviour across all groups ($\varphi = 100.00\%, f = 98.94\%$; Figure 2.7). The model recovery analysis indicated that the two hypotheses were distinguishable.

	Research Question 1 Psychosis			Research Question 2 Jumping-to-Conclusions			Research Question 3 Delusions			Treatment Response to Metacognitive Training		
	HC $n = 56$	**PSY** $n = 261$	Statistic	**JTC-** $n = 174$	**JTC+** $n = 143$	Statistic	**D-** $n = 131$	**D+** $n = 126$	Statistic	**R-** $n = 63$	**R+** $n = 43$	Statistic
Age$_n$ median $[25^{th}, 75^{th}]$	29_{56} [27, 38]	33_{261} [26, 45]	$\eta^2 = 0.003$ $p = 0.306$	32_{174} [27, 43]	34_{143} [26, 44]	$\eta^2 = 0.001$ $p = 0.500$	33_{131} [26, 43]	33_{126} [27, 45]	$\eta^2 = 0.002$ $p = 0.427$	35_{63} [30, 45]	36_{43} [30, 43]	$\eta^2 = 0.001$ $p = 0.750$
Sex f/m	26/30	104/157	$\chi^2 = 0.826$ $p = 0.364$	75/99	55/88	$\chi^2 = 0.699$ $p = 0.403$	49/82	52/74	$\chi^2 = 0.402$ $p = 0.526$	31/32	19/24	$\chi^2 = 0.258$ $p = 0.611$
Education$_n$ median $[25^{th}, 75^{th}]$	13_{56} [10, 13]	12_{258} [10, 13]	$\eta^2 = 0.005$ $p = 0.200$	13_{174} [10, 13]	12_{140} [10, 13]	$\eta^2 = 0.015$ **p = 0.047**	13_{129} [10, 13]	12_{125} [10, 13]	$\eta^2 = 0.010$ $p = 0.120$	13_{62} [10, 13]	12_{43} [10, 13]	$\eta^2 = 0.013$ $p = 0.252$
IQ$_n$ median $[25^{th}, 75^{th}]$	107_{56} [101, 118]	104_{181} [94, 112]	$\eta^2 = 0.033$ **p = 0.005**	107_{133} [100, 118]	101_{104} [94, 112]	$\eta^2 = 0.022$ **p = 0.022**	104_{83} [95, 114]	101_{96} [93, 112]	$\eta^2 = 0.001$ $p = 0.628$	107_{45} [101, 118]	103_{27} [96, 112]	$\eta^2 = 0.016$ $p = 0.280$
Draws-to-decision$_n$ median $[25^{th}, 75^{th}]$	3_{56} [2, 5]	3_{181} [1, 5]	$\eta^2 = 0.007$ $p = 0.138$	5_{174} [3, 6]	1_{143} [1, 2]	$\eta^2 = 0.772$ **p < 0.001**	2_{131} [1, 5]	3_{126} [2, 5]	$\eta^2 = 0.010$ $p = 0.109$	3_{63} [2, 5]	2_{43} [1, 4]	$\eta^2 = 0.013$ $p = 0.249$
Medication n/y	56/0	24/219	$\chi^2 = 188.630$ **p < 0.001**	54/110	16/109	$\chi^2 = 7.058$ **p = 0.008**	6/118	17/98	$\chi^2 = 6.783$ **p = 0.009**	5/54	1/41	$\chi^2 = 1.630$ $p = 0.202$
PANSS Positive$_n$ median $[25^{th}, 75^{th}]$		14_{260} [9, 19]		14_{136} [10, 19]	14_{124} [9, 19]	$\eta^2 < 0.001$ $p = 0.802$	10_{131} [7, 13]	19_{125} [15, 22]	$\eta^2 = 0.507$ **p < 0.001**	11_{63} [8, 16]	15_{43} [13, 19]	$\eta^2 = 0.104$ **p < 0.001**
PANSS Negative$_n$ median $[25^{th}, 75^{th}]$		12_{261} [9, 16]		11_{137} [9, 15]	13_{124} [9, 17]	$\eta^2 = 0.010$ $p = 0.111$	10_{131} [8, 16]	13_{126} [10, 18]	$\eta^2 = 0.053$ **p < 0.001**	11_{63} [8, 17]	14_{43} [11, 19]	$\eta^2 = 0.061$ **p = 0.012**
PANSS Excitement$_n$ median $[25^{th}, 75^{th}]$		11_{261} [9, 13]		10_{137} [9, 13]	11_{124} [9, 15]	$\eta^2 = 0.013$ $p = 0.067$	10_{131} [8, 12]	12_{126} [10, 15]	$\eta^2 = 0.092$ **p < 0.001**	10_{63} [9, 11]	12_{43} [10, 15]	$\eta^2 = 0.044$ **p = 0.031**
PANSS Distress$_n$ median $[25^{th}, 75^{th}]$		16_{261} [12, 20]		15_{137} [12, 19]	16_{124} [12, 21]	$\eta^2 = 0.005$ $p = 0.250$	12_{131} [10, 16]	18_{126} [15, 22]	$\eta^2 = 0.227$ **p < 0.001**	14_{63} [11, 17]	18_{43} [13, 23]	$\eta^2 = 0.113$ **p < 0.001**
PANSS Disorganisation$_n$ median $[25^{th}, 75^{th}]$		14_{261} [12, 18]		14_{137} [11, 17]	14_{124} [12, 19]	$\eta^2 = 0.009$ $p = 0.135$	12_{131} [11, 15]	16_{126} [14, 20]	$\eta^2 = 0.200$ **p < 0.001**	13_{63} [11, 17]	14_{43} [12, 18]	$\eta^2 = 0.015$ $p = 0.212$
PSYRATS Delusions$_n$ median $[25^{th}, 75^{th}]$		8_{257} [0, 14]		9_{136} [0, 14]	5_{121} [0, 14]	$\eta^2 = 0.005$ $p = 0.270$	0_{131} [0, 1]	14_{126} [11, 17]	$\eta^2 = 0.787$ **p < 0.001**	2_{61} [0, 11]	0_{43} [0, 12]	$\eta^2 = 0.019$ $p = 0.164$
PSYRATS Hallucinations$_n$ median $[25^{th}, 75^{th}]$		0_{257} [0, 4]		0_{136} [0, 6]	0_{121} [0, 4]	$\eta^2 < 0.001$ $p = 0.739$	0_{128} [0, 0]	0_{125} [0, 21]	$\eta^2 = 0.799$ **p < 0.001**	0_{61} [0, 0]	0_{43} [0, 12]	$\eta^2 = 0.039$ **p = 0.045**

Table 2.2: Sociodemographic and clinical characteristics. Reported are uncorrected p-values and test statistics of either χ^2-tests for categorical variables or Kruskal-Wallis tests for all other variables for healthy controls (**HC**) or patients with psychotic disorders (**PSY**), participants without (**JTC+**) or with (**JTC+**) jumping-to-conclusion bias (decision after ≤ 2 fish), patients with low (**D-**) or high (**D+**) current delusions (split half based on median of Psychotic Symptom Rating Scales[103] (**PSYRATS**): Delusion subscale), and patients without (**R-**) or with (**R+**) treatment response to Metacognitive Training[167] defined as at least 20% decrease in the Positive and Negative Syndrome Scale[126] (**PANSS**) positive factor[244] compared to baseline. Bold print highlights p-values significant at: $p < 0.05$, uncorrected. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

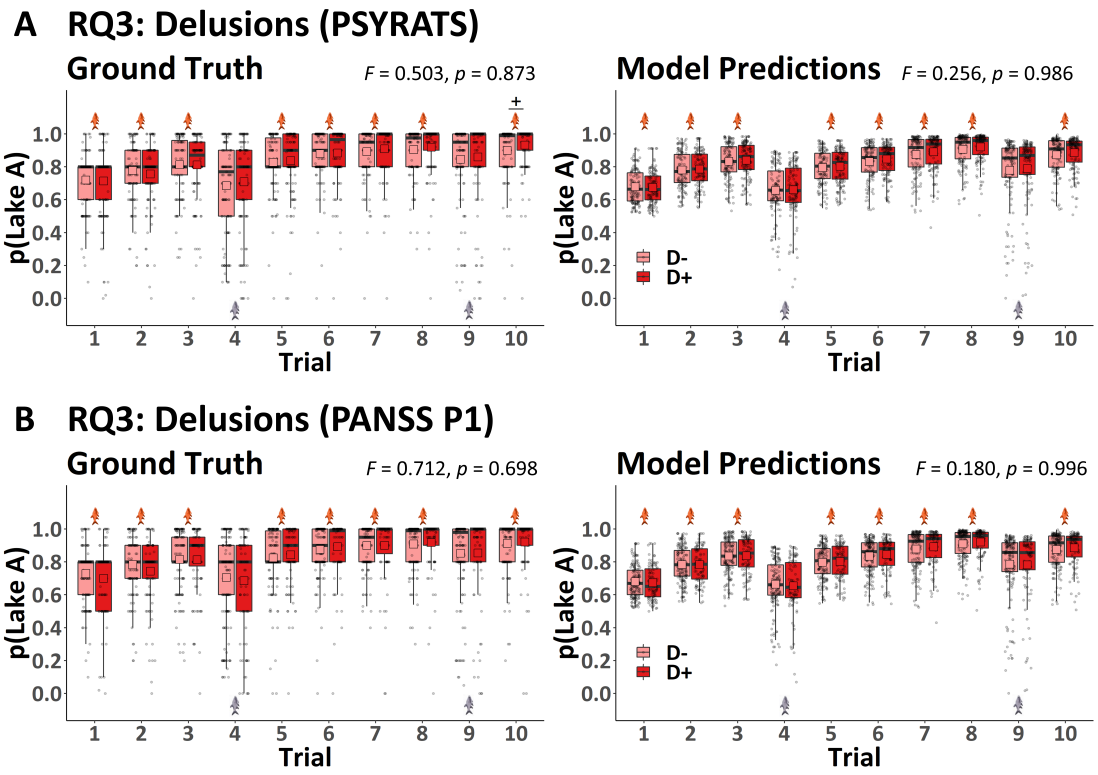

Figure 2.5: Behavioural effects versus model predictions for RQ3: Delusions. **A** Comparing behaviour between patients with low (**D-**) and high (**D+**) current delusions (split half based on median of Psychotic Symptom Rating Scales [103] (PSYRATS): Delusion subscale). **B** Behavioural effects based on an alternative definition of delusions: Comparing behaviour between patients without (**D-**) and with (**D+**) *any* current clinically relevant delusions, i.e., Positive and Negative Syndrome Scale [126] (**PANSS**) item P1: *Delusions* ≥ 3. F- and p-values indicate results of ANCOVAs corrected for education, medication, sex, and study. **Y-axis**: Participants' estimates of the probability that the fisherman was fishing from lake A (see question (1) in Figure 2.1). **Left panels**: Behavioural effects. **Right panels**: Model prediction of the winning model. **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: $^{+}p < 0.05$ uncorrected. Note, that Bonferroni correction is likely to be too conservative as responses were correlated across trials. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*. [105]

Posterior predictive checks and parameter recovery

To confirm that the winning model captured the behavioural effects of interest, we conducted posterior predictive checks. Repeating the behavioural analysis on the winning model's predictions confirmed that this model recapitulated the interaction effects observed in patients with psychotic disorders (RQ1) and individuals with JTC (RQ2), as well as the absence of a delusion effect (RQ3) in accordance with the behavioural analysis (see Figures 2.3, 2.4, and 2.5). Our parameter recovery analysis indicated good recovery for all parameters in all simulations (i.e., Cohen's $f^2 > 0.35$; Figure 2.7D). Next, we tested for group differences in model

parameters.

2.3.3 Parameter group effects

RQ1: Parameter effects between HC and patients with psychotic disorders

First, we found that patients were characterised by significantly larger belief instability κ_1 compared to HC ($\eta^2 = 0.033, p = 0.005$; Figure 2.8A and Table 2.3), which was reproduced in IQ-matched subsamples (Figure 2.9A). Increased belief instability κ_1 likely explained the increased updating in response to disconfirmatory evidence that was observed behaviourally (Figure 2.2B). None of the other parameters showed a significant effect.

RQ2: Parameter effects between individuals with and without JTC

Second, individuals with JTC displayed significantly larger belief instability κ_1 ($\eta^2 = 0.038$, $p = 0.002$; Figure 2.8B and Table 2.3), but also increased prior uncertainty $\sigma_2^{(0)}$ ($\eta^2 = 0.0208, p < 0.050$ ($p = 0.0499$); Figure 2.8C), which likely accounted for the initial increase in belief updating found in individuals with JTC (Figure 2.2B). Both effects remained significant, when comparing IQ-matched subsamples (Figure 2.9B and C). None of the other parameters significantly differed across JTC groups.

RQ3: Parameter effects between patients with low and high current delusions

Lastly, we found no significant effect of current delusions on any model parameters (Table 2.3). Based on the alternative definition of delusions, we identified a trend-effect of increased decision noise ν in patients with *any* current clinically-relevant delusions (i.e., PANSS P1: *Delusions* ≥ 3; $\eta^2 = 0.0126, p_{uncorr} = 0.046, p = 0.186$; Figure 2.8D). To assess the relationship with other symptoms, we computed Kendall rank correlations between all four model parameters and the five PANSS factors,[244] or the PSYRATS[103] delusion and hallucination subscales. We found only a trend-effect suggesting that increased decision noise ν was associated with higher PSYRATS hallucination scores ($\tau = -0.114, p_{uncorr} = 0.016, p = 0.130$).

	Research Question 1 Psychosis				Research Question 3 Jumping-to-conclusions				Research Question 3 Delusions			
Parameter	η^2	p_{uncorr}	p_{fdr}	p_{bf}	η^2	p_{uncorr}	p_{fdr}	p_{bf}	η^2	p_{uncorr}	p_{fdr}	p_{bf}
belief instability κ_1	0.033	**0.001**	**0.005**	**0.005**	0.038	**<0.001**	**0.002**	**0.002**	<0.001	0.816	0.816	1.000
evolution rate ω_2	<0.001	0.882	0.882	1.000	0.006	0.174	0.232	0.697	0.007	0.128	0.256	0.512
prior uncertainty $\sigma_2^{(0)}$	0.010	0.073	0.145	0.291	0.020	**0.012**	**0.025**	**<0.050**[a]	0.009	0.097	0.256	0.388
decision noise ν	0.005	0.189	0.252	0.755	0.004	0.284	0.284	1.000	0.005	0.216	0.288	0.863

Table 2.3: Overview of parameter effects. Displayed are results of Krushkal-Wallis tests to test for group differences in model parameters. We report p-values adjusted for multiple comparisons across the four model parameters using FDR correction (p_{fdr}), Bonferroni correction (p_{bf}), as well as uncorrected p-values (p_{uncorr}). [a] $p_{bf} = 0.0499$.

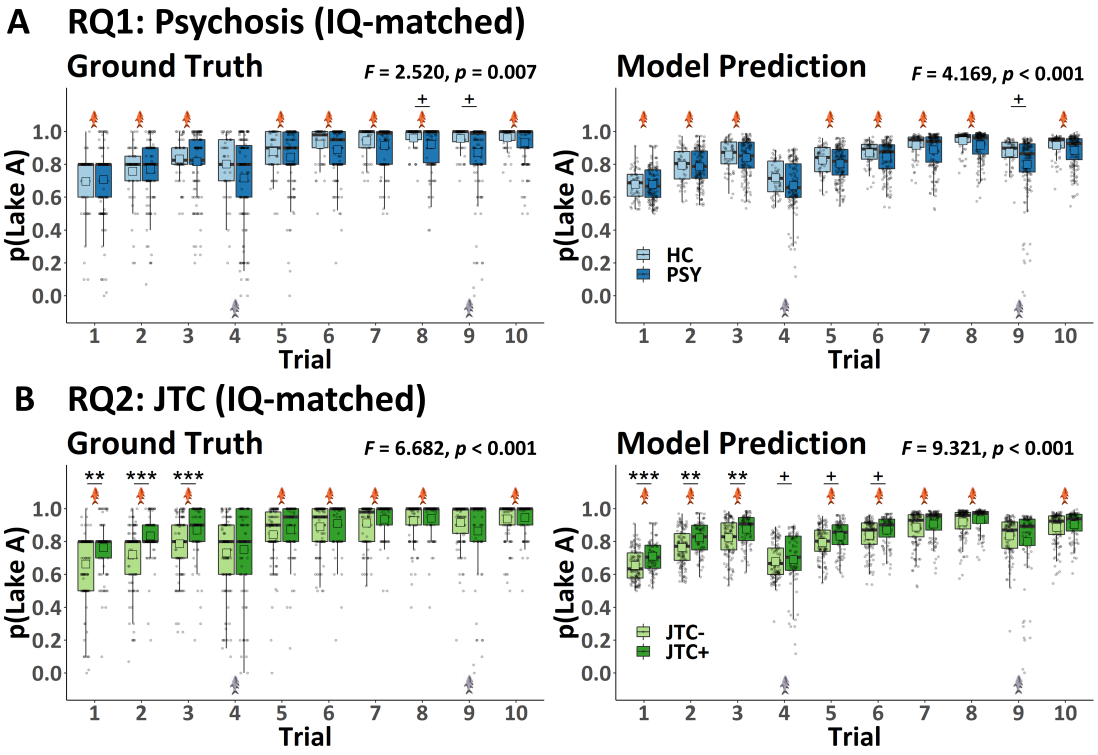

Figure 2.6: Behavioural effects versus model predictions for IQ-matched subsamples. We repeated the behavioural analysis with an IQ-matched subsample of patients. **A** Behavioural effects for RQ1: Psychosis. Comparing behaviour in the fish-task between healthy controls (**HC**) and patients with psychotic disorder (**PSY**). **B** Behavioural effects for RQ2: JTC. Comparing behaviour between individuals without (**JTC-**) and with (**JTC+**) jumping-to-conclusion bias (decision after seeing $\leq$ 2 fish). *F*- and *p*-values indicate results of ANCOVAs corrected for education, medication, gender, and study. **Y-axis**: Participants' estimates of the probability that the fisherman was fishing from lake A (see question (1) in Figure 2.1). **Left panels**: Behavioural effects. **Right panels**: Model prediction of the winning model. **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: *** $p < 0.001$, and ** $p < 0.01$, using Bonferroni correction, or at $^{+}$ $p < 0.05$ uncorrected. Note, that Bonferroni correction is likely to be too conservative as responses were correlated across trials. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

2.3.4 Treatment response prediction

Increased belief instability κ_1 was significantly associated with better treatment response at the group level ($\eta^2 = 0.074$, $p = 0.021$, Figure 2.8E). Subsequently, we also investigated whether treatment response could be predicted at the individual level.

