JNeurosci THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Attractor-like dynamics in belief updating in schizophrenia

Rick A Adams[1,2], Gary Napier[1], Jonathan P Roiser[1], Christoph Mathys[3,4,5] and James Gilleen[6,7]

[1]Institute of Cognitive Neuroscience, UCL, 17 Queen Square, London, WC1N 3AZ, UK

[2]Division of Psychiatry, UCL, 6th floor, 149 Tottenham Court Road, London, W1T 7NF, UK

[3]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy

[4]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

[5]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, 10-12 Russell Square, London, WC1B 5EH, UK

[6]Department of Psychology, University of Roehampton, London, SE15 4JD.

[7]Department of Psychosis Studies; Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, SE5 8AF.

DOI: 10.1523/JNEUROSCI.3163-17.2018

Received: 2 November 2017

Revised: 3 May 2018

Accepted: 27 June 2018

Published: 5 September 2018

Author contributions: R.A.A., J.R., C.M., and J.G. designed research; R.A.A., G.N., and J.G. performed research; R.A.A. and G.N. analyzed data; R.A.A. wrote the first draft of the paper; R.A.A., G.N., J.R., C.M., and J.G. wrote the paper; C.M. contributed unpublished reagents/analytic tools.

Conflict of Interest: The authors declare no competing financial interests.

The authors are very grateful to Dr Emmanuelle Peters for providing them with dataset 1. Dr Rick Adams is funded by the Academy of Medical Sciences (AMS-SGCL13-Adams) and the National Institute of Health Research (CL-2013-18-003). JG was supported in his contribution to this project by the British Academy

Correspondence should be addressed to corresponding author: rick.adams@ucl.ac.uk

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.3163-17.2018

Attractor-like dynamics in belief updating in schizophrenia

Rick A Adams[1,2*†], Gary Napier[1†], Jonathan P Roiser[1], Christoph Mathys[3,4,5¶], James Gilleen[6,7¶]

[1]Institute of Cognitive Neuroscience, UCL, 17 Queen Square, London, WC1N 3AZ, UK

[2]Division of Psychiatry, UCL, 6th floor, 149 Tottenham Court Road, London, W1T 7NF, UK

[3]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy

[4]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

[5]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, 10-12 Russell Square, London, WC1B 5EH, UK

[6]Department of Psychology, University of Roehampton, London, SE15 4JD.

[7]Department of Psychosis Studies; Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, SE5 8AF.

[†]equal contribution

[¶]joint senior authors

*corresponding author:
rick.adams@ucl.ac.uk

Key words: schizophrenia; psychosis; Bayesian; disconfirmatory bias; beads task; attractor model

Abstract word count: 247 words
Introduction word count: 610 words

Discussion word count: 1500

3 Tables, 10 Figures

Abstract

Subjects with a diagnosis of schizophrenia (Scz) overweight unexpected evidence in probabilistic inference: such evidence becomes ‘aberrantly salient’. A neurobiological explanation for this effect is that diminished synaptic gain (e.g. hypofunction of cortical *N*-methyl-D-aspartate receptors) in Scz destabilizes quasi-stable neuronal network states (or ‘attractors’). This attractor instability account predicts that i) Scz would overweight unexpected evidence but underweight consistent evidence, ii) belief updating would be more vulnerable to stochastic fluctuations in neural activity, and iii) these effects would correlate.

Hierarchical Bayesian belief updating models were tested in two independent datasets (n=80 and n=167, male and female) comprising human subjects with schizophrenia, and both clinical and non-clinical controls (some tested when unwell and on recovery) performing the ‘probability estimates’ version of the beads task (a probabilistic inference task). Models with a standard learning rate, or including a parameter increasing updating to ‘disconfirmatory evidence’, or a parameter encoding belief instability were formally compared.

The ‘belief instability’ model (based on the principles of attractor dynamics) had most evidence in all groups in both datasets. Two of four parameters differed between Scz and non-clinical controls in each dataset: belief instability and

response stochasticity. These parameters correlated in both datasets. Furthermore, the clinical controls showed similar parameter distributions to Scz when unwell, but were no different to controls once recovered.

These findings are consistent with the hypothesis that attractor network instability contributes to belief updating abnormalities in Scz, and suggest that similar changes may exist during acute illness in other psychiatric conditions.