The classifier trained on model parameters predicted treatment response with 64% BAC, which was significantly greater than chance, indicated by a permutation test ($p = 0.001$, Figure 2.8F; AUC: 0.67, SE: 0.53, SP: 0.76, PPV: 0.60, NPV: 0.71). This model's performance was mainly driven by belief instability κ_1, followed by decision noise ν (Figure 2.8G).

To evaluate whether the modelling step was necessary for this performance, we also trained a classifier directly on the raw behavioural data. This model could not predict treatment response above chance (BAC: 0.55, $p = 0.127$, Figure 2.8F; AUC: 0.63, SE: 0.38, SP: 0.71, PPV: 0.49, NPV: 0.63).

Lastly, to investigate whether treatment response could be equally well or even better predicted using clinical measures that are more readily available in clinical practice, we trained the third model on clinical baseline information. Despite differences in symptom expression at baseline, this model did not predict treatment response above chance (BAC: 0.55, $p = 0.139$, Figure 2.8F; AUC: 0.58, SE: 0.41, SP: 0.68, PPV: 0.47, NPV: 0.63) suggesting that the model-based analysis indeed uncovered additional clinically-relevant information.

2.4 Discussion

We employed a computational modelling approach to understand belief updating dynamics during the fish task and their relationship with psychotic disorder diagnosis (RQ1), JTC (RQ2), and current delusions (RQ3). Comparing two competing mechanisms, we found that *belief updating subject to belief instability* best explained participants' behaviour in our study. This model was well-recoverable and could reproduce differences in probabilistic reasoning associated with psychotic disorders and a propensity to jump to conclusions. Analysing parameters of the winning model, we obtained two major results: First, we found that probabilistic reasoning in patients with psychotic disorders was explained by the model through increased belief instability. Second, our results suggest that belief instability differentiated patients who responded from those who did not respond to a Metacognitive Training intervention, both at the group level and the individual level.

2.4.1 Learning mechanisms underlying psychotic disorders and jumping-to-conclusions

Despite analysing a different task in a more heterogeneous patient population, we replicated previous findings by Adams et al.,[1] which suggested that abnormal belief updating in patients with schizophrenia performing the beads task may be explained by increased belief instability κ_1. Our results also offer a possible explanation for JTC as a general cognitive trait across HC and patients as we found an increase in prior uncertainty $\sigma_2^{(0)}$ associated with JTC that explained this effect. Importantly, both associations held in a subsample, which was matched for IQ (Figure 2.9) and were not accounted for by differences in education, or medication. Additionally, we found a significant increase in belief instability in participants with JTC, which remains challenging to interpret. Based on simulations (Figure 2.2B), the most likely explanation is that this increase in belief instability explained differences in belief updating when participants were faced with disconfirmatory evidence (fish 9) that the behavioural analysis did not identify due to a lack of power. However, we cannot rule out that κ_1 also partially

explained increased initial updating for those participants, where the parameter assumed very low values.

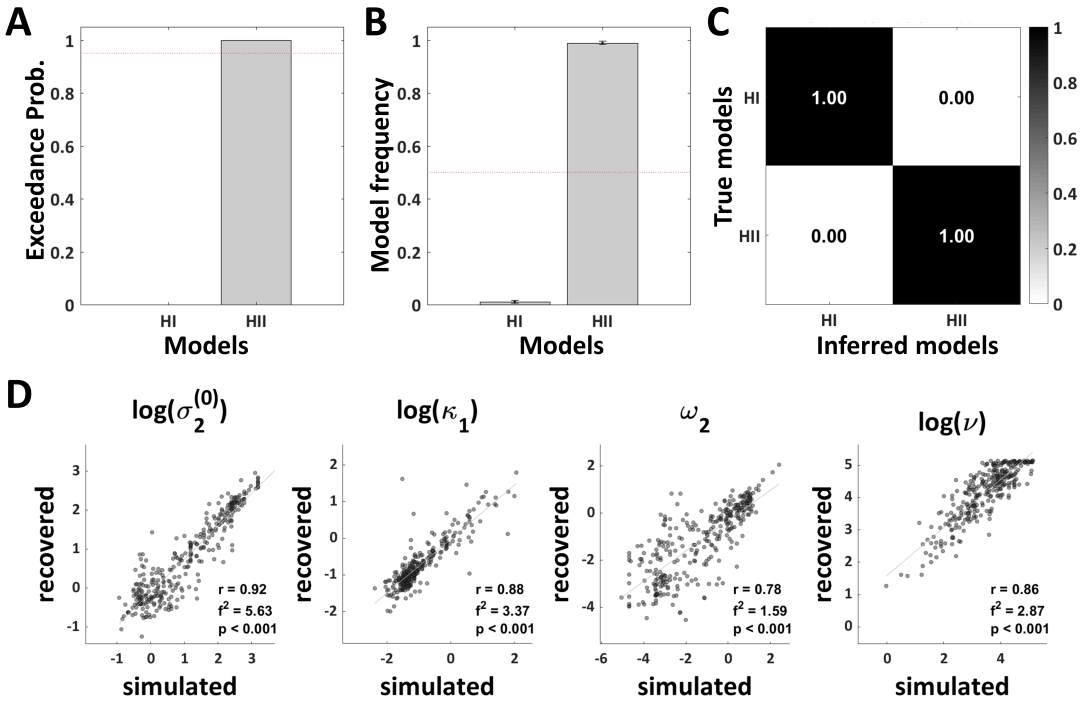

Figure 2.7: Bayesian model selection and recovery analyses. Results of random-effects Bayesian model selection.[225,186] **A**: Protected exceedance probabilities. The dashed line indicates 95% exceedance probability. **B**: Expected model frequencies as a measure of effect size. The dashed line indicates chance model frequencies (i.e., $1/\#\text{models} = 50\%$ with two models). **C** Model recovery analysis. The grey scale indicates protected exceedance probability averaged across all random seeds. **D** Parameter recovery analysis for the winning model. This figure displays the recovery results for one of the random seeds, but all other seeds were comparable. In all simulations recovery was good for all parameters (i.e., Cohen's $f^2 \geq 0.35$). Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*.[105]

2.4.2 Related modelling work

Although we replicated Adams and colleagues' findings[1] of a relative increase in belief instability in patients with psychotic disorders, we note that absolute belief instability in our sample was smaller. Furthermore, unlike others,[1,170] we only found trend-effects linking increased decision noise with symptom severity, although feature importance measures indicated that decision noise was relevant for treatment response prediction. This divergence may be explained by differences between clinical populations (schizophrenia vs psychotic disorders) or different tasks that were used (beads vs fish task), and possibly ensuing differences in task comprehension.

In contrast to other results,[61,156,187] Baker et al.[14] found that delusion severity correlated with more conservative behaviour, primarily, in a condition with high uncertainty

(60:40 beads ratio), which their model explained through increased reliance on priors. Intuitively, this agrees with belief rigidity — by definition a hallmark of delusions. The authors used a performance-contingent monetised beads task with endowment. The impact of performance-contingent versus flat payments (as in our case) is not entirely clear. Some authors argued that the payment mode may affect cognitive strategies employed, for example by setting new goals or spending cognitive resources on strategy development.[23,144] Furthermore, it is possible that endowments led to more loss-averse (conservative) instead of risk-seeking (liberal) behaviour. Due to differences in environmental uncertainty and payment structure, a direct comparison is difficult. However, our model appears to capture a mechanism underlying psychotic disorders in general and not specifically related to current delusions.

Other computational approaches were employed to characterise belief updating in schizophrenia.[116,117] Using a task related to ours, but without any sequential updating, Jardri et al.[117] suggested that schizophrenia is likely characterised by an overcounting of sensory information. Increased prior uncertainty has a comparable effect in early trials, because it increases the magnitude of belief updates, leading to stronger weighing of sensory information early on (Figure 2.2B). However, we found that belief instability, rather than prior uncertainty, differentiated patients with psychotic disorders from HC. Increasing belief instability primarily results in exaggerated belief updates, when faced with disconfirmatory evidence specifically, not an overcounting of any evidence.

2.4.3 Can belief instability be leveraged to predict treatment response?

At the group level, we found that belief instability significantly differed in patients who responded to an intervention targeting cognitive biases. Intriguingly, greater belief instability (i.e., more extreme pathology) related to better treatment response. One speculative explanation for this is that increased belief instability may indicate a vulnerable cognitive system, which places individuals at higher risk of being susceptible to delusional ideas,[4] but also more amenable to a therapy designed to make cognition more robust.

Subsequent analyses suggested that model parameters also predicted individual treatment response with 64% accuracy. Bearing in mind that treatment response prediction constitutes one of the most challenging problems in psychiatry and that MCT was merely an add-on treatment in patients already treated with antipsychotics, we believe this to be an encouraging result. Given previous evidence,[140] it is interesting to note that neither JTC nor clinical baseline measures predicted individual treatment response above chance. This finding may suggest that the model-derived computational fingerprint contains additional clinically-relevant information about inference mechanisms. This prognostic model may be a valuable screening instrument for clinical trials, or help reduce the therapy load on patients with motivational deficits. However, the accuracy based on model parameters alone is likely not sufficient to justify clinical implementation. Nonetheless, this model can provide a valuable component of a sequential prognostic test battery, together with other clinical or neurophysiological predictors, as proposed previously for transition-to-psychosis[34] or negative symptom prediction.[106]

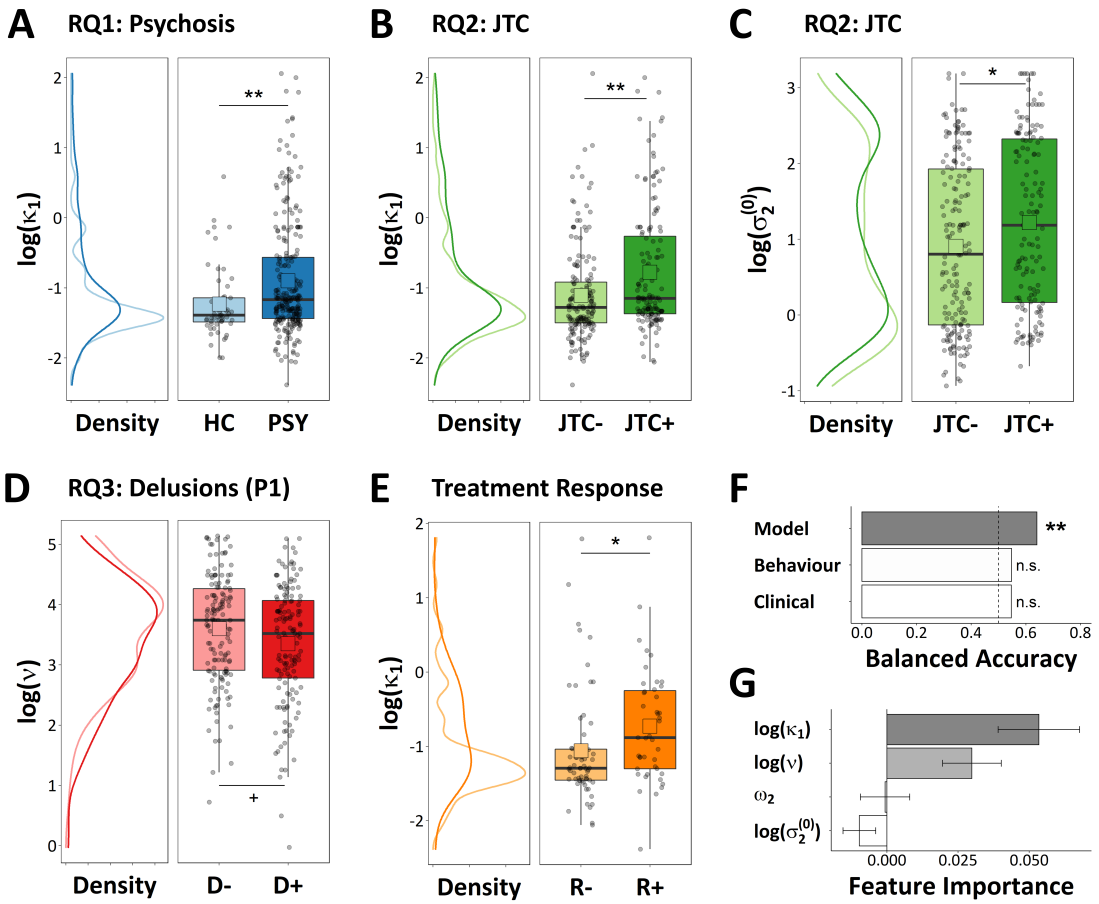

Figure 2.8: Parameter group effects and treatment response prediction. **A** Belief instability κ_1 across healthy controls (**HC**) and patients with psychotic disorders (**PSY**). **B** Belief instability κ_1 and **C** prior uncertainty $\sigma_2^{(0)}$ across individuals without (**JTC-**) and with (**JTC+**) jumping-to-conclusions bias (decision after ≤ 2 fish). **D** Decision noise ν across patients without (**D-**) and with (**D+**) any current delusions (Positive and Negative Syndrome Scale [126] (PANSS) item P1: Delusions ≥ 3). **E** Belief instability κ_1 across patients, who showed either no response (**R-**) or a response (**R+**) to Metacognitive Training defined as 20% decrease compared to baseline in the PANSS [126] positive factor according to factor.[244] **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: $^{**}p < 0.01$, and $^{*}p < 0.05$, using Bonferroni correction, or at $^{+}p < 0.05$ uncorrected. **F** Classification performance of random forest trained on either the winning models' parameters (**Model**), raw behavioural data (probability estimates and choices) and a jumping-to-conclusion bias indicator (**Behaviour**), or on PANSS baseline items (**Clinical**) to predict treatment response. Asterisks indicate significant permutation test with 1000 label permutations at: $^{**}p < 0.01$. **n.s.**: not significant. **G** Feature importance for the random forest trained on winning models' parameters. Bar size corresponds to mean and error bars to standard deviation across cross-validation folds. Adapted from Hauke et al. (2022), *Schizophrenia Bulletin*. [105]

To summarise, two notable benefits of this approach are (1) the interpretability of the predictors and (2) the simplicity of the assessment, since the model relies on very little data per participant. Task and model fitting can be performed fast rendering it attractive for clinical

applications, but the results still need to be replicated in different research sites.

A striking aspect of our results is that despite evident relationships between psychotic disorders and both behavioural and computational measures – and the potential for computational parameters to predict treatment outcome – we did not find any relationship between these measures and current delusions, even though these tasks were designed to assess reasoning biases thought to contribute to delusions themselves. Our findings add to a growing literature including meta-analyses,[13] large case-control,[239] and population-based studies[44] that find weak or absent correlations between delusion and beads task measures.

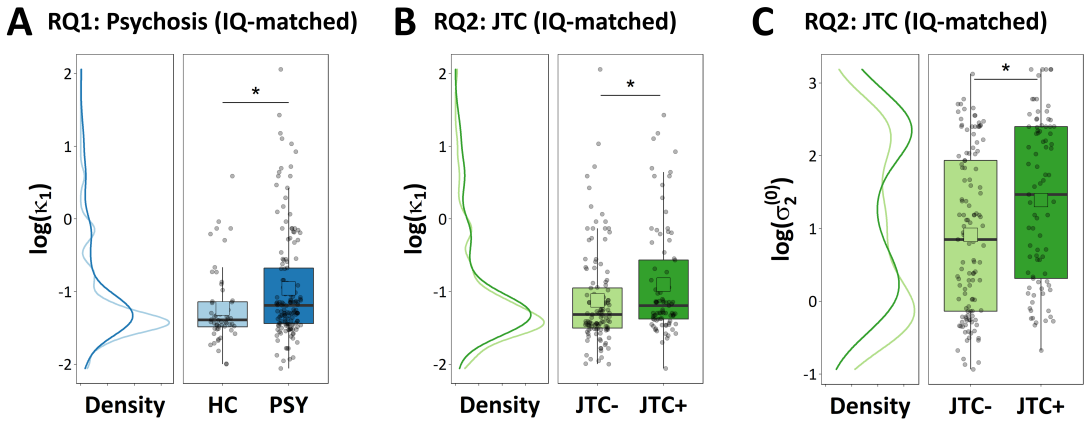

Figure 2.9: Parameter effects for IQ-matched subsamples. We repeated the parameter analysis with an IQ-matched subsample of patients. **A** Belief instability κ_1 across healthy controls (**HC**) and patients with psychotic disorders (**PSY**). **B** Belief instability κ_1 and **C** prior uncertainty $\sigma_2^{(0)}$ across individuals without (**JTC-**) and with (**JTC+**) jumping-to-conclusions bias (decision after seeing ≤ 2 fish). **RQ**: Research question. Horizontal lines and squares in boxplots represent median and mean, respectively. Boxes span the 25^{th} to 75^{th} quartiles and whiskers extend from hinges to the largest and smallest value that lies within $1.5\times$ interquartile range. Asterisks indicate significance of non-parametric Kruskal-Wallis tests at: * $p < 0.05$, using Bonferroni correction.

2.4.4 Limitations

Certain limitations merit attention: First, we only modelled ten trials per subject. While this increases clinical applicability, obtaining precise parameter estimates from such sparse data is challenging. Surprisingly, we could still recover parameters and were able to pinpoint computational mechanisms. Second, although we carefully controlled several confounders (education, medication, premorbid IQ), other confounders cannot be ruled out (e.g., socioeconomic status). More fine-grained measures of socioeconomic status should be included in future studies.[14] Thirdly, participants were not incentivised to respond quickly. Fast decisions could reflect patients' desire to end the experiment soon. However, participants were required to complete all trials rendering this unlikely. Furthermore, it is unclear, how monetisation affects the cognitive processes involved. Fourth, even though we defined treatment response as change scores and despite our finding that baseline symptoms did not predict treatment

response above chance, we cannot exclude influence of regression-to-the-mean effects on the treatment response prediction analysis presently. Lastly, without a clinical control group, we could not assess the specificity of increased belief instability, which is an important avenue for future research.

2.4.5 Future directions

Future studies are required to examine the physiological basis of belief instability. A candidate mechanism is NMDA receptor hypofunction [122,162] as a recent pharmacological study suggests that NMDA receptor functioning is linked to probabilistic reasoning during the beads task. [228] If this relationship can be confirmed, treatment response prediction to pharmacological interventions targeting glutamate metabolism (e.g., d-serine or glycine), may be a promising avenue of research. Furthermore, future research is required to assess, whether model parameters allow stratifying patients for clinical trials using MCT or similar interventions. Lastly, this model-based approach can also inform the design of new interventions that target belief instability specifically to assess whether such interventions can improve patients' well-being.

2.4.6 Conclusions

In conclusion, our results suggest that increased belief instability may be a key computational mechanism underlying probabilistic reasoning in patients with psychotic disorders. Furthermore, we provide a proof-of-concept that this computational parameter can potentially be leveraged to predict clinically-relevant outcomes.

For my part I know nothing with any certainty,
but the sight of the stars makes me dream.

Vincent van Gogh (1888)[245]

3 Modelling Sensory Learning

The previous chapters investigated the computational mechanisms underlying different symptoms of psychotic disorders, such as paranoid delusions (Chapter 1) and reasoning biases (Chapter 2). Moreover, Chapter 2 provided a test of the clinical utility of casting different symptoms of psychosis as instances of hierarchical Bayesian inference. This last experimental chapter will examine whether this approach is not only conceptually useful, but also biologically plausible.

3.1 Introduction

Often without our awareness, our brain continuously learns about the environment that surrounds us. The MMN is a biological index of such implicit learning. It refers to a brain response that occurs when a sensory stimulus violates a statistical regularity in the environment,[174] for example when a sequence of low tones is unexpectedly interrupted by a high tone (see Figure 3.1A). The MMN can be measured with EEG – a cost efficient diagnostic tool with high temporal resolution that is readily available in clinical practice. Formally, the MMN is defined as the difference waveform obtained when subtracting the electrophysiological response to a predictable stimulus (*standard*) from the response to an unpredictable stimulus (*deviant*, see Figure 3.1B).

Consistent reductions in MMN amplitude have been replicated in numerous studies with patients suffering from psychosis, rendering it one of the most reliable biomarkers of psychosis.[67] A number of pharmacological studies investigating the neuronal basis of the MMN were able to reproduce the effects found in patients with NMDA receptor antagonists (e.g.,

phencyclidine or ketamine), in different animals such as monkeys,[120] and rodents,[65] as well as in humans (see[238] for an overview). Together, these results support the notion that psychosis reflects an NMDA receptor dysfunction leading to reductions in the MMN and characteristic symptoms.[40,39,87,221,227,162,145]

The MMN has gained increasing interest in recent years as an early warning sign for psychosis. MMN reductions are already present in CHR individuals, likely reflecting vulnerability for progression of the disease as opposed to genetic risk.[67] The abnormalities increase when individuals progress towards a psychotic disorder, but saturate during chronification of the disease.[67] Importantly, MMN amplitude reductions were found to be more pronounced in those high risk individuals that later converted to a psychotic disorder.[67] Other studies highlighted the MMN's role in early stages of psychosis further, showing that it was predictive of the transition from a clinical high risk state to a first episode of psychosis.[22,177] Despite its great clinical potential, the mechanisms that account for these MMN alterations in the clinical high risk population remain poorly understood.

One of the biggest challenges in early detection and intervention research lies in determining the most effective medication to delay or even prevent a psychotic episode in a given patient.[49] This difficulty has been attributed to a lack of mechanistic models of pathophysiological processes, especially in the CHR population.[49] Here, we adopt a computational approach[253] to understand how information processing is altered in early psychosis and develop a mechanistic model of MMN reductions in the clinical high risk state.

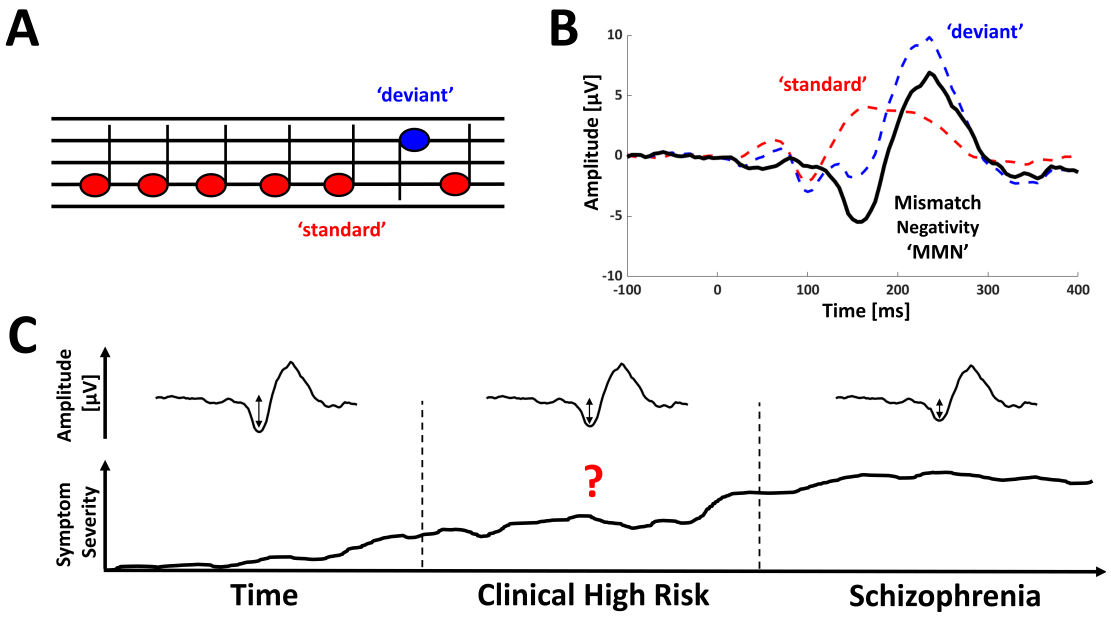

Figure 3.1: Mismatch negativity. **A** Example stimulus sequence to elicit auditory mismatch negativity (**MMN**). Violation of statistical regularity (established through repetition of low tones) elicits MMN. **B** MMN waveform is obtained by subtracting response to the predictable tone (*standard*) from the response to the unpredicted tone (*deviant*). **C** Research question.

3.2 Methods

To understand the computational mechanism underlying the MMN, we modelled EEG data from $N = 101$ participants, previously published by Perez et al.[177] For convenience, we will briefly recapitulate the most important information about the study.

3.2.1 Participants

A total of 19 early-illness schizophrenia patients (SCZ; ≤ 5 years since initial hospitalisation or initiation of antipsychotic medication), 38 CHR, and 44 HC were included in the study.[177] Out of the 38 CHR, 15 CHR later converted to psychosis. In a second analysis, we compared these 15 individuals to 16 CHR that did not transition to a psychotic disorder during a follow-up period of at least 12 months ($n = 7$ CHR dropped out of the study before the 12 month follow-up and were excluded from the second analysis). SCZ were referred by community physicians. CHR were recruited from the Yale Psychosis Prodrome Research Clinic and HC through advertisements and word-of-mouth. The study was approved by the Institutional Review Board of Yale University and all adult participants provided informed written consent. For minor participants, parents provided informed written consent and minor participants written assent.

3.2.2 In- and exclusion criteria

Schizophrenia diagnoses were assessed with the Structured Clinical Interview for DSM-IV[73] and CHR criteria based on the SIPS.[157,158] Participants of all groups were excluded from the study, if they fulfilled the following criteria: substance dependence or abuse within the past year, a history of significant medical or neurological illness or a head injury resulting in loss of consciousness, and abnormal audiometric testing (see[177]). Additionally, HC, who met criteria for any past or current DSM-IV Axis I disorder or had a first-degree relative with a psychotic disorder were excluded.

3.2.3 Task

Participants performed an unrelated primary task (silently reading a book) while undergoing three different auditory MMN paradigms presented in fixed order. Each paradigm comprised two runs with 875 tones each (1750 tones in total) including 90% standard tones (50 ms, 633 Hz) and either (1) 10% duration (100 ms), (2) 10% frequency (1000 Hz), or (3) 10% duration + frequency double deviants (100 ms *and* 1000 Hz). All tones were presented at 78 dB in fixed pseudorandomized order with 5 ms rise/fall times and 510 ms stimulus onset asynchrony through Etymotic ER3-A insert earphones (Etymotic Research, Inc., Elk Grove Village, Illinois).

3.2.4 EEG data processing

EEG was recorded using a 20-channel electrode cap with a standard 10-20 montage (Physiometrix, Inc., North Billerica, Massachusetts) and additional mastoid and nose electrodes with linked-ear reference and an FPz ground. Signals were digitised at 1000 Hz with a Neuroscan Synamps amplifier (Neuroscan, Herndon, Virginia). Electro-oculograms were recorded from electrodes located above and below the left eye and at the outer canthi of both eyes.

EEG preprocessing consisted of downsampling (256 Hz), bandpass filtering between 0.5 and 30 Hz using a Butterworth filter, epoching into 500 ms segments around tone onsets (-100 to 400 ms), baseline correction (-100 to 0 ms), and eyeblink correction using principal component analysis with 1 component. Eyeblink components of all participants were manually inspected and eyeblink detection thresholds adjusted if necessary, followed by rejection of remaining artefactual trials (using a $\pm 100\,\mu$V amplitude threshold). Preprocessing and statistical analyses were implemented in Matlab (version: 2020b; `https://mathworks.com`) using the SPM12 toolbox (`https://www.fil.ion.ucl.ac.uk/spm/software/spm12/`).

3.2.5 Computational modelling

Perceptual model

We modelled implicit sensory learning about the tone sequences using a 3-level binary HGF.[153,154] This model assumes that participants infer on a number of hidden states of the environment (Figure 3.2, left). In the context of the MMN paradigm, the states that participants' need to infer on based on the experimental input on each trial (*standard* or *deviant* tones) are structured as follows: The lowest level state corresponds to the *tone probability*. On each trial k a tone can either be deviant ($x_1^{(k)} = 1$) or a standard tone ($x_1^{(k)} = 0$). This state can be described by a Bernoulli distribution that is linked to the state at the second level $x_2^{(k)}$ through the unit sigmoid transformation:

$$p(x_1^{(k)}|x_2^{(k)}) = s(x_2^{(k)})^{x_1^{(k)}}(1 - s(x_2^{(k)}))^{1-x_1^{(k)}} \sim \text{Bernoulli}(x_1^{(k)}; s(x_2^{(k)})), \tag{3.1}$$

with

$$s(z) = \frac{1}{1 + e^{-z}}. \tag{3.2}$$

$x_2^{(k)}$ represents the unbounded tendency towards standard or deviant tones $(-\infty, +\infty)$ or the *tone tendency* and is specified by a normal distribution:

$$p(x_2^{(k)}|x_2^{(k-1)}, x_3^{(k)}, \kappa_2, \omega_2) \sim \mathcal{N}(x_2^{(k)}; x_2^{(k-1)}, \exp(\kappa_2 x_3^{(k)} + \omega_2)). \tag{3.3}$$

The state at the third level $x_3^{(k)}$ expresses the (log) volatility of the environment over time and is also specified by a normal distribution:

$$p(x_x^{(k)}|x_3^{(k-1)}, \theta) \sim \mathcal{N}(x_3^{(k)}; x_3^{(k-1)}, \theta). \tag{3.4}$$

Participants' beliefs about these hidden states at level i of the hierarchy and on trial k are denoted with $\mu_i^{(k)}$ and updated after each new tone according to the following update equation:

$$\underbrace{\Delta\mu_i^{(k)}}_{\text{Belief Update}} \propto \overbrace{\frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}}}^{\text{Precision-Weight}} \underbrace{\delta_{i-1}^{(k)}}_{\text{Prediction Error}}, \tag{3.5}$$

where $\mu_i^{(k)}$ is the expectation or belief at trial k and level i of the hierarchy, $\hat{\pi}_{i-1}^{(k)}$ is the precision (inverse of the variance) from the level below (the hat symbol denotes that this precision has not been updated yet and is associated with the prediction before hearing a new tone), $\pi_i^{(k)}$ is the updated precision at the current level, and $\delta_{i-1}^{(k)}$ is a PE expressing the discrepancy between the expected and the experienced outcome.

In line with a previous study examining the effects of ketamine on sensory learning in a roving paradigm,[253] we focused our analysis on low-level precision-weighted PEs about the tone tendency (ε_2) and high-level precision-weighted PEs about the volatility of the environment (ε_3), where the precision-weighted PE $\varepsilon_i^{(k)}$ on each trial k and at level i of the hierarchy is defined as (cf. Eq. 3.5):

$$\varepsilon_i^{(k)} = \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)}. \tag{3.6}$$

We implemented this model using the 'tapas_ehgf_binary' function from the HGF toolbox (version 6.0), which is made available as open-source code as part of the TAPAS[79] software collection (version: `5.1.0`; `https://github.com/translationalneuromodeling/tapas/releases/tag/v5.1.0`) in Matlab (version: `2020b`; `https://mathworks.com`). We used this recently developed enhanced version of the HGF to improve sensitivity to learning about environmental volatility. The main distinction with respect to earlier versions of the HGF is that the posterior means $\mu_i^{(k)}$ are updated before the precisions $\pi_i^{(k)}$ at level i of the hierarchy. For more details on the update equations, see[153,154] for the original HGF and the 'tapas_ehgf_binary' function for the eHGF.

Since the MMN paradigm is a passive task that does not require participants to make re-

sponses, we optimised the parameters of the perceptual model assuming an ideal Bayesian observer that minimises the cumulative Shannon surprise for a given input sequence using the 'tapas_bayes_optimal_binary' function. The prior settings (mean, variance) for this optimisation were $(-3, 4)$ for the *evolution rate* ω_2 and $(2, 4)$ for *meta-volatility* θ. The *coupling strength* κ_2 was fixed to $\log(1)$. Posterior parameter estimates are summarised in Table 3.1.

	HC $n = 44$	**CHR** $n = 38$	**SCZ** $n = 19$	**CHR-C** $n = 15$	**CHR-NC** $n = 16$
evolution rate ω_2 mean [SD]	-0.20 [0.78]	-0.19 [0.77]	-0.36 [0.85]	-0.16 [0.82]	-0.22 [0.78]
meta-volatility θ mean [SD]	4.80 [0.30]	4.86 [0.29]	4.80 [0.31]	4.76 [0.29]	4.83 [0.31]

Table 3.1: Summary of posterior parameter estimates. **HC**: Healthy controls. **CHR**: Individuals at clinical high risk for psychosis. **SCZ**: Early illness schizophrenia patients ($\leq$ 5 years since initial hospitalisation or initiation of antipsychotic medication). **CHR-C**: Converters. **CHR-NC**: Non-converters.

3.2.6 Statistical analysis

Demographic and clinical variables were analysed in R (version: 4.04; `https://www.r-project.org/`) using R-Studio (version: 1.4.1106; `https://www.rstudio.com/`). We report uncorrected p-values for either ANOVAs or χ^2-tests where appropriate. Post hoc tests were Bonferroni-corrected.

First level analysis

We extracted the trajectories of low-level precision-weighted PEs about the tone tendency ε_2 and high-level precision-weighted PEs about the volatility of the environment ε_3. Trial-by trial magnitude estimates of the absolute value of low-level precision-weighted PEs $|\varepsilon_2|$ or high-level precision-weighted PEs ε_3 were included as parametric regressors to explain trial-by-trial variation in EEG amplitude (see Figure 3.2) as done previously.[253] The absolute value of ε_2 was chosen, because it expresses Bayesian surprise independent of the physical characteristics of a tone such as a specific frequency. The general linear model at the first level consisted of an intercept term and either (z-standardised) low or high-level precision-weighted PE trajectories as predictors and EEG amplitude across sensors and peristimulus time as the response variable. For each precision-weighted PE, we tested the null hypothesis that the parameter estimate was zero at each sensor and time point using an F-test. Statistical analyses were restricted to 100 ms to 400 ms post-stimulus time.