Significance Statement

Subjects with a diagnosis of schizophrenia (Scz) make large adjustments to their beliefs following unexpected evidence, but also smaller adjustments than controls following consistent evidence. This has previously been construed as a bias towards 'disconfirmatory' information, but a more mechanistic explanation may be that in Scz, neural firing patterns ('attractor states') are less stable and hence easily altered in response to both new evidence and stochastic neural firing. We model belief updating in Scz and controls in two independent datasets using a hierarchical Bayesian model, and show that all subjects are best fit by a model containing a belief instability parameter. Both this and a response stochasticity parameter are consistently altered in Scz, as the unstable attractor hypothesis predicts.

Introduction

Subjects with a diagnosis of schizophrenia (Scz) tend to use less evidence to make decisions in probabilistic tasks than healthy controls (Garety et al., 1991; Dudley et al., 2016). The paradigm most commonly used to demonstrate this effect is the 'beads' or 'urn' task, in which subjects are shown two urns, each containing opposite ratios of coloured beads (e.g. 85% blue and 15% red and vice versa), which are then hidden. A sequence of beads is then drawn (with replacement) from one urn, and the subject either has to stop the sequence when they are sure which urn it is coming from (the 'draws to decision' task) or the subject must rate the probability of the sequence coming from either urn after seeing each bead, without having to make any decision (the 'probability estimates' task). Bayesian analysis of these tasks has indicated that Scz are more stochastic in their responding (Moutoussis et al., 2011) and that they overweight recent evidence and thus update their beliefs (in the probabilistic sense) more rapidly (Jardri et al., 2017).

Several belief-updating abnormalities have been found in Scz using the 'probability estimates' task. The most consistent finding is that Scz (or just Scz with delusions (Moritz and Woodward, 2005)) change their beliefs *more* than non-psychiatric controls in response to changes in evidence (Langdon et al., 2010) – particularly 'disconfirmatory' evidence, i.e. evidence contradicting a current belief (Garety et al., 1991; Fear and Healy, 1997; Young and Bentall, 1997; Peters and Garety, 2006). Another is that probability ratings at the start of the sequence are higher in currently psychotic (but not in recovered) Scz than in both clinical and healthy controls (Peters and Garety, 2006), similar to the 'jumping to conclusions' bias in the 'draws to decision' version of the task. Others have also found that Scz update *less* than controls to more *consistent* evidence, in this (Horga, in preparation) and other paradigms (Averbeck et al., 2010).

These findings can potentially be understood in the light of the 'unstable attractor network' hypothesis of Scz. An attractor network is a neural network that can occupy numerous stable states that are learned from experience, via adjustments to synaptic weights. It can revisit these states if presented with inputs that resemble previous patterns of synaptic weights, or through spontaneous fluctuations in neural activity: either way, the activity of all nodes is 'attracted' to a quasi-stable state because the network energy is lower at these states, and network firing patterns evolve to minimise energy. Attractor networks were originally developed to model the storage and reactivation of memories (Hopfield, 1982), but related network models also offer mechanistic explanations for working memory storage (e.g. Brunel and Wang, 2001), decision-making (Wang, 2013) and interval timing (Standage et al., 2013), as well as Bayesian belief updating (Gepperth and Lefort, 2016).

In Scz, attractor states in prefrontal cortex are thought to be less stable, so it is easier for the network to switch between them, but harder to become more confident about (i.e. increase the stability of) any particular one (Rolls et al., 2008). This loss of stable neuronal states – recently demonstrated in two animal models of Scz (Hamm et al., 2017) – is thought to be due to hypofunction of *N*-methyl-D-aspartate receptors (NMDARs) or cortical dopamine 1 receptors in Scz

(Figure 1). Interestingly, healthy volunteers given ketamine (an NMDAR antagonist) show a decrement in updating to consistent stimulus associations and an increase in decision stochasticity in this context (Vinckier et al., 2016). Attractor network perturbations have been linked to working memory problems in Scz using a bistable (i.e. a stable 'up' state corresponding to persistent neuronal activity, and a 'down' state corresponding to background activity) model (Murray et al., 2014), but not as yet to a computational understanding of belief updating.