Second level analysis

First-level statistics were converted into images and smoothed using a Gaussian kernel (FWHM: 16 mm x 16 mm) to ensure that the assumptions of Gaussian random field theory were met.[129,264] Smoothed images were carried to the second level to compare groups using different factorial designs for each precision-weighted PE to obtain statistical parametric maps over 2D sensor space and peristimulus time (see Figure 3.2). Each factorial design included group as between- and MMN paradigm as within-subject factor, as well as age as a covariate. To ensure that the equal slope assumption for age was met, we masked out voxels that showed a significant group-by-age interaction. Multiple testing correction was implemented using Gaussian random field theory[129,264] and we report p-values corrected for peak- (p_{pFWE}) or cluster-level (p_{cFWE}) family-wise error rates using a cluster defining threshold of $p < 0.001$.[75]

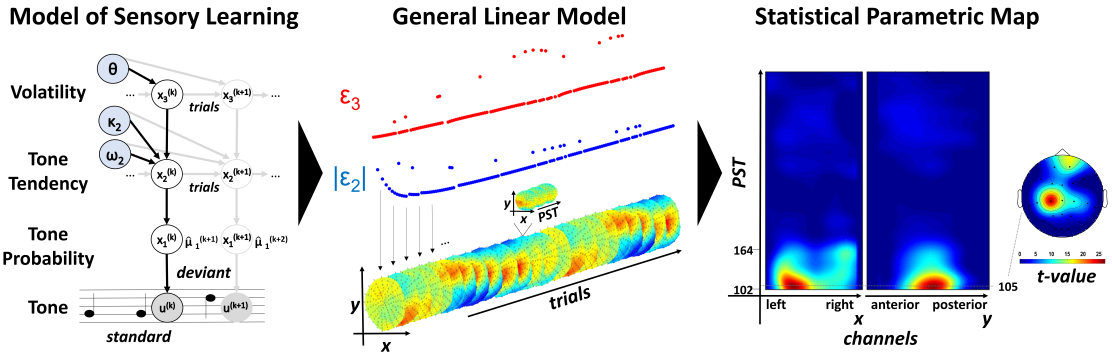

Figure 3.2: Computational analysis pipeline. Trial-by trial trajectories of low- and high-level precision-weighted were computed using the Hierarchical Gaussian Filter[153,154] (**left**). In a first level analysis, precision-weighted prediction errors were used as parametric regressors to explain EEG amplitude variations at each point in sensor space and peristimulus time (**PST**) across trials within each participant (**middle**). First level statistics were carried to the second level to obtain statistical parametric maps over 2D sensor space and peristimulus time (**right**). **EEG**: Electroencephalography.

3.3 Results

3.3.1 Sociodemographic and clinical characteristics

Demographic and clinical characteristics are displayed in Table 3.2 (see also[177] for more information).

3.3.2 Group differences in the expression of low-level precision-weighted prediction errors

We observed a significant group effect on the expression of low-level precision-weighted PEs about the tone tendency ε_2 peaking at 105 ms over left, central channels ($F = 20.795, p_{cFWE} < 0.001$) and at 109 ms over frontal channels ($F = 15.656, p_{pFWE} < 0.001$). Closer inspection

of the first effect revealed that low-level precision-weighted PEs correlated with more positive EEG amplitudes in central channels in SCZ compared to CHR (peak: 152 ms, $t = 4.923$, $p_{cFWE} = 0.001$; Figure 3.3) and SCZ vs HC (peak: 105 ms, $t = 6.427$, $p_{cFWE} < 0.001$; Figure 3.3). The second effect suggested that increased low-level precision-weighted PEs correlated with more negative EEG amplitudes over frontal channels in SCZ vs HC (peak: 109 ms, $t = 5.594$, $p_{pFWE} < 0.001$; Figure 3.3).

	HC $n = 44$	**CHR** $n = 38$	**SCZ** $n = 19$	Test statistic	Post hoc contrasts	**CHR-C** $n = 15$	**CHR-NC** $n = 16$	Test statistic
Age	19.97	17.40	23.91	$F = 10.838$	SCZ > HC	17.47	15.88	$F = 2.475$
mean [SD]	[5.50]	[3.50]	[6.17]	$\mathbf{p < 0.001}$	SCZ > CHR	[2.18]	[3.27]	$p = 0.127$
Sex	17/27	15/23	4/15	$\chi^2 = 2.178$		6/9	7/9	$\chi^2 = 0.045$
f/m				$p = 0.337$				$p = 0.833$
Handedness[a]	37/5/2	31/3/4	16/1/2	$\chi^2 = 1.765$		13/1/1	12/1/3	$\chi^2 = 1.009$
r/l/a				$p = 0.779$				$p = 0.604$
High risk type[b]								
APS		38				15	16	
BLIP		1				1	0	
GRD		1				1	0	
Diagnostic type								
Paranoid			11					
Disorganised			1					
Undifferentiated			2					
Catatonic			1					
Residual			1					
Schizoaffective			3					
Antipsychotic type								
Atypical only		10	13			5	3	
Typical only		0	0			0	0	
Atypical and typical		1	3			0	0	
None		27	2			10	13	
Unknown		0	1			0	0	
PANSS Positive			18.71					
mean [SD]			[5.78]					
PANSS Negative			17.14					
mean [SD]			[6.11]					
SOPS Positive		11.03				12.47	9.00	$F = 3.739$
mean [SD]		[4.96]				[5.07]	[4.91]	$p = 0.063$
SOPS Negative		10.74				14.40	6.69	$F = 15.767$
mean [SD]		[6.35]				[5.05]	[5.71]	$\mathbf{p < 0.001}$

Table 3.2: Demographic and clinical characteristics. All p-values are uncorrected. **HC**: Healthy controls. **CHR**: Individuals at clinical high risk for psychosis. **SCZ**: Early illness schizophrenia patients (≤ 5 years since initial hospitalisation or initiation of antipsychotic medication). **CHR-C**: Converters. **CHR-NC**: Non-converters. **APS**: Attenuated psychotic symptoms. **BLIP**: Brief and limited intermittent psychotic symptoms. **GRD**: Genetic risk and deterioration syndrome. **PANSS**: Positive and Negative Syndrome Scale.[126] **SOPS**: Scale of Prodromal Symptoms.[157,158] Bold print highlights p-values significant at: $p < 0.05$, uncorrected. [a]Crovitz-Zener questionnaire for handedness (right, left, or ambidextrous). [b]High risk types are not mutually exclusive.

3.3.3 Group differences in the expression of high-level precision-weighted prediction errors

The expression of high-level precision-weighted PEs about the volatility of the environment ε_3 also showed a significant effect of group peaking at 125 ms over right, central channels ($F = 17.277, p_{cFWE} = 0.005$). Pairwise comparisons revealed that larger high-level precision-weighted PEs correlated with more positive EEG amplitudes in HC compared to SCZ over frontal channels (peak: 125 ms, $t = 3.931, p_{pFWE} = 0.027$) and during a later time window over posterior central channels (peak: 344 ms, $t = 3.821, p_{cFWE} = 0.018$; Figure 3.4), which was also significant when comparing CHR to SCZ (peak: 340 ms, $t = 3.621, p_{cFWE} = 0.046$; Figure 3.4). Furthermore, we found that stronger precision-weighted PEs correlated with more negative amplitudes during an early time window in SCZ vs CHR (peak: 129 ms, $t = 5.014, p_{cFWE} = 0.008$; Figure 3.4) and in SCZ vs HC (peak: 125 ms, $t = 5.728, p_{cFWE} = 0.002$; Figure 3.4).

3.3.4 Group differences between converters and non-converters

Lastly, when comparing CHR converters to non-converters, we found a significant group effect on the expression of low-level precision-weighted PEs ε_2 peaking at 137 ms over left, central channels ($F = 12.722, p_{cFWE} = 0.040$; small-volume corrected for the group effect on ε_2 between HC and SCZ). In CHR individuals that later transitioned to psychosis, larger low-level precision-weighted PEs were correlated with more positive EEG amplitudes (peak: 137 ms, $t = 3.567, p_{cFWE} = 0.022$; small-volume corrected for the group effect on ε_2 between HC and SCZ; Figure 3.3).

3.4 Discussion

The objective of this study was to understand how information processing is altered in early psychosis and test a mechanistic model of the MMN, one of the most reliable biomarkers of psychosis.[67] We obtained three major findings: First, we observed altered expression of low-level precision-weighted PEs about the tone tendency between HC and SCZ and in CHR compared to SCZ. Second, we also identified changes in the expression of high-level precision-weighted PEs about the volatility of the environment in SCZ compared to both HC and CHR during an early time window (at about 100–175 ms peristimulus time), as well as during a later time window (∼320–380 ms). Third, the expression of low-level precision-weighted PEs was significantly altered in those CHR that later converted to a psychotic disorder compared to non-converters suggesting that this computational model appears to capture relevant pathophysiological mechanisms and may constitute a useful tool to predict transition to psychosis in individual patients.

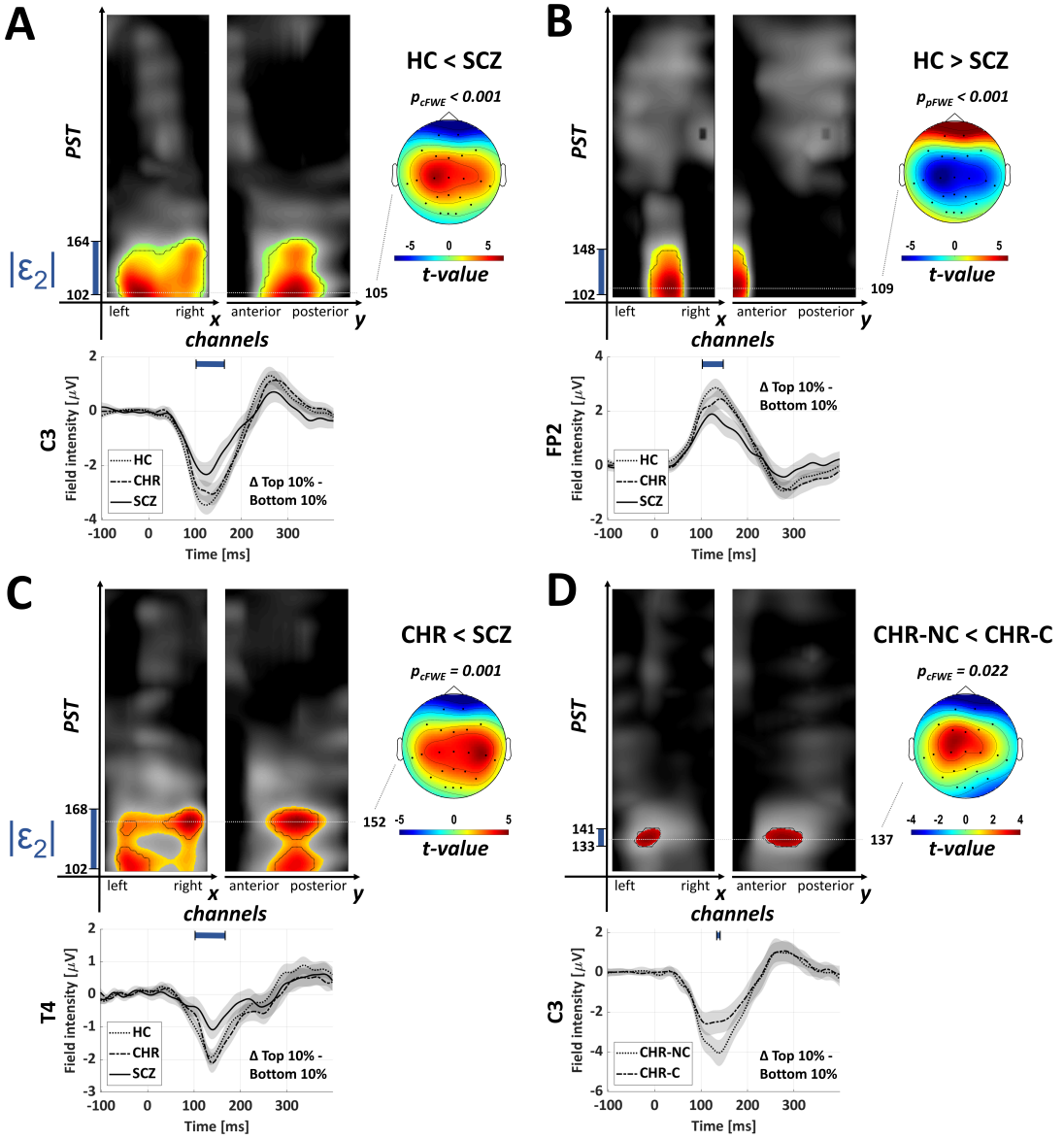

Figure 3.3: Expression of low-level precision-weighted prediction errors in early psychosis. **A-D**: Displayed are maximum intensity projections highlighting significant voxels of t-contrasts testing for pairwise group differences in the expression of low-level prediction errors ε_2 about the tone tendency in emerging psychosis. Times displayed on y-axis indicate earliest and latest significant voxel. p-values were corrected for peak- (p_{pFWE}; black dashed-line) or cluster-level (p_{cFWE}) family-wise error rates (**FWE**) using a cluster defining threshold of $p < 0.001$ (highlighted by coloured area). Note, that p-values in **D** are small-volume corrected for the group effect on ε_2 between HC and SCZ. For illustration, difference waveforms (10% highest - 10% lowest ε_2 trials) are shown across groups for a channel close to the peak effect. **HC**: Healthy controls. **CHR**: Individuals at clinical high risk for psychosis. **CHR-C**: Converters. **CHR-NC**: Non-converters. **SCZ**: Early-illness schizophrenia patients (illness duration ≤ 5 years).

3.4.1 Theoretical implications for the predictive coding account of psychosis

Our results are in line with the predictive coding account of psychosis that postulates that disturbances in hierarchical PE processing may contribute to psychotic symptoms.[76,227] Our finding of alterations in the expression of precision-weighted PEs in central channels in patients with schizophrenia may suggest that patients experience aberrantly salient PEs[125] in response to familiar stimuli (*standard* tones). Furthermore, changed expression of hierarchical PEs in frontal channels could signal a decrease in prior precision in frontal regions as proposed in the predictive coding account.[227]

3.4.2 Possible cortical generators of aberrant PEs in early psychosis

The network of cortical regions involved in generating the MMN response is well-understood. It includes bilateral primary auditory cortex (A1), superior temporal gyrus (STG) and inferior frontal gyrus (IFG).[59,99,149,176] It is possible that the different spatiotemporal clusters that were identified in our study may be caused by different cortical generators, for example the correlation between precision-weighted PEs and more positive EEG amplitudes in SCZ expressed in central channels may originate in A1 or STG, while the second cluster, which was identified over frontal regions, possibly suggests involvement of IFG. However, due to volume-conduction effects, we will have to formally test this hypothesis using source modelling in the future.

Adams and colleagues[2] recently investigated the neural mechanisms of schizophrenia using dynamic causal modelling and found remarkably consistent findings across a wide range of paradigms pointing at a loss of synaptic gain of pyramidal cells in schizophrenia. Notably, this study also included an MMN paradigm, in which the authors identified loss of pyramidal gain in IFG specifically. This finding suggests that the expression of hierarchical PEs over frontal channels may be altered due to a reduction of the precision-weight rather than changes in the PE component of the precision-weighted PEs.

3.4.3 Are aberrant precision-weighted prediction errors related to alterations in neurotransmission?

The dysconnectivity hypothesis[87,90,91,221,222] postulates that NMDA receptor-mediated modulation of synaptic gain is altered in psychosis. In line with this account, Weber et al.[253] found that ketamine administration led to a reduced expression of high-level precision-weighted PEs about the volatility of the environment in central channels similar to our results. However, their results suggest that low-level precision-weighted PEs are unaffected by ketamine.

Several neurotransmitters interact with NMDA receptors to dynamically control synaptic gain and neuroplasticity. Altered expression of precision-weighted PEs in SCZ as identified in our study over early auditory regions could reflect changes in cholinergic neurotransmission. Two recent studies implicate acetylcholine in regulating synaptic gain or – according to the

predictive coding account and the dysconnectivity hypothesis – regulating sensory precision in early auditory regions.[163,203] The first study employed a Kalman filter (i.e., a 2-level HGF) to model changes in participants that were administered galantamine, which enhances cholinergic neurotransmission.[163] The authors argued that galantamine may increase the precision of sensory PEs. The second study modelled changes in between- and within-region connectivity including synaptic gain during a pharmacological manipulation using muscarinic receptor antagonist scopolamine or muscaranic receptor agonist pilocarpine in rats.[203] Schöbi and colleagues[203] found dose-dependent changes in synaptic gain, but also changes in inter-regional connectivity between A1 and secondary auditory cortex. Moreover, changes in muscarinic receptor density among schizophrenia patients have even been frequently reported[51,191,192,193] and Scarr et al.[191] proposed that there may be a subgroup of schizophrenia patients specifically characterised by decreased cortical muscarine receptor expression. These results support a potential role of cholinergic neurotransmission in precision-weighted PE signalling. Beyond cholinergic processes, glutamatergic neurotransmission at AMPA receptors may be involved, but its precise role still needs to be clarified.

3.4.4 Clinical implications

Interestingly, our results suggest that the expression of low-level PEs aggregated across three different MMN designs is significantly altered in CHR that later converted to a psychotic disorder compared to CHR that did not convert. This finding highlights potential applications of this computational approach to transition-to-psychosis prediction. Furthermore, if the neurotransmitter systems that are involved in computing precision-weighted PEs during the MMN paradigm can be identified, this approach may be useful for predicting treatment response to pharmacological interventions that target either glutamatergic neurotransmission like d-serine, which has shown promising results in a recent clinical trial,[124] or cholinergic neurotransmission such as clozapine or olanzapine.