We analysed belief updating in Scz using the Hierarchical Gaussian Filter (HGF; Mathys et al., 2011), a variational Bayesian model with individual priors, in two independent 'probability estimates' beads task datasets. We asked: given the larger belief updates in Scz compared with controls, can these be explained by group differences in i) general learning rate and/or ii) response stochasticity, or by adding parameters encoding iii) the variance (i.e. uncertainty) of beliefs at the start of the sequence, iv) a propensity to overweight disconfirmatory evidence specifically, or v) patterns of belief updating typical of unstable attractor states in a Hopfield-type network, i.e. greater instability and stochasticity, which correlate with each other? (Note that the HGF does not contain attractor states: the model in (v) is designed to simulate the effects on inference that unstable neuronal attractors may have.) Furthermore, are these findings consistent within Scz tested at different illness phases, and are they unique to Scz or also present in other non-psychotic mood disorders?

Methods and Materials

Subject characteristics

Dataset 1 comprised 23 patients with delusions (18 Scz), 22 patients with non-psychotic mood disorders, and 35 non-clinical controls (overall, 50 male and 30 female – see Table 1 for details of the groups); the first two groups were selected from inpatient wards at the Maudsley and the Bethlem Royal Hospitals. All groups were tested twice (with loss of n=25 from the groups – see Table 1); the clinical groups were tested once when they were unwell ('baseline'), and again once they had recovered ('follow-up'). The mean time between testing sessions was 17.4 (range 6 to 41) weeks in the deluded group, 33.4 (range 4 to 68) weeks in the clinical control group, and 35.6 (range 27 to 46) weeks in the non-clinical control group. The deluded group's shorter inter-test interval was due to their shorter admission period and to the prioritization of their follow-up over the non-clinical control group. Dataset 1 is described in detail elsewhere (Peters and Garety, 2006).

Dataset 2 comprised 56 subjects with a diagnosis of schizophrenia (Scz) and 111 controls (overall, 83 male and 84 female – see Table 1). All subjects provided informed, written consent, and ethical permission for the study was obtained from the local NHS Research Ethics Committee (Reference 14/LO/0532). Given the National Adult Reading Test (Nelson, 1982) was used to estimate IQ in these participants, a recruitment condition was that English was their first language.

Measures of cognitive function and delusion-proneness (or schizotypy) were collected in all subjects; clinical symptom ratings were collected in clinical subjects only (see Table 1 for details).

Experimental design

Subjects in dataset 1 performed the 'probability estimates' beads task as used previously (Garety et al., 1991), with two urns with ratios of 85:15 and 15:85 blue and red beads respectively, and viewing a single sequence of ten beads (Figure 2); after each bead they had to mark an analogue scale (from 1 to 100) denoting the probability the urn was 85% red.

Subjects in dataset 2 performed the 'probability estimates' beads task, with two urns with ratios of 80:20 and 20:80 red and blue beads respectively. They each viewed four separate sequences (two identical pairs of sequences with the colours swapped within each pair) of ten beads (Figure 2); after each bead they had to mark a Likert scale (from 1 to 7) denoting the probability the urn was the 80% blue one. Two sequences contained an apparent change of jar. The order of the four sequences was randomised.

We used some of the behavioural measures employed in the original analysis of dataset 1 (Peters and Garety, 2006) to analyse dataset 2. These were 'disconfirmatory updating', the mean change in belief on seeing a bead of a different colour to the ≥2 beads preceding it and 'final certainty' (the response to the last bead). We altered their 'initial certainty' measure from the mean response to the first three beads to the response to the first bead, which comes closer to capturing the classic 'jumping to conclusions' bias (in which around 50% of Scz decide on the jar colour after seeing only one bead; (Garety et al., 1991), although the results of both measures are presented below.