3.4.5 Limitations

A few limitations of this study merit attention. First, the MMN paradigms in this study were not well-suited to separate low- and high-level PEs, because environmental volatility was not manipulated explicitly throughout the task. Future studies should include explicit manipulations of volatility to better distinguish between different levels of hierarchical inference. Secondly, we were not able to fit individual responses to determine subject-specific parameters. This is an inherent limitation of the MMN paradigm, which is usually administered as a passive task. Without estimating subject-specific parameters, our results are more challenging to interpret. Group differences could arise because different groups are better explained by different models as highlighted in Chapter 1 or by different parameter values as was the case in Chapter 2. While the MMN is one of the most reliable biomarkers for psychosis,[67] future studies should also investigate the representation of precision-weighted PEs using odd-

ball paradigms that require an explicit response of participants, such as the paradigm used in a recent study,[104] which found that P3b amplitudes were predictive of conversion to psychosis.

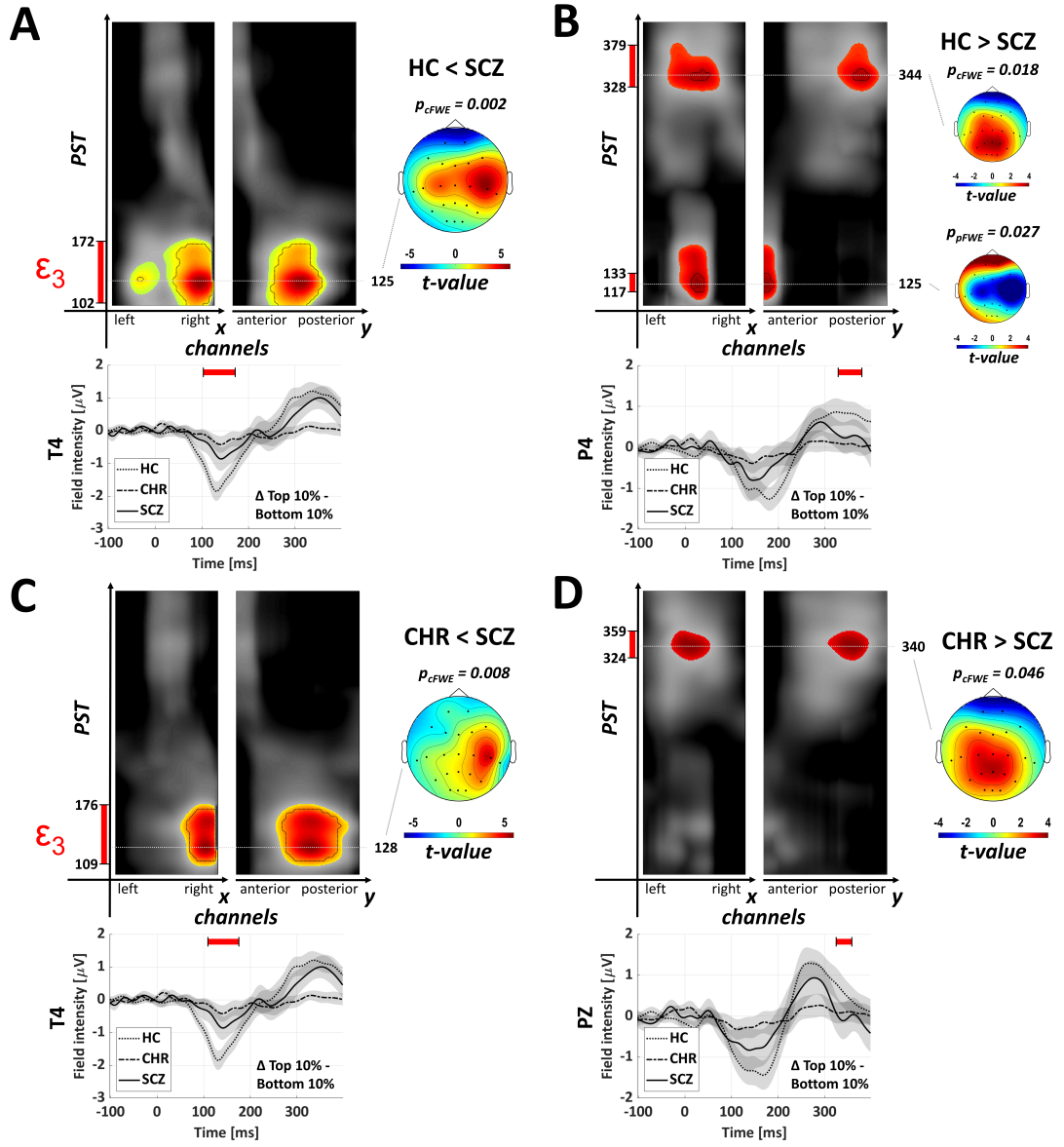

Figure 3.4: Expression of high-level precision-weighted prediction errors in early psychosis. A-D: Displayed are maximum intensity projections highlighting significant voxels of t-contrasts testing for pairwise group differences in the expression of high-level prediction errors ε_3 about environmental volatility in emerging psychosis. Times displayed on y-axis indicate earliest and latest significant voxel. p-values were corrected for peak- (p_{pFWE}; black dashed-line) or cluster-level (p_{cFWE}) family-wise error rates (**FWE**) using a cluster defining threshold of $p < 0.001$ (highlighted by coloured area). For illustration, difference waveforms (10% highest - 10% lowest ε_3 trials) are shown across groups for a channel close to the peak effect. **HC**: Healthy controls. **CHR**: Individuals at clinical high risk for psychosis. **SCZ**: Early-illness schizophrenia patients (illness duration ≤ 5 years).

3.4.6 Future directions

Future studies should determine the cortical generators of changes in the expression of hierarchical precision-weighted PEs in early psychosis. Moreover, the biological implementation of these computations needs to be clarified further, for example through the use of models that include greater physiological detail to bridge the algorithmic description that our modelling approach offers and Marr's implementational level of analysis.[151] Dynamic causal models for electrophysiological data in general and conductance-based dynamic causal models specifically have been highlighted as computational assays that may allow inference on receptor densities of neuronal populations.[221,224] These models have been validated in studies investigating NMDA receptor antibody encephalitis,[234] dopaminergic action on NMDA receptors,[164] and manipulations of cholinergic neurotransmission[203] and thus constitute a promising way forward. Additionally, there is a need for more pharmacological studies in both animals and humans to map the relationship between hierarchical precision-weighted PEs and different neurotransmitter systems that are targeted by antipsychotic medication.

3.4.7 Conclusions

In this study, we examined the computational mechanisms underlying MMN reductions in emerging psychosis and found evidence for aberrant expression of precision-weighted PEs at different levels of hierarchical inference. Our results suggest that the expression of low-level precision-weighted PEs is significantly altered in individuals at clinical high risk for psychosis that will later transition to psychosis highlighting that this computational modelling approach captures relevant pathophysiological mechanisms and may prove useful for predicting transition to psychosis in individual patients.

4 Conclusions

The goal of this thesis was to understand symptoms of psychosis through the computational lens of Bayesian inference and assess the clinical utility and biological plausibility of this approach. Chapter 1 investigated the emergence of paranoid delusions. Our results suggest that emerging psychosis may be accompanied by an altered perception of environmental volatility. Chapter 2 applied this modelling approach to delusions more broadly and examined their relationship with reasoning biases. We found that beliefs of patients with psychotic disorders were more unstable, which explained increased belief updating in light of disconfirmatory evidence. Furthermore, the parameters of the computational model could predict treatment response to a psychotherapeutic intervention, providing support for the clinical utility of this computational framework. Chapter 3 assessed the biological plausibility of this approach. Analysing EEG data during a sensory learning task, we identified signatures of precision-weighted PEs derived from the model in EEG recordings. This result highlights the possibility that this approach may not only be conceptually or clinically useful, but also biologically plausible, although further investigation is warranted.

4.1 Theoretical implications

4.1.1 Implications for the dopamine hypothesis

It has been hypothesised that chaotic firing of dopamine neurons may lead to assigning aberrant salience to otherwise irrelevant stimuli.[125] In line with this proposal, we found in Chapter 1 that early psychosis was accompanied by changes in parameters that govern learning about environmental volatility. Specifically, changing these parameters results in higher learning

rates or increased belief updating at lower levels of an inferential hierarchy during a social learning task (Figure 1.6). Importantly, aberrant salience could be captured either by an increase in the precision associated with low-level beliefs, i.e., through increasing the numerator in the HGF update equation, or through a decrease in the precision associated with higher level beliefs, i.e., decreasing the denominator (cf., Eq. 1.5 and Figure 1.4). Our results and other recent findings[184,231,53] suggest the latter mechanism. A possible involvement of dopamine is further supported by indirect evidence from two pharmacological studies in healthy controls, which found drug-induced changes in activation of dopaminergic midbrain regions during the same task.[57]

Increased belief instability as identified in Chapter 2 has also been discussed in relation to dopaminergic processing previously,[1] although recent pharmacological studies, which administered dopamine antagonist haloperidol and dopamine precursor levodopa to healthy participants, did not report an impact of dopamine on jumping-to-conclusions.[7,8] However, the authors did not examine other reasoning biases like increased updating in light of disconfirmatory evidence, which is primarily reflected in the belief instability parameter. Moreover, previous biophysical modelling work suggested that shifts in the D1:D2 receptor *ratio* can lead to a more stable D1-dominated regime, in which switching between different attractor states (i.e., states that a dynamic system approaches over time) becomes less likely, or a less stable D2-dominated state, in which switching between attractor states becomes more likely.[62] This may be captured by the belief instability parameter introduced in Chapter 2 and would imply that differential receptor binding affinities for D1 vs D2 receptors should be taken into account in future pharmacological studies. Future studies should examine the impact of substances such as asenapine, olanzapine, zotepine, or chlorpromazine, which have higher affinity for D1 receptors, specifically.[216]

Lastly, it is unlikely that dopamine mediates precision-weighted PEs during the oddball paradigm (Chapter 3) as most studies did not find an association between dopamine and MMN amplitudes[238] (but also see[215,268]).

4.1.2 Implications for the glutamate hypothesis

The glutamate hypothesis of schizophrenia postulates that alterations in glutamatergic neurotransmission may be a core pathophysiology of schizophrenia.[162] This hypothesis was based on the observation that NMDA receptor antagonists like ketamine do not only produce effects reminiscent of core symptoms observed in patients, but also MMN amplitude reductions.[162,67] Chapter 3 studied the expression of precision-weighted PEs during three different MMN paradigms and identified alterations in both low- and high-level precision-weighted PEs in early psychosis. Weber et al.[253] reported an effect of ketamine on the expression of high-level precision-weighted PEs implicating NMDA receptor functioning. This finding points towards a potential explanation of changes in high-level precision-weighted PEs in emerging psychosis, namely alterations in NMDA receptor functioning. However, we also found that low-level precision-weighted PEs differentiated converters from non-converters. Moreover,

others report a positive association between low-level precision-weighted PEs and prodromal symptom severity in CHR.[32] Thus, it will be important to clarify the neuropharmacological basis for low-level precision-weighted PEs during the oddball task. Acetylcholine appears to be a promising candidate mechanism since previous studies showed an impact of cholinergic interventions on both the MMN as well as the computational mechanisms underlying it.[163,203,254]

4.1.3 Implications for the dysconnectivity hypothesis

The results presented in the three experimental chapters are in line with the dysconnectivity hypothesis, which emphasises the interaction between neuromodulators and NMDA-receptors.[91,87,221,222,90] Importantly, although different neuromodulatory systems may be involved in computing precisions across different cognitive tasks, these neuromodulators may still converge on NMDA receptors.[102] Once our results have been replicated, it will be important to investigate whether the changes in the expression of precision-weighted PEs in emerging psychosis (Chapter 3) are driven by changes in the computation of precisions or PEs and secondly, to determine the neurotransmitters involved in computing these computational quantities. However, importantly, a recent study suggests that the relationship between computational and neural mechanisms may not be straightforward[113] highlighting that the interaction between different neurotransmitter systems remains an important field of future research.

4.1.4 Implications for the predictive coding account

Our results are in line with the predictive coding account of psychosis, which emphasises perturbations in the precisions associated with incoming sensory information and prior expectations.[76,227] We found more uncertain high-level priors (Chapter 1), increased belief instability (Chapter 2), and altered expression of precision-weighted PEs in patients with psychotic disorders (Chapter 3). Jointly, our results suggest that this modelling approach may be conceptually useful to understand how information processing is changed in psychotic disorders. It can be employed to map developmental changes in information processing across the lifespan and different phases of psychosis.

Furthermore, this approach allows us to examine whether symptoms of psychosis are caused by a uniform deficit in predictive coding or whether different symptoms are explained by different computational mechanisms, for example changes in low- or high-level precisions (see Eq. 1.5). While the literature has consistently found changes in parameters that relate to learning about environmental volatility across different (social and non-social) tasks and even species,[184,231,53] the study of other symptoms such as hallucinations has yielded more heterogeneous results (see[227] for an overview). Sterzer and colleagues[227] have argued that this may be resolved by localising the deficits at different levels of a processing hierarchy.

Moreover, it is important to consider along which dimension information is hierarchically organised in the brain.[260] In the HGF, hierarchical levels are coupled via their variances, which implies representation of a hierarchy in *time*, where the third level reflects the (more slowly changing) rate of change of the level below. Other fields have assumed a hierarchy of *causes* or of what is represented, for example letters nested in words nested in sentences associated with a specific meaning in hierarchical Bayesian schemes for semantic processing.[76] Importantly, these different possibilities are not necessarily mutually exclusive as larger objects, for example a face, will also remain longer in the receptive field of a cell than a simple visual feature like a short edge as detected by cells in primary visual cortex. Nonetheless, it is important to clarify what information is hierarchically organised, for example by comparing different hierarchical models.

4.2 Clinical implications

There are several avenues to employ the computational approach presented in this thesis for clinical applications. First of all, Chapters 1 and 3 outlined the possibility of using this approach to predict transition to a psychotic disorder from a clinical high risk state. In Chapter 1, we observed heterogeneity in model attributions, especially in the high risk group, in which the model formalising the notion of an altered perception of environmental volatility was favoured for some individuals, whereas the standard HGF was better accounting for behaviour in others. In Chapter 3, we found that the expression of low-level precision-weighted PEs significantly differed between individuals who later converted to a psychotic disorders and non-converters.

While there have been successful approaches to predict transition to psychosis based on anatomical data,[45,138] the approach proposed here bares the advantage that it is mechanistically interpretable. Using computational model parameters as features for transition-to-psychosis prediction does not only drastically reduce the dimensionality of the feature space, but it facilitates explaining why a prognostic model arrives at an individual risk prediction. Understanding why an algorithm comes to the conclusion that a given individual has a high risk of transitioning to psychosis is essential to both clinical practitioners and patients. It enables clinical practitioners to identify the most suitable treatment strategy, but also to spot potential errors of a predictive model, while also helping patients to understand the (biological) basis of their condition better.

Secondly, this approach could prove useful for predicting treatment response in individual patients with psychotic disorders as demonstrated in Chapter 2, in which we predicted treatment response to a psychotherapeutic intervention. This treatment was specifically tailored to different reasoning biases that we modelled and therefore constituted a promising candidate for treatment response prediction. In future, computational mechanisms may themselves constitute new targets for psychotherapeutic interventions. Advanced understanding of the computational mechanisms underlying behavioural differences possibly culminating

in reasoning biases could also help to expand psychoeducation programs.

Provided that computational quantities like low- and high-level precisions can be associated with specific neurotransmitter systems, it may be possible to employ this approach to predict treatment response to pharmacological interventions as well. However, this relationship is likely complex[113] and may vary across brain regions, whose recruitment depends on the experimental task that is utilised. Reinforcement learning tasks or the social learning task employed in Chapter 1 may be appropriate to probe dopaminergic processes, while cholinergic signaling may be better probed with an oddball task. If a relationship between cholinergic modulation and precisions can be confirmed, the approach outlined in Chapter 3 may be useful for predicting treatment response to antipsychotic medication that affects the cholinergic system such as clozapine or olanzapine. However, it is possible that models, which include more biophysical detail are required to uncover changes in specific neurotransmitter systems, for example conductance-based dynamic causal models[152] (see[164,234] for recent applications).

Lastly, this approach may be used for computational phenotyping to identify more homogeneous subgroups of patients, some of which may display changes in computing low- others in high-level precision. This may also be an alternative explanation for heterogeneous results suggesting either weak[214] or strong priors[181] observed in studies investigating hallucinations outlined in the preceding section. Understanding the heterogeneity in both symptom expression and treatment response across patients and stratifying patients based on pathophysiological mechanisms may increase the chance of identifying effective treatments for these subgroups of patients. For example, if a relationship between high-level precision-weighted PEs and NMDA receptor function can be confirmed,[253] this modelling approach may be used to identify high risk patients that would benefit specifically from pharmacological interventions targeting NMDA receptor-related deficits like d-serine administration. In addition to different subgroups, such computational assays may also be well-suited to identify critical time windows to alter the trajectory of the disease progression. The possibility of such critical time windows for pharmacological interventions has been outlined for example in a recent revision of the glutamate hypothesis.[64]

4.3 Limitations and future directions

There are a number of limitations to the computational approach presented in this thesis. First, especially for clinical applications like treatment response prediction, the reliability of parameters estimates and the specificity of parameter effects with respect to other psychiatric conditions need to be carefully assessed. This will require more challenging and costly repeated-measure designs and studies that include other clinical control groups such as patients with affective disorders, respectively.

Secondly, the HGF focuses on a detailed model of perception, but only considers a very simple mapping from perception to action. There are other modelling approaches that should be explored further, since they take actions more explicitly into account. One such approach

is Q-learning, which assumes that agents learn about Q-values, i.e., the values of state-action pairs,[252,251] or active inference, which proposes to solve the exploration-exploitation dilemma by minimising expected free energy or surprise[24,88,89] (see[4,74] for recent examples). Since some symptoms of schizophrenia are expressed in action choices – for example, anhedonia may manifest as the choice not to go to work or meet with friends – taking action policies explicitly into account may provide additional insight.