Computational modelling

The optimal way to use sensory information to update one's beliefs under conditions of uncertainty is to use Bayesian inference. Neural systems are likely to approximate Bayesian inference using schemes of simple update equations (Rao and Ballard, 1999; Friston, 2005); one such model is the Hierarchical Gaussian Filter (HGF). The HGF is a hierarchical Bayesian inference scheme that gives a principled account of how beliefs are updated on acquiring new data, using variational Bayes and individual priors. Variational Bayesian schemes (e.g. (Beal, 2003) use analytic equations to derive an exact solution to an approximation of the posterior distribution over the latent variables and parameters (as opposed to sampling methods which approximate a solution to the exact posterior). The HGF has been used as a generic state model for learning under uncertainty and has repeatedly been shown to outperform similar approaches, such as reinforcement learning models with fixed (e.g. Rescorla-Wagner) or dynamic (e.g. (Sutton, 1992) learning rates (Iglesias et al., 2013; Diaconescu et al., 2014; Hauser et al., 2014; Vossel et al., 2014). One advantage of the HGF is that it contains subject-specific parameters (and prior beliefs) that can account for between-subject differences in learning whilst preserving the (Bayes) optimality of any individual's learning (relative to his/her model parameters and prior beliefs). These parameters may be encoded by tonic levels of neuromodulators such as dopamine (Marshall et al., 2016), or by the intrinsic properties of neuronal networks (e.g. the ratio of excitatory to inhibitory neural activity can affect both the speed of evidence accumulation (Lam et al., 2017) – analogous to the evolution rate in the HGF – and also response stochasticity (Murray et al., 2014)). Differences in model parameters between Scz and controls may therefore explain, in computational terms, how pathophysiology leads to abnormal inference (Adams et al., 2015).

In general, when modelling behaviour under Bayesian assumptions, it is necessary to distinguish between the model of the world used by the subject (the perceptual model) and a model of how a subject's beliefs translated into observed behaviour (the observation or response model). Most of the parameters pertain to the perceptual model (here, all parameters except response stochasticity ν – see Table 2) and reflect (inferred) neuronal processing. In contrast, the parameters of the response model link subjective states to behavioural outcomes, and thus may reflect stochasticity in neuronal processing, measurement noise (in some paradigms), or non-random effects that have not been captured by the perceptual model. This and related learning models are freely available from http://www.translationalneuromodeling.org/tapas/ (version 5.1.0): this analysis used the perceptual models 'hgf_binary' or 'hgf_ar1_binary' and the response model 'beta_obs'.

At the bottom of the model (Figure 3 shows some simulated responses) is the bead drawn $u^{(k)}$ on trial k and the probability $x_1^{(k)}$ that draws are coming from the blue jar. At the level above this is x_2, the tendency towards the blue jar (a transform of the probability, bounded by ±∞); by definition, $x_1 = s(x_2)$, where $s(\bullet)$ is the logistic sigmoid function. As x_2 approaches infinity, the probability of the blue jar approaches 1; as it approaches minus infinity, the probability of the blue jar approaches 0. For $x_2 = 0$, both jars are equally probable. This quantity is hidden from the subject and must be inferred: the subject's posterior estimate of x_2 is μ_2, and the subject's posterior estimate of the probability of the jar being blue on trial k is $s\left(\mu_2^{(k)}\right)$ – equivalent to the prediction (denoted by ^) on the next trial $\hat{\mu}_1^{(k+1)}$.

Before seeing any new input on trial k the model's expected jar probability $\hat{\mu}_1^{(k)}$ and precisions (inverse variances) $\hat{\pi}_1^{(k)}, \hat{\pi}_2^{(k)}$ of the expectations at each level are given by:

$$\hat{\mu}_1^{(k)} \equiv s\left(\kappa_1 \mu_2^{(k-1)}\right)$$

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

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

Note that in Models 1-4, κ_1 is fixed to 1. A new input $u^{(k)} \equiv \mu_1^{(k)}$ generates a prediction error $\delta_1^{(k)}$ and the model updates and generates a new prediction as follows:

$$\delta_1^{(k)} \equiv \mu_1^{(k)} - \hat{\mu}_1^{(k)}$$

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

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

$$\hat{\mu}_1^{(k+1)} \equiv s\left(\kappa_1 \mu_2^{(k)}\right)$$

The subject's response $y^{(k)}$ (i.e. where on the continuous or Likert scale they responded) is determined by $\hat{\mu}_1^{(k+1)}$ and the precision of the response model's beta distribution ν.

We parameterize the beta distribution in terms of its mean μ and precision ν. These sufficient statistics relate to the conventional parameterization in terms of the sufficient statistics α and β by the following bijection:

$$\mu := \frac{\alpha}{\alpha + \beta}$$

$$\nu := \alpha + \beta$$

Note that updates to μ_2 are driven by the product of the prediction error from Bayesian updating explained above and a learning rate which, crucially, can change over time: this is an important aspect of the HGF in contrast to learning models such as Rescorla-Wagner that have a fixed learning rate. Parameters which affect the degree to which μ_2 can change during the experiment include ω, φ, κ_1 and $\sigma_2^{(0)}$. The contributions of φ and κ_1 are illustrated in Figure 4 (left panels).