Thirdly, hierarchical Bayesian inference – as a reductionist account – fails to provide an explanation for all characteristics of specific symptoms, such as delusional themes[155] and phenomenological aspects that accompany psychotic experiences.[72] Delusions in psychosis often involve recurring themes, for example paranoid beliefs, which also distinguishes them from delusions occurring in other disorders like major depressive disorder, where delusions of guilt are reported more frequently.[180] Recent studies have suggested that broader social and cultural contexts may play a role in shaping both delusions and hallucinations.[139,146,230] For example, Luhrmann and colleagues[146] reported that across three groups of participants, who experienced auditory hallucinations, individuals from the US were more likely to report hearing violent commands than participants from India or Ghana, who more frequently reported positive experiences with their voices. More research involving patients with diverse cultural backgrounds is needed and future models may benefit from taking these contextual factors into account.

Furthermore, as Feyaerts and colleagues[72] recently pointed out, patients often do not describe delusions as reasoned conclusions based on particular experiences (e.g., of aberrant salience); instead, they perceive them as sudden and spontaneous revelations. Delusions are also frequently accompanied by radical changes in the basic structure of human experience, for example alterations in the perception of *time*, *space*, or *causality*.[72] Future models may need to consider these additional aspects of the phenomenology of psychotic experiences.

Additionally, more studies are needed to understand the dynamics of psychosis in individual patients. Schizophrenia and other psychiatric illnesses like depression[258] are not only characterised by dynamic processes at a macro-scale such as the transition from a prodromal to a psychotic state, but also by fluctuations on smaller timescales (e.g., month-to-month or even hour-to-hour). Here, we investigated different groups including high risk individuals, first-episode, early-illness and chronic patients in a cross-sectional manner. While this approach provided important insights into some of the mechanisms underlying these different populations, more studies that employ longitudinal designs and new methods like experience sampling[173] are needed to fully address the dynamic nature of psychosis. These approaches may be invaluable for understanding how symptom dynamics in individual patients predict critical events, including not only transition-to-psychosis, but also relapse and recovery following psychotic episodes. This approach has produced first promising results in depression research[242,258,259] and may provide important insights into psychotic disorders as well.

Lastly, more pharmacological studies are urgently needed to determine the precise relationship between computational mechanisms (e.g., changes in high- and low-level precisions and

PEs) and neurophysiological mechanisms such as reduced pyramidal gain[2] and alterations in specific neurotransmitter systems (see[113,163,203,254] for recent examples). Clarifying this relationship will be crucial to determine the most promising clinical applications for this modelling approach, for example treatment response prediction to medication that modulates specific neurotransmitter systems or for stratifying individual patients into more homogeneous subgroups based on the underlying pathophysiological mechanisms.[223]

4.4 Final remarks

As outlined in the introduction of this thesis, psychotic disorders like schizophrenia can impose a tremendous burden on patients and their families. The field has come a long way from locking patients away in asylums to both psychotherapeutic and pharmacological treatments, which allow some patients to return to a relatively normal life. However, for other patients, treatments are unsuccessful as residual symptoms persist or relapses occur. Some patients experience strong side effects of medication that can in and of themselves drastically reduce life expectancy and quality of life. Thus, there is still a long road ahead to further improve treatment for psychotic disorders.

It is also important to note that patients are still confronted with stigmatisation today, which is deeply ingrained in our culture.[190] This may possibly be rooted in an elusive mind-body dualism, that is still institutionalised in medicine, for example in the distinction between neurology (the study of the nervous system) and psychiatry (the study of the psyche). Fortunately, with the advent of new technologies, this boundary between bodily and mental illnesses has blurred. With the computational approach presented in this thesis, I sought to blur this boundary even further and advance a quantifiable and mathematically rigorous way to understand psychiatric symptoms and link them to biological mechanisms.

In this thesis, I cast several symptoms of psychotic disorders as instances of hierarchical Bayesian inference, including paranoid delusions, reasoning biases and aberrant sensory learning. This approach provided new conceptual insights into the emergence of paranoid delusions and reasoning biases such as increased belief updating in light of disconfirmatory evidence in patients with psychotic disorders. Furthermore, I provided a proof-of-concept that this approach may be clinically useful by predicting treatment response to a psychotherapeutic intervention based on computational model parameters. Lastly, I assessed the biological plausibility of this model by using it to explain EEG amplitude fluctuations during a sensory learning task. Jointly, this work demonstrates that this approach may not only be conceptually and clinically useful, but also biologically plausible. Hopefully, approaches such as the one outlined here will deliver computational assays that may be used as clinical tests to identify specific pathophysiological mechanisms that can be targeted by treatments and lead to de-stigmatisation of mental illnesses and better treatments for psychotic disorders.

Acronyms

A1 primary auditory cortex

AMPA α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid

APS attenuated psychotic symptoms

AUC area under the curve

BAC balanced accuracy

BLIP brief and limited intermittent psychotic symptoms

CBT cognitive-behavioural therapy

CHR individuals at clinical high risk for psychosis

COGIDS cognitive disturbances

COPER cognitive-perceptive basic symptoms

CRT cognitive remediation therapy

DCM dynamic causal modelling

EEG electroencephalography

FEP first-episode psychosis patients

fMRI functional magnetic resonance imaging

GABA gamma-aminobutyric acid

GRD genetic risk and deterioration syndrome

HC healthy controls

HGF Hierarchical Gaussian Filter

IFG inferior frontal gyrus

JTC jumping-to-conclusions bias

M.I.N.I. Mini-International Neuropsychiatric Interview

MCT Metacognitive Training

MMN mismatch negativity

NMDA *N*-methyl-D-aspartate

NPV negative predictive value

PANSS Positive and Negative Syndrome Scale

PCL Paranoia Checklist

PCP phencyclidine

PEs prediction errors

PET positron emission tomography

PPV positive predictive value

PSYRATS Psychotic Symptom Rating Scales

SCZ early-illness schizophrenia patients

SE sensitivity

SIPS Structured Interview for Prodromal Symptoms

sMRI structural magnetic resonance imaging

SP specificity

SPECT single photon emission computerised tomography

SPI-A Schizophrenia Proneness Instrument, adult version

SPI-CY Schizophrenia Proneness Instrument, child and youth version

STG superior temporal gyrus

References

[1] Adams, R. A., Napier, G., Roiser, J. P., Mathys, C., & Gilleen, J. (2018). Attractor-like dynamics in belief updating in schizophrenia. *Journal of Neuroscience*, 38(44), 9471–9485.

[2] Adams, R. A., Pinotsis, D., Tsirlis, K., Unruh, L., Mahajan, A., Horas, A. M., Convertino, L., Summerfelt, A., Sampath, H., Du, X. M., et al. (2022). Computational modeling of electroencephalography and functional magnetic resonance imaging paradigms indicates a consistent loss of pyramidal cell synaptic gain in schizophrenia. *Biological Psychiatry*, 91(2), 202–215.

[3] Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.

[4] Adams, R. A., Vincent, P., Benrimoh, D., Friston, K. J., & Parr, T. (2021). Everything is connected: inference and attractors in delusions. *Schizophrenia Research*.

[5] Allen, N. C., Bagade, S., McQueen, M. B., Ioannidis, J., Kavvoura, F. K., Khoury, M. J., Tanzi, R. E., & Bertram, L. (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics*, 40(7), 827–834.

[6] An, S. K., Kang, J. I., Park, J. Y., Kim, K. R., Lee, S. Y., & Lee, E. (2010). Attribution bias in ultra-high risk for psychosis and first-episode schizophrenia. *Schizophrenia Research*, 118(1-3), 54–61.

[7] Andreou, C., Moritz, S., Veith, K., Veckenstedt, R., & Naber, D. (2014). Dopaminergic modulation of probabilistic reasoning and overconfidence in errors: a double-blind study. *Schizophrenia Bulletin*, 40(3), 558–565.

[8] Andreou, C., Schneider, B. C., Braun, V., Kolbeck, K., Gallinat, J., & Moritz, S. (2015). Dopamine effects on evidence gathering and integration. *Journal of Psychiatry and Neuroscience*, 40(6), 422–428.

[9] Andreou, C., Steinmann, S., Leicht, G., Kolbeck, K., Moritz, S., & Mulert, C. (2018). fMRI correlates of jumping-to-conclusions in patients with delusions: connectivity patterns and effects of metacognitive training. *NeuroImage: Clinical*, 20, 119–127.

[10] Andreou, C., Wittekind, C. E., Fieker, M., Heitz, U., Veckenstedt, R., Bohn, F., & Moritz, S. (2017). Individualized metacognitive therapy for delusions: a randomized controlled rater-blind study. *Journal of Behavior Therapy and Experimental Psychiatry*, 56, 144–151.

[11] Anticevic, A., Hu, X., Xiao, Y., Hu, J., Li, F., Bi, F., Cole, M. W., Savic, A., Yang, G. J., Repovs, G., et al. (2015). Early-course unmedicated schizophrenia patients exhibit elevated prefrontal connectivity associated with longitudinal change. *Journal of Neuroscience*, 35(1), 267–286.

[12] Appelbaum, P. S., Robbins, P. C., & Roth, L. H. (1999). Dimensional approach to delusions: comparison across types and diagnoses. *American Journal of Psychiatry*, 156(12), 1938–1943.

[13] Ashinoff, B. K., Singletary, N. M., Baker, S. C., & Horga, G. (2021). Rethinking delusions: a selective review of delusion research through a computational lens. *Schizophrenia Research*.

[14] Baker, S. C., Konova, A. B., Daw, N. D., & Horga, G. (2019). A distinct inferential mechanism for delusions in schizophrenia. *Brain*, 142(6), 1797–1812.

[15] Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), 1214–1221.

[16] Benamer, N., Marti, F., Lujan, R., Hepp, R., Aubier, T. G., Dupin, A., Frébourg, G., Pons, S., Maskos, U., Faure, P., et al. (2018). GluD1, linked to schizophrenia, controls the burst firing of dopamine neurons. *Molecular Psychiatry*, 23(3), 691–700.

[17] Bendfeldt, K., Smieskova, R., Koutsouleris, N., Klöppel, S., Schmidt, A., Walter, A., Harrisberger, F., Wrege, J., Simon, A., Taschler, B., et al. (2015). Classifying individuals at high-risk for psychosis based on functional brain activity during working memory processing. *NeuroImage: Clinical*, 9, 555–563.

[18] Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N., & Kinderman, P. (2001). Persecutory delusions: a review and theoretical integration. *Clinical Psychology Review*, 21(8), 1143–1192.

[19] Bergstein, M., Weizman, A., & Solomon, Z. (2008). Sense of coherence among delusional patients: Prediction of remission and risk of relapse. *Comprehensive Psychiatry*, 49(3), 288–296.

[20] Berridge, K. C. & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3), 309–369.

[21] Blackwood, N. J., Howard, R. J., Bentall, R. P., & Murray, R. M. (2001). Cognitive neuropsychiatric models of persecutory delusions. *American Journal of Psychiatry*, 158(4), 527–539.

[22] Bodatsch, M., Ruhrmann, S., Wagner, M., Müller, R., Schultze-Lutter, F., Frommann, I., Brinkmeyer, J., Gaebel, W., Maier, W., Klosterkötter, J., et al. (2011). Prediction of psychosis by mismatch negativity. *Biological Psychiatry*, 69(10), 959–966.

[23] Bonner, S. E. & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society*, 27(4-5), 303–345.

[24] Botvinick, M. & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 16(10), 485–488.

[25] Bowie, C. R. & Harvey, P. D. (2006). Cognitive deficits and functional outcome in schizophrenia. *Neuropsychiatric Disease and Treatment*, 2(4), 531–536.

[26] Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.

[27] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[28] Cannon, T. D., Cadenhead, K., Cornblatt, B., Woods, S. W., Addington, J., Walker, E., Seidman, L. J., Perkins, D., Tsuang, M., McGlashan, T., et al. (2008). Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. *Archives of General Psychiatry*, 65(1), 28–37.

[29] Catalan, A., Tognin, S., Kempton, M. J., Stahl, D., de Pablo, G. S., Nelson, B., Pantelis, C., Riecher-Rössler, A., Bressan, R., Barrantes-Vidal, N., et al. (2020). Relationship between jumping to conclusions and clinical outcomes in people at clinical high-risk for psychosis. *Psychological Medicine*, (pp. 1–9).

[30] Chadwick, P. K. (1993). The stepladder to the impossible: a first hand phenomenological account of a schizoaffective psychotic crisis. *Journal of Mental Health*, 2(3), 239–250.

[31] Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., McGrath, J. J., & Whiteford, H. A. (2018). Global epidemiology and burden of schizophrenia: findings from the global burden of disease study 2016. *Schizophrenia Bulletin*, 44(6), 1195–1203.

[32] Charlton, C. E., Lepock, J. R., Hauke, D. J., Mizrahi, R., Kiang, M., & Diaconescu, A. O. (2022). Atypical prediction error learning is associated with psychosis-like symptoms in patients at clinical high risk for psychosis: A computational single-trial analysis of the mismatch negativity. *Under Review*.

[33] Chen, G. M. & Weston, J. K. (1960). The analgesic and anesthetic effect of lN-(l-Phenylcyclohexyl) Piperidine HCl on the monkey. *Anesthesia & Analgesia*, 39(2), 132–137.

[34] Clark, S. R., Schubert, K. O., & Baune, B. T. (2015). Towards indicated prevention of psychosis: using probabilistic assessments of transition risk in psychosis prodrome. *Journal of Neural Transmission*, 122(1), 155–169.

[35] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.

[36] Coid, J. W., Ullrich, S., Kallis, C., Keers, R., Barker, D., Cowden, F., & Stamps, R. (2013). The relationship between delusions and violence: findings from the East London first episode psychosis study. *JAMA Psychiatry*, 70(5), 465–471.

[37] Cole, D. M., Diaconescu, A. O., Pfeiffer, U. J., Brodersen, K. H., Mathys, C. D., Julkowski, D., Ruhrmann, S., Schilbach, L., Tittgemeyer, M., Vogeley, K., et al. (2020). Atypical processing of uncertainty in individuals at risk for psychosis. *NeuroImage: Clinical*, 26, 102239.

[38] Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4), 515–530.

[39] Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: an updated model. *Journal of Psychopharmacology*, 30(11), 1145–1155.

[40] Corlett, P. R., Honey, G. D., Krystal, J. H., & Fletcher, P. C. (2011). Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology*, 36(1), 294–315.

[41] Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R., 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(9), 2387–2400.

[42] Corlett, P. R., Taylor, J., Wang, X.-J., Fletcher, P., & Krystal, J. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, 92(3), 345–369.

[43] Creese, I., Burt, D. R., & Snyder, S. H. (1976). Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science*, 192(4238), 481–483.

[44] Croft, J., Teufel, C., Heron, J., Fletcher, P., David, A. S., Lewis, G., Moutoussis, M., FitzGerald, T. H., Linden, D. E., Thompson, A., et al. (2021). A computational analysis of abnormal belief-updating processes and their association with psychotic experiences and childhood trauma in a UK birth cohort. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

[45] Das, T., Borgwardt, S., Hauke, D. J., Harrisberger, F., Lang, U. E., Riecher-Rössler, A., Palaniyappan, L., & Schmidt, A. (2018). Disorganized gyrification network properties during the transition to psychosis. *JAMA Psychiatry*, 75(6), 613–622.

[46] Daunizeau, J., Adam, V., & Rigoux, L. (2014). Vba: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, 10(1), e1003441.

[47] Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Friston, K. J., & Stephan, K. E. (2010a). Observing the observer (II): deciding when to decide. *PLoS one*, 5(12), e15555.

[48] Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010b). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS one*, 5(12), e15554.

[49] Davies, C., Cipriani, A., Ioannidis, J. P., Radua, J., Stahl, D., Provenzani, U., McGuire, P., & Fusar-Poli, P. (2018). Lack of evidence to favor specific preventive interventions in psychosis: a network meta-analysis. *World Psychiatry*, 17(2), 196–209.

[50] Davis, K. L., Kahn, R. S., Ko, G., & Davidson, M. (1991). Dopamine in schizophrenia: a review and reconceptualization. *The American Journal of Psychiatry*, 148(11), 1474–1486.

[51] Dean, K. & Murray, R. M. (2005). Environmental risk factors for psychosis. *Dialogues in Clinical Neuroscience*, 7(1), 69–80.

[52] Delay, J., Deniker, P., & Harl, J. (1952). Utilisation en thérapeutique psychiatrique d'une phénothiazine d'action centrale élective (4560 RP) [Therapeutic use in psychiatry of phenothiazine of central elective action (4560 RP)]. *Annales Medico-Psychologiques*, 110(21), 112–117.

[53] Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K. E., Heinz, A., & Schlagenhauf, F. (2020). Volatility estimates increase choice switching and relate to prefrontal activity in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(2), 173–183.

[54] Diaconescu, A. O., Hauke, D. J., & Borgwardt, S. (2019). Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Molecular Psychiatry*, 24(9), 1258–1267.

[55] Diaconescu, A. O., Litvak, V., Mathys, C., Kasper, L., Friston, K. J., & Stephan, K. E. (2017a). A computational hierarchy in human cortex. *arXiv: 1709.02323*.

[56] Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, 10(9), e1003810.

[57] Diaconescu, A. O., Mathys, C., Weber, L. A., Kasper, L., Mauer, J., & Stephan, K. E. (2017b). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, 12(4), 618–634.

[58] Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., & Stephan, K. E. (2020). Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, 129(6), 556–569.

[59] Doeller, C. F., Opitz, B., Mecklinger, A., Krick, C., Reith, W., & Schröger, E. (2003). Prefrontal cortex involvement in preattentive auditory deviance detection: neuroimaging and electrophysiological evidence. *NeuroImage*, 20(2), 1270–1282.

[60] Doya, K., Ishii, S., Pouget, A., & Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge: MIT press.