The model usually has a third level, at which x_3 encodes the phasic volatility of x_2 (this determines the probability of the jar changing at any point): given the very short sequences employed in our datasets, from which volatility cannot be reliably estimated, we omitted this level. In any case, volatility could not account for the rapid changes in learning rate (from trial to trial, following confirmatory vs disconfirmatory evidence) present in the Scz group in these datasets.

In Models 1 and 2, changes in x_2 from trial to trial occur only according to the evolution rate ω, the variance of the random process at the second level. These models were equivalent to the subsequent models with either φ (Models 3 and 4) fixed to 0 or κ_1 (Models 5 and 6) fixed to 1.

In Models 3 and 4, changes in x_2 from trial to trial occur according to an autoregressive (AR(1)) process that is controlled by three parameters: m, the level to which x_2 is attracted, φ, the rate of change of x_2 towards m, and ω, the variance of the random process:

$$p\left(x_2^{(k+1)}\right) \sim \mathcal{N}\left(x_2^{(k)} + \varphi(m - x_2^{(k)}), \exp(\omega)\right)$$

After inversion, the evolution of x_2 according to this equation is reflected in the prediction of μ_2:

$$\hat{\mu}_2^{(k+1)} = \mu_2^{(k)} + \varphi(m - \mu_2^{(k)})$$

In this study, given there was no bias towards one jar or the other, m was fixed to 0, so φ always acted to shift the model's beliefs back towards maximum uncertainty (i.e. disconfirm the current belief) about the jars. Figure 4 (upper left panel) illustrates the effect of φ on $s\left(\mu_2^{(k)}\right)$ over time.

In Models 5 and 6, changes in μ_2 from trial to trial occur according to two parameters: ω, the variance of the random process, and κ_1, a scaling factor that changes the size of updates when $\hat{\mu}_1$ = 0.5, or maximum uncertainty, relative to when $\hat{\mu}_1$ is closer to 0 or 1, i.e. when the subject is more confident about either jar. Figure 4 (lower left panel) illustrates the effect of κ_1 on $\hat{\mu}_1$ over time. Formally, the scaling occurs as:

$$\hat{\mu}_1^{(k+1)} \equiv s\left(\mu_2^{(k)}\kappa_1\right)$$

When $\kappa_1 > 1$, updating towards 1 on observing a blue bead ($u = 1$) is greatest (i.e. switching between jars becomes more likely) when $\hat{\mu}_1 < 0.3$; when $\kappa_1 < 1$, updating is comparatively far lower when $\hat{\mu}_1 < 0.3$. This is illustrated in Figure 4 (middle panel): for high values of κ_1 (brown line), belief updates that cross the $\hat{\mu}_1 = 0.5$ line encounter little resistance (i.e. little evidence is required to cause a large shift), while approaching the extremes of $\hat{\mu}_1 = 0$ and $\hat{\mu}_1 = 1$ in response to confirmatory evidence is resisted (belief shifts are very small for $\hat{\mu}_1$ near 1). By contrast, for low values of κ_1 (black line, Figure 4 middle panel), there is relatively less resistance against approaching the extremes while it takes more evidence for beliefs to cross the $\hat{\mu}_1 = 0.5$ line.

Figure 4 (right panel) illustrates the average absolute shifts in beliefs on observing beads of either colour. This 'vulnerability to updating' is highly reminiscent of the 'energy state' of a neural network model – i.e. in low energy states, less updating occurs. The effect of increasing κ_1 is to convert confident beliefs about the jar (near 0 and 1) from low to high 'energy states', i.e. to make them much more unstable. This recapitulates the attractor network properties illustrated in Figure 1: an unstable network easily switches from one state to another but has difficulty stabilising any one state, whereas a stable network requires more energy (here, information) to overcome the boundary between two states (here, beliefs). Models 5 and 6 therefore capture the effects of attractor (in)stability on belief updating, or at least the kind of updating for which (un)stable attractor states are a good analogy.