[61] Dudley, R., Taylor, P., Wickham, S., & Hutton, P. (2016). Psychosis, delusions and the "jumping to conclusions" reasoning bias: a systematic review and meta-analysis. *Schizophrenia Bulletin*, 42(3), 652–665.

[62] Durstewitz, D. & Seamans, J. K. (2008). The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. *Biological Psychiatry*, 64(9), 739–749.

[63] Egerton, A., Chaddock, C. A., Winton-Brown, T. T., Bloomfield, M. A., Bhattacharyya, S., Allen, P., McGuire, P. K., & Howes, O. D. (2013). Presynaptic striatal dopamine dysfunction in people at ultra-high risk for psychosis: findings in a second cohort. *Biological Psychiatry*, 74(2), 106–112.

[64] Egerton, A., Grace, A. A., Stone, J., Bossong, M. G., Sand, M., & McGuire, P. (2020). Glutamate in schizophrenia: neurodevelopmental perspectives and drug development. *Schizophrenia Research*, 223, 59–70.

[65] Ehrlichman, R. S., Maxwell, C. R., Majumdar, S., & Siegel, S. J. (2008). Deviance-elicited changes in event-related potentials are attenuated by ketamine in mice. *Journal of Cognitive Neuroscience*, 20(8), 1403–1414.

[66] Erdmann, T. & Mathys, C. (2021). A generative framework for the study of delusions. *Schizophrenia Research*.

[67] Erickson, M. A., Ruffle, A., & Gold, J. M. (2016). A meta-analysis of mismatch negativity in schizophrenia: from clinical risk to disease specificity and progression. *Biological Psychiatry*, 79(12), 980–987.

[68] Esterberg, M. L. & Compton, M. T. (2009). The psychosis continuum and categorical versus dimensional diagnostic approaches. *Current Psychiatry Reports*, 11(3), 179–184.

[69] Fear, C. F. & Healy, D. (1997). Probabilistic reasoning in obsessive–compulsive and delusional disorders. *Psychological Medicine*, 27(1), 199–208.

[70] Feinberg, I. (1978). Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophrenia Bulletin*, 4(4), 636–640.

[71] Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.

[72] Feyaerts, J., Henriksen, M. G., Vanheule, S., Myin-Germeys, I., & Sass, L. A. (2021). Delusions beyond beliefs: a critical overview of diagnostic, aetiological, and therapeutic schizophrenia research from a clinical-phenomenological perspective. *The Lancet Psychiatry*, 8(3), 237–249.

[73] First, M. B. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders (SCID), Research Version, Patient Edition with Psychotic Screen*. New York: New York State Psychiatric Institute, Biometrics Research.

[74] FitzGerald, T. H., Schwartenbeck, P., Moutoussis, M., Dolan, R. J., & Friston, K. (2015). Active inference, evidence accumulation, and the urn task. *Neural Computation*, 27(2), 306–328.

[75] Flandin, G. & Friston, K. J. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, 40(7), 2052–2054.

[76] Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48–58.

[77] Fornito, A. & Bullmore, E. T. (2015). Reconciling abnormalities of brain network structure and function in schizophrenia. *Current Opinion in Neurobiology*, 30, 44–50.

[78] Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience*, 33(8), 3602–3611.

[79] Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., Harrison, S. J., Heinzle, J., Iglesias, S., Kasper, L., et al. (2021). TAPAS: an open-source software package for translational neuromodeling and computational psychiatry. *Frontiers in Psychiatry*, 12, 857.

[80] Freeman, D. (2016). Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7), 685–692.

[81] Freeman, D. & Garety, P. (2014). Advances in understanding and treating persecutory delusions: a review. *Social Psychiatry and Psychiatric Epidemiology*, 49(8), 1179–1189.

[82] Freeman, D. & Garety, P. A. (2000). Comments on the content of persecutory delusions: does the definition need clarification? *British Journal of Clinical Psychology*, 39(4), 407–414.

[83] Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Fowler, D., Kuipers, E., Ray, K., & Dunn, G. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *The British Journal of Psychiatry*, 186(5), 427–435.

[84] Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, 41(4), 331–347.

[85] Freeman, D., Startup, H., Dunn, G., Wingham, G., Černis, E., Evans, N., Lister, R., Pugh, K., Cordwell, J., & Kingdon, D. (2014). Persecutory delusions and psychological well-being. *Social Psychiatry and Psychiatric Epidemiology*, 49(7), 1045–1050.

[86] Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836.

[87] Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia Research*, 176(2-3), 83–94.

[88] Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology*, 100(1-3), 70–87.

[89] Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 7, 598.

[90] Friston, K. J. (1998). The disconnection hypothesis. *Schizophrenia Research*, 30(2), 115–125.

[91] Friston, K. J., Frith, C. D., et al. (1995). Schizophrenia: a disconnection syndrome. *Clinical Neuroscience*, 3(2), 89–97.

[92] Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302.

[93] Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417–458.

[94] Frith, C. D. & Done, D. J. (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine*, 19(2), 359–363.

[95] Frith, C. D. & Friston, K. J. (2013). False perceptions and false beliefs: understanding schizophrenia. *Neurosciences and the Human Person: New Perspectives on Human Activities*, 121, 1–15.

[96] Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., Keshavan, M., Wood, S., Ruhrmann, S., Seidman, L. J., et al. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry*, 70(1), 107–120.

[97] Garety, P. (1991). Reasoning and delusions. *The British Journal of Psychiatry*, 159(S14), 14–18.

[98] Garety, P., Hemsley, D., Wessely, S., et al. (1991). Reasoning in deluded schizophrenic and paranoid patients. *Journal of Nervous and Mental Disease*, 179(4), 194–201.

[99] Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology*, 120(3), 453–463.

[100] Geisler, W. S. & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1420), 419–448.

[101] Grace, A. A. (1993). Cortical regulation of subcortical dopamine systems and its possible relevance to schizophrenia. *Journal of Neural Transmission*, 91(2), 111–134.

[102] Gu, Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience*, 111(4), 815–835.

[103] Haddock, G., McCarron, J., Tarrier, N., & Faragher, E. (1999). Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). *Psychological Medicine*, 29(4), 879–889.

[104] Hamilton, H. K., Roach, B. J., Bachman, P. M., Belger, A., Carrion, R. E., Duncan, E., Johannesen, J. K., Light, G. A., Niznikiewicz, M. A., Addington, J., et al. (2019). Association between P300 responses to auditory oddball stimuli and clinical outcomes in the psychosis risk syndrome. *JAMA Psychiatry*, 76(11), 1187–1197.

[105] Hauke, D. J., Roth, V., Karvelis, P., Adams, R. A., Moritz, S., Borgwardt, S., Diaconescu, A. O., & Andreou, C. (2022). Increased belief instability in psychosis predicts treatment response to metacognitive training. *Schizophrenia Bulletin*, 48(4), 826–838.

[106] Hauke, D. J., Schmidt, A., Studerus, E., Andreou, C., Riecher-Rössler, A., Radua, J., Kambeitz, J., Ruef, A., Dwyer, D. B., Kambeitz-Ilankovic, L., et al. (2021). Multimodal prognosis of negative symptom severity in individuals at increased risk of developing psychosis. *Translational Psychiatry*, 11(1), 1–11.

[107] Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *The Lancet Psychiatry*, 4(4), 295–301.

[108] Homayoun, H. & Moghaddam, B. (2007). NMDA receptor hypofunction produces opposite effects on prefrontal cortex interneurons and pyramidal neurons. *Journal of Neuroscience*, 27(43), 11496–11500.

[109] Howes, O., Bose, S., Turkheimer, F., Valli, I., Egerton, A., Stahl, D., Valmaggia, L., Allen, P., Murray, R., & McGuire, P. (2011). Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. *Molecular Psychiatry*, 16(9), 885–886.

[110] Howes, O. D. & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III—the final common pathway. *Schizophrenia Bulletin*, 35(3), 549–562.

[111] Howes, O. D., Montgomery, A. J., Asselin, M.-C., Murray, R. M., Valli, I., Tabraham, P., Bramon-Bosch, E., Valmaggia, L., Johns, L., Broome, M., et al. (2009). Elevated striatal dopamine function linked to prodromal signs of schizophrenia. *Archives of General Psychiatry*, 66(1), 13–20.

[112] Huq, S., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, 40(4), 801–812.

[113] Iglesias, S., Kasper, L., Harrison, S. J., Manka, R., Mathys, C., & Stephan, K. E. (2021). Cholinergic and dopaminergic effects on prediction error and uncertainty responses during sensory associative learning. *NeuroImage*, 226, 117590.

[114] Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, 80(2), 519–530.

[115] Iglesias, S., Tomiello, S., Schneebeli, M., & Stephan, K. E. (2017). Models of neuromodulation for computational psychiatry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(3), e1420.

[116] Jardri, R. & Deneve, S. (2013). Circular inferences in schizophrenia. *Brain*, 136(11), 3227–3241.

[117] Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, 8(1), 1–13.

[118] Jaspers, K. (1913). *Allgemeine Psychopathologie*. Berlin: J. Springer.

[119] Javitt, D. (2004). Glutamate as a therapeutic target in psychiatric disorders. *Molecular Psychiatry*, 9(11), 984–997.

[120] Javitt, D. C., Steinschneider, M., Schroeder, C. E., & Arezzo, J. C. (1996). Role of cortical N-methyl-D-aspartate receptors in auditory sensory memory and mismatch negativity generation: implications for schizophrenia. *Proceedings of the National Academy of Sciences*, 93(21), 11962–11967.

[121] Javitt, D. C. & Zukin, S. R. (1991). Recent advances in the phencyclidine model of schizophrenia. *The American Journal of Psychiatry*.

[122] Javitt, D. C., Zukin, S. R., Heresco-Levy, U., & Umbricht, D. (2012). Has an angel shown the way? Etiological and therapeutic implications of the PCP/NMDA model of schizophrenia. *Schizophrenia Bulletin*, 38(5), 958–966.

[123] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620–630.

[124] Kantrowitz, J. T., Woods, S. W., Petkova, E., Cornblatt, B., Corcoran, C. M., Chen, H., Silipo, G., & Javitt, D. C. (2015). D-serine for the treatment of negative symptoms in individuals at clinical high risk of schizophrenia: a pilot, double-blind, placebo-controlled, randomised parallel group mechanistic proof-of-concept trial. *The Lancet Psychiatry*, 2(5), 403–412.

[125] Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry*, 160(1), 13–23.

[126] Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 13(2), 261–276.

[127] Kendler, K. S., Gallagher, T. J., Abelson, J. M., & Kessler, R. C. (1996). Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a us community sample: the national comorbidity survey. *Archives of General Psychiatry*, 53(11), 1022–1031.

[128] Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, 4(11), e1000209.

[129] Kiebel, S. J. & Friston, K. J. (2004). Statistical parametric mapping for event-related potentials: I. Generic considerations. *NeuroImage*, 22(2), 492–502.

[130] King, D. J., Hodgekins, J., Chouinard, P. A., Chouinard, V.-A., & Sperandio, I. (2017). A review of abnormalities in the perception of visual illusions in schizophrenia. *Psychonomic Bulletin & Review*, 24(3), 734–751.

[131] King, R., Barchas, J. D., & Huberman, B. (1984). Chaotic behavior in dopamine neurodynamics. *Proceedings of the National Academy of Sciences*, 81(4), 1244–1247.

[132] King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, 308(5718), 78–83.

[133] Klosterkötter, J., Hellmich, M., Steinmeyer, E. M., & Schultze-Lutter, F. (2001). Diagnosing schizophrenia in the initial prodromal phase. *Archives of General Psychiatry*, 58(2), 158–164.

[134] Knowles, R., McCarthy-Jones, S., & Rowse, G. (2011). Grandiose delusions: a review and theoretical integration of cognitive and affective perspectives. *Clinical Psychology Review*, 31(4), 684–696.

[135] Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244–247.

[136] Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Möller, H.-J., & Riecher-Rössler, A. (2012). Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophrenia Bulletin*, 38(6), 1234–1246.

[137] Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., et al. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry*, 66(7), 700–712.

[138] Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., Von Saldern, S., Cabral, C., Reiser, M., Falkai, P., et al. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin*, 41(2), 471–482.

[139] Larøi, F., Luhrmann, T. M., Bell, V., Christian Jr, W. A., Deshpande, S., Fernyhough, C., Jenkins, J., & Woods, A. (2014). Culture and hallucinations: overview and future directions. *Schizophrenia Bulletin*, 40(Suppl_4), S213–S220.

[140] Leanza, L., Studerus, E., Bozikas, V. P., Moritz, S., & Andreou, C. (2020). Moderators of treatment efficacy in individualized metacognitive training for psychosis (MCT+). *Journal of Behavior Therapy and Experimental Psychiatry*, 68, 101547.

[141] Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7), 1434–1448.

[142] Lehrl, S., Triebig, G., & Fischer, B. (1995). Multiple choice vocabulary test MWT as a valid and short test to estimate premorbid intelligence. *Acta Neurologica Scandinavica*, 91(5), 335–345.

[143] Lieberman, J., Kane, J., & Alvir, J. (1987). Provocative tests with psychostimulant drugs in schizophrenia. *Psychopharmacology*, 91(4), 415–433.

[144] Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological Bulletin*, 90(1), 125.

[145] Luby, E. D., Gottlieb, J. S., Cohen, B. D., Rosenbaum, G., & Domino, E. F. (1962). Model psychoses and schizophrenia. *American Journal of Psychiatry*, 119(1), 61–67.

[146] Luhrmann, T. M., Padmavati, R., Tharoor, H., & Osei, A. (2015). Differences in voice-hearing experiences of people with psychosis in the USA, India and Ghana: interview-based study. *The British Journal of Psychiatry*, 206(1), 41–44.

[147] Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., & Bullmore, E. (2010). Functional connectivity and brain networks in schizophrenia. *Journal of Neuroscience*, 30(28), 9477–9487.

[148] Maia, T. V. & Frank, M. J. (2017). An integrative perspective on the role of dopamine in schizophrenia. *Biological psychiatry*, 81(1), 52–66.

[149] Marco-Pallarés, J., Grau, C., & Ruffini, G. (2005). Combined ICA-LORETA analysis of mismatch negativity. *NeuroImage*, 25(2), 471–477.

[150] Markov, N. T., Ercsey-Ravasz, M. M., Ribeiro Gomes, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M.-A., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex*, 24(1), 17–36.

[151] Marr, D. (1982). *Vision*. San Francisco: Freeman.

[152] Marreiros, A. C., Kiebel, S. J., Daunizeau, J., Harrison, L. M., & Friston, K. J. (2009). Population dynamics under the Laplace assumption. *NeuroImage*, 44(3), 701–714.

[153] Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5, 39.

[154] 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. *Frontiers in Human Neuroscience*, 8, 825.

[155] McKay, R. (2019). Measles, magic and misidentifications: a defence of the two-factor theory of delusions. *Cognitive Neuropsychiatry*, 24(3), 183–190.

[156] McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. *Schizophrenia Bulletin*, 43(2), 344–354.

[157] Miller, T. J., McGlashan, T. H., Rosen, J. L., Cadenhead, K., Ventura, J., McFarlane, W., Perkins, D. O., Pearlson, 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. *Schizophrenia Bulletin*, 29(4), 703–715.

[158] 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. *American Journal of Psychiatry*, 159(5), 863–865.

[159] Mittal, V. A., Ellman, L. M., & Cannon, T. D. (2008). Gene-environment interaction and covariation in schizophrenia: the role of obstetric complications. *Schizophrenia Bulletin*, 34(6), 1083–1094.

[160] Moghaddam, B. (2003). Bringing order to the glutamate chaos in schizophrenia. *Neuron*, 40(5), 881–884.

[161] Moghaddam, B. (2004). Targeting metabotropic glutamate receptors for treatment of the cognitive symptoms of schizophrenia. *Psychopharmacology*, 174(1), 39–44.

[162] Moghaddam, B. & Javitt, D. (2012). From revolution to evolution: the glutamate hypothesis of schizophrenia and its implication for treatment. *Neuropsychopharmacology*, 37(1), 4–15.

[163] Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., & Friston, K. J. (2013). Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience*, 33(19), 8227–8236.

[164] Moran, R. J., Symmonds, M., Stephan, K. E., Friston, K. J., & Dolan, R. J. (2011). An in vivo assay of synaptic function mediating human cognition. *Current Biology*, 21(15), 1320–1325.

[165] Moritz, S., Veckenstedt, R., Bohn, F., Hottenrott, B., Scheu, F., Randjbar, S., Aghotor, J., Köther, U., Woodward, T. S., Treszl, A., et al. (2013). Complementary group metacognitive training (MCT) reduces delusional ideation in schizophrenia. *Schizophrenia Research*, 151(1-3), 61–69.

[166] Moritz, S., Veckenstedt, R., Hottenrott, B., Woodward, T. S., Randjbar, S., & Lincoln, T. M. (2010a). Different sides of the same coin? Intercorrelations of cognitive biases in schizophrenia. *Cognitive Neuropsychiatry*, 15(4), 406–421.

[167] Moritz, S., Veckenstedt, R., Randjbar, S., Vitzthum, F., & Woodward, T. (2011). Antipsychotic treatment beyond antipsychotics: metacognitive intervention for schizophrenia patients improves delusional symptoms. *Psychological Medicine*, 41(9), 1823–1832.

[168] Moritz, S., Vitzthum, F., Randjbar, S., Veckenstedt, R., & Woodward, T. S. (2010b). Detecting and defusing cognitive traps: metacognitive intervention in schizophrenia. *Current Opinion in Psychiatry*, 23(6), 561–569.

[169] Moritz, S. & Woodward, T. S. (2007). Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. *Current Opinion in Psychiatry*, 20(6), 619–625.

[170] Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of jumping-to-conclusions bias in delusional patients. *Cognitive Neuropsychiatry*, 16(5), 422–447.

[171] Murray, G., Corlett, P., Clark, L., Pessiglione, M., Blackwell, A., Honey, G., Jones, P., Bullmore, E., Robbins, T., & Fletcher, P. (2008). Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Molecular Psychiatry*, 13(3), 267–276.

[172] Murray, R. M. & Lewis, S. W. (1987). Is schizophrenia a neurodevelopmental disorder? *British Medical Journal (Clinical Research Ed.)*, 295(6600), 681–682.

[173] Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & Van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological Medicine*, 39(9), 1533–1547.

[174] Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology*, 118(12), 2544–2590.

[175] Nakao, K., Jeevakumar, V., Jiang, S. Z., Fujita, Y., Diaz, N. B., Pretell Annan, C. A., Eskow Jaunarajs, K. L., Hashimoto, K., Belforte, J. E., & Nakazawa, K. (2019). Schizophrenia-like dopamine release abnormalities in a mouse model of NMDA receptor hypofunction. *Schizophrenia Bulletin*, 45(1), 138–147.

[176] Opitz, B., Rinne, T., Mecklinger, A., Von Cramon, D. Y., & Schröger, E. (2002). Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. *NeuroImage*, 15(1), 167–174.

[177] Perez, V. B., Woods, S. W., Roach, B. J., Ford, J. M., McGlashan, T. H., Srihari, V. H., & Mathalon, D. H. (2014). Automatic auditory processing deficits in schizophrenia and clinical high-risk patients: forecasting psychosis risk with mismatch negativity. *Biological Psychiatry*, 75(6), 459–469.

[178] Peters, E. & Garety, P. (2006). Cognitive functioning in delusions: a longitudinal analysis. *Behaviour Research and Therapy*, 44(4), 481–514.

[179] Phillips, L. D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3), 346––354.

[180] Picardi, A., Fonzi, L., Pallagrosi, M., Gigantesco, A., & Biondi, M. (2018). Delusional themes across affective and non-affective psychoses. *Frontiers in Psychiatry*, 9, 132.

[181] Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351), 596–600.

[182] Raihani, N. J. & Bell, V. (2017). Paranoia and the social representation of others: a large-scale game theory approach. *Scientific Reports*, 7(1), 1–9.

[183] Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

[184] Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *Elife*, 9, e56345.

[185] Rescorla, R. A. & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York: Appleton Century Crofts.

[186] Rigoux, L., Stephan, K., Friston, K., & Daunizeau, J. (2014). Bayesian model selection for group studies — revisited. *NeuroImage*, 84, 971–985.

[187] Ross, R. M., McKay, R., Coltheart, M., & Langdon, R. (2015). Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophrenia Bulletin*, 41(5), 1183–1191.

[188] Salvatore, G., Lysaker, P. H., Popolo, R., Procacci, M., Carcione, A., & Dimaggio, G. (2012). Vulnerable self, poor understanding of others' minds, threat anticipation and cognitive biases as triggers for delusional experience in schizophrenia: a theoretical model. *Clinical Psychology & Psychotherapy*, 19(3), 247–259.

[189] Sartorius, N., Jablensky, A., Korten, A., Ernberg, G., Anker, M., Cooper, J. E., & Day, R. (1986). Early manifestations and first-contact incidence of schizophrenia in different cultures: a preliminary report on the initial evaluation phase of the who collaborative study on determinants of outcome of severe mental disorders. *Psychological Medicine*, 16(4), 909–928.

[190] Scarf, D., Zimmerman, H., Winter, T., Boden, H., Graham, S., Riordan, B. C., & Hunter, J. A. (2020). Association of viewing the films joker or terminator: dark fate with prejudice toward individuals with mental illness. *JAMA Network Open*, 3(4), e203423–e203423.

[191] Scarr, E., Cowie, T., Kanellakis, S., Sundram, S., Pantelis, C., & Dean, B. (2009). Decreased cortical muscarinic receptors define a subgroup of subjects with schizophrenia. *Molecular Psychiatry*, 14(11), 1017–1023.

[192] Scarr, E., Craig, J., Cairns, M., Seo, M., Galati, J., Beveridge, N., Gibbons, A., Juzva, S., Weinrich, B., Parkinson-Bates, M., et al. (2013). Decreased cortical muscarinic M1 receptors in schizophrenia are associated with changes in gene promoter methylation, mRNA and gene targeting microRNA. *Translational Psychiatry*, 3(2), e230–e230.

[193] Scarr, E., Hopper, S., Vos, V., Seo, M. S., Everall, I. P., Aumann, T. D., Chana, G., & Dean, B. (2018). Low levels of muscarinic M1 receptor–positive neurons in cortical layers III and V in brodmann areas 9 and 17 from individuals with schizophrenia. *Journal of Psychiatry and Neuroscience*, 43(5), 338–346.

[194] Schmack, K., de Castro, A. G.-C., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., Heinz, A., Petrovic, P., & Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience*, 33(34), 13701–13712.

[195] Schmack, K., Rothkirch, M., Priller, J., & Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human Brain Mapping*, 38(4), 1767–1779.

[196] Schmack, K., Schnack, A., Priller, J., & Sterzer, P. (2015). Perceptual instability in schizophrenia: Probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophrenia Research: Cognition*, 2(2), 72–77.

[197] Schmidt, A., Antoniades, M., Allen, P., Egerton, A., Chaddock, C. A., Borgwardt, S., Fusar-Poli, P., Roiser, J. P., Howes, O., & McGuire, P. (2017a). Longitudinal alterations in motivational salience processing in ultra-high-risk subjects for psychosis. *Psychological Medicine*, 47(2), 243–254.

[198] Schmidt, A., Cappucciati, M., Radua, J., Rutigliano, G., Rocchetti, M., Dell'Osso, L., Politi, P., Borgwardt, S., Reilly, T., Valmaggia, L., et al. (2017b). Improving prognostic accuracy in subjects at clinical high risk for psychosis: systematic review of predictive models and meta-analytical sequential testing simulation. *Schizophrenia Bulletin*, 43(2), 375–388.

[199] Schmidt, A., Palaniyappan, L., Smieskova, R., Simon, A., Riecher-Rössler, A., Lang, U. E., Fusar-Poli, P., McGuire, P., & Borgwardt, S. J. (2016). Dysfunctional insular connectivity during reward prediction in patients with first-episode psychosis. *Journal of Psychiatry and Neuroscience*, 41(6), 367–376.

[200] 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(9), 903–912.

[201] Schmidt, A., Smieskova, R., Simon, A., Allen, P., Fusar-Poli, P., McGuire, P. K., Bendfeldt, K., Aston, J., Lang, U. E., Walter, M., et al. (2014). Abnormal effective connectivity and psychopathological symptoms in the psychosis high-risk state. *Journal of Psychiatry and Neuroscience*, 39(4), 239–248.

[202] Schmidt, K. H. & P., M. (1992). *Wortschatztest*. Göttingen: Hogrefe.

[203] Schöbi, D., Homberg, F., Frässle, S., Endepols, H., Moran, R. J., Friston, K. J., Tittgemeyer, M., Heinzle, J., & Stephan, K. E. (2021). Model-based prediction of muscarinic receptor function from auditory mismatch negativity responses. *NeuroImage*, 237, 118096.

[204] Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2), 241–263.

[205] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599.

[206] Schultze-Lutter, F. (2009). Subjective symptoms of schizophrenia in research and the clinic: the basic symptom concept. *Schizophrenia Bulletin*, 35(1), 5–8.

[207] Schultze-Lutter, F., Addington, J., Ruhrmann, S., & Klosterkötter, J. (2007). *Schizophrenia Proneness Instrument, adult version (SPI-A)*. Rome: Giovanni Fioriti.

[208] Schultze-Lutter, F. & Koch, E. (2010). *Schizophrenia Proneness Instrument: child and youth version (SPI-CY)*. Rome: Giovanni Fioriti.

[209] Seeman, P. & Lee, T. (1975). Antipsychotic drugs: direct correlation between clinical potency and presynaptic action on dopamine neurons. *Science*, 188(4194), 1217–1219.

[210] Seeman, P., Lee, T., Chau-Wong, M., & Wong, K. (1976). Antipsychotic drug doses and neuroleptic/dopamine receptors. *Nature*, 261(5562), 717–719.

[211] Selten, J.-P., Van Der Ven, E., & Termorshuizen, F. (2020). Migration and psychosis: a meta-analysis of incidence studies. *Psychological Medicine*, 50(2), 303–313.

[212] Shaner, A. (1999). Delusions, superstitious conditioning and chaotic dopamine neurodynamics. *Medical Hypotheses*, 52(2), 119–123.

[213] Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G. C., et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, 59(20), 22–33.

[214] Shergill, S. S., White, T. P., Joyce, D. W., Bays, P. M., Wolpert, D. M., & Frith, C. D. (2014). Functional magnetic resonance imaging of impaired sensory prediction in schizophrenia. *JAMA psychiatry*, 71(1), 28–35.

[215] Shiga, T., Horikoshi, S., Kanno, K., Kanno-Nozaki, K., Hikita, M., Itagaki, S., Miura, I., & Yabe, H. (2020). Plasma levels of dopamine metabolite correlate with mismatch negativity in patients with schizophrenia. *Psychiatry and Clinical Neurosciences*, 74(5), 289–293.

[216] Siafis, S., Tzachanis, D., Samara, M., & Papazisis, G. (2018). Antipsychotic drugs: from receptor-binding profiles to metabolic side effects. *Current Neuropharmacology*, 16(8), 1210–1223.

[217] Smieskova, R., Roiser, J. P., Chaddock, C. A., Schmidt, A., Harrisberger, F., Bendfeldt, K., Simon, A., Walter, A., Fusar-Poli, P., McGuire, P. K., et al. (2015). Modulation of motivational salience processing during the early stages of psychosis. *Schizophrenia Research*, 166(1-3), 17–23.

[218] Snyder, S. H. (1976). The dopamine hypothesis of schizophrenia: focus on the dopamine receptor. *The American Journal of Psychiatry*, 133(2), 197–202.

[219] Speechley, W. J., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience*, 35(1), 7–17.

[220] Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112, 92–97.

[221] Stephan, K. E., Baldeweg, T., & Friston, K. J. (2006). Synaptic plasticity and dysconnection in schizophrenia. *Biological Psychiatry*, 59(10), 929–939.

[222] Stephan, K. E., Friston, K. J., & Frith, C. D. (2009a). Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin*, 35(3), 509–527.

[223] Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015). Translational perspectives for computational neuroimaging. *Neuron*, 87(4), 716–732.

[224] Stephan, K. E. & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25, 85–92.

[225] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009b). Bayesian model selection for group studies. *NeuroImage*, 46(4), 1004–1017.

[226] Stephan, K. E., Schlagenhauf, F., Huys, Q. J., Raman, S., Aponte, E. A., Brodersen, K. H., Rigoux, L., Moran, R. J., Daunizeau, J., Dolan, R. J., et al. (2017). Computational neuroimaging strategies for single patient predictions. *NeuroImage*, 145, 180–199.

[227] 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. *Biological Psychiatry*, 84(9), 634–643.

[228] Strube, W., Marshall, L., Quattrocchi, G., Little, S., Cimpianu, C. L., Ulbrich, M., Schneider-Axmann, T., Falkai, P., Hasan, A., & Bestmann, S. (2020). Glutamatergic contribution to probabilistic reasoning and jumping to conclusions in schizophrenia: a double-blind, randomized experimental trial. *Biological Psychiatry*, 88(9), 687–697.

[229] Stuke, H., Weilnhammer, V. A., Sterzer, P., & Schmack, K. (2019). Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. *Schizophrenia Bulletin*, 45(1), 80–86.

[230] Suhail, K. & Cochrane, R. (2002). Effect of culture and environment on the phenomenology of delusions and hallucinations. *International Journal of Social Psychiatry*, 48(2), 126–138.

[231] Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., Litman, L., Robinson, J., Moss, A. J., Taylor, J. R., et al. (2021). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*, 5(9), 1190–1202.

[232] Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts, Amherst, MA, USA.

[233] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1), 9–44.

[234] Symmonds, M., Moran, C. H., Leite, M. I., Buckley, C., Irani, S. R., Stephan, K. E., Friston, K. J., & Moran, R. J. (2018). Ion channels in EEG: isolating channel dysfunction in NMDA receptor antibody encephalitis. *Brain*, 141(6), 1691–1702.

[235] Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Frontiers in Psychology*, 4, 127.

[236] Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., & Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Current Biology*, 19(15), 1274–1277.

[237] Tien, A. Y. (1991). Distribution of hallucinations in the population. *Social Psychiatry and Psychiatric Epidemiology*, 26(6), 287–292.

[238] Todd, J., Harms, L., Michie, P., & Schall, U. (2013). Mismatch negativity: translating the potential. *Frontiers in Psychiatry*, 4, 171.

[239] Tripoli, G., Quattrone, D., Ferraro, L., Gayer-Anderson, C., Rodriguez, V., La Cascia, C., La Barbera, D., Sartorio, C., Seminerio, F., Tarricone, I., et al. (2021). Jumping to conclusions, general intelligence, and psychosis liability: findings from the multi-centre EU-GEI case-control study. *Psychological Medicine*, 51(4), 623–633.

[240] Uhlenbeck, G. E. & Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical Review*, 36(5), 823.

[241] Valli, I., Marquand, A. F., Mechelli, A., Raffin, M., Allen, P., Seal, M. L., & McGuire, P. (2016). Identifying individuals at high risk of psychosis: predictive utility of support vector machine using structural and functional MRI data. *Frontiers in Psychiatry*, 7, 52.

[242] van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., et al. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1), 87–92.

[243] van den Heuvel, M. P., Sporns, O., Collin, G., Scheewe, T., Mandl, R. C., Cahn, W., Goñi, J., Pol, H. E. H., & Kahn, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry*, 70(8), 783–792.

[244] van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel, B., van Harten, P. N., Valmaggia, L., de Hert, M., Cuijpers, A., & Wiersma, D. (2006). The five-factor model of the Positive and Negative Syndrome Scale II: A ten-fold cross-validation of a revised model. *Schizophrenia Research*, 85(1), 280–287.

[245] van Gogh, V. (1888, July 9/10). [Letter from Vincent van Gogh to Theo van Gogh]. Accessible online at: `https://vangoghletters.org/vg/letters/let638/letter.html`.

[246] van Oosterhout, B., Smit, F., Krabbendam, L., Castelein, S., Staring, A., & Van der Gaag, M. (2016). Metacognitive training for schizophrenia spectrum patients: a meta-analysis on outcome studies. *Psychological Medicine*, 46(1), 47–57.

[247] van Os, J., Rutten, B. P., & Poulton, R. (2008). Gene-environment interactions in schizophrenia: review of epidemiological findings and future directions. *Schizophrenia Bulletin*, 34(6), 1066–1082.

[248] Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-analysis of the association of urbanicity with schizophrenia. *Schizophrenia Bulletin*, 38(6), 1118–1123.

[249] von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.

[250] Walter, A., Suenderhauf, C., Smieskova, R., Lenz, C., Harrisberger, F., Schmidt, A., Vogel, T., Lang, U. E., Riecher-Rössler, A., Eckert, A., et al. (2016). Altered insular function during aberrant salience processing in relation to the severity of psychotic symptoms. *Frontiers in Psychiatry*, 7, 189.

[251] Watkins, C. J. & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3), 279–292.

[252] Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, United Kingdom.

[253] Weber, L. A., Diaconescu, A. O., Mathys, C., Schmidt, A., Kometer, M., Vollenweider, F., & Stephan, K. E. (2020). Ketamine affects prediction errors about statistical regularities: a computational single-trial analysis of the mismatch negativity. *Journal of Neuroscience*, 40(29), 5658–5668.

[254] Weber, L. A., Tomiello, S., Schöbi, D., Wellstein, K. V., Müller, D., Iglesias, S., & Stephan, K. (2022). Auditory mismatch responses are differentially sensitive to changes in muscarinic acetylcholine versus dopamine receptor function. *eLife*, 11, e74835.

[255] Wechsler, D. (1981). *Wechsler adult intelligence scale-revised (WAIS-R)*. San Antonio: Psychological Corporation.

[256] Weinberger, D. R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. *Archives of General Psychiatry*, 44(7), 660–669.

[257] Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E. A., Ullrich, J., & Stephan, K. E. (2020). Inflexible social inference in individuals with subclinical persecutory delusional tendencies. *Schizophrenia Research*, 215, 344–351.

[258] Wichers, M. (2014). The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, 44(7), 1349–1360.

[259] Wichers, M., Groot, P. C., Psychosystems, E., Group, E., et al. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, 85(2), 114–116.

[260] Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition*, 61, 129–147.

[261] Winton-Brown, T. T., Fusar-Poli, P., Ungless, M. A., & Howes, O. D. (2014). Dopaminergic basis of salience dysregulation in psychosis. *Trends in Neurosciences*, 37(2), 85–94.

[262] Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, 28(4), 605–617.

[263] Woodward, T. S., Munz, M., LeClerc, C., & Lecomte, T. (2009). Change in delusions is associated with change in "jumping to conclusions". *Psychiatry Research*, 170(2-3), 124–127.

[264] Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1), 58–73.

[265] Xu, F. & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3), 288–297.

[266] Yaniv, I. & Kleinberger, E. (2000). Advice taking in decision making: egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260–281.

[267] Young, H. F. & Bentall, R. P. (1997). Probabilistic reasoning in deluded, depressed and normal subjects: effects of task difficulty and meaningful versus non-meaningful material. *Psychological Medicine*, 27(2), 455–465.

[268] Zhou, Z., Zhu, H., & Chen, L. (2013). Effect of aripiprazole on mismatch negativity (MMN) in schizophrenia. *PLoS One*, 8(1), e52186.