As group differences in initial updating had been observed in dataset 1, we also estimated the standard deviation of μ_2 before the sequence begins, $\sigma_2^{(0)}$, in Models 2, 4 and 6.

NB for intermediate values of κ_1, Models 5 and 6 produce similar belief updating trajectories to Models 3 and 4 (containing the disconfirmatory updating parameter φ): both make greater updates following disconfirmatory evidence. For more extreme values of κ_1, however, Models 5 and 6 produce trajectories that Models 3 and 4 cannot: φ cannot pull beliefs far towards certainty in the opposite jar (c.f. brown line in Figure 4, lower left panel), and neither can it make it *more* difficult to update to disconfirmatory evidence (c.f. black line in Figure 4, lower left panel).

The parameters ω and ν +/- $\sigma_2^{(0)}$ +/- φ or κ_1 were estimated individually for each subject. If estimated, the prior probability distributions for their values are given in Table 2. The means given here refer to the parameters' native space, but the variances refer not to the parameters' native space, which in many cases is bounded, but to the unbounded space they were transformed to for estimation purposes. Otherwise they were fixed as $\varphi = 0$ (Models 1 and 2) and $\sigma_2^{(0)} = 0.006$ (Models 1, 3 and 5). The model's prior beliefs about the jars at the start of the sequence were fixed at $\mu_2^{(0)} = 0$ (i.e. believing each to be equally likely). The priors were sufficiently uninformative to be easily updated by the data: all prior means are standard for the HGF except $\sigma_2^{(0)}$, which had to be increased from 0.006 to 0.8 to allow the data to change it. The latter change ensured that group differences in initial belief updating alone would cause group differences in $\sigma_2^{(0)}$ rather than κ_1.

Model fitting and statistical analysis

We tested models with different combinations of parameters ω, ν, φ or κ_1 and $\sigma_2^{(0)}$ (see Table 2). In analysing dataset 2, we concatenated all four sequences for each subject in order to estimate the model parameters as accurately as possible (resetting the beliefs about the jars at the start of each sequence).

After fitting the six models to each subject's data, we performed Bayesian model selection on all groups separately in both dataset 1 (at baseline and follow-up) and dataset 2. This procedure weights models according to their accuracy but penalises them for complexity (i.e. unnecessary extra parameters) to prevent overfitting (Stephan et al., 2009; Rigoux et al., 2014). The winning model in all eight groups was Model 6 (Figure 6), although around a third of psychotic subjects and non-clinical controls in dataset 1 (at baseline) and in dataset 2 were better fit by Model 4. It is unclear why this change occurs, but given that Model 6 can produce very similar trajectories to Model 4 for intermediate values of κ_1 (Figure 4), any increase in response stochasticity is likely to diminish the strength of evidence for one model over a similar one.

In order to confirm we could reliably estimate the parameters of the winning model, Model 6, we simulated 100 datasets using the modal values of the parameters for both control and Scz groups (Figure 5, upper and lower rows respectively; an example simulated dataset is shown in Figure 3). We then estimated the parameters for the simulated data, and showed that in most cases, the parameters are recovered reasonably accurately. The exception was $\sigma_2^{(0)}$ in the Scz group simulation, which was distributed around the prior mean of 0.8 rather than the true value of 1.5. We retained a prior mean of 0.8 for $\sigma_2^{(0)}$ because using a higher prior mean led to overestimation of $\sigma_2^{(0)}$ in other simulations (not shown).

Results

Behavioural results: dataset 1

Each group's mean responses are plotted in Figure 2A, and statistical tests detailed in Table 1 (*p(adj)* refers to the adjusted *p* value of Tukey's HSD *post hoc* test). As described previously (Peters and Garety, 2006), at baseline there was a significant difference in disconfirmatory updating between the groups ($F(2,77) = 6$, $p = 0.004$, ANOVA), and the psychotic group had greater disconfirmatory updating than the non-clinical controls (*p(adj)* = 0.003) but not the clinical controls (*p(adj)* = 0.4). There was no difference between the clinical and non-clinical controls (*p(adj)* = 0.13). There were also significant differences in initial certainty across the three groups ($F(2,77) = 8.7$, $p = 0.0004$, ANOVA); the psychotic group's initial certainty was higher than the non-clinical controls' (*p(adj)* = 0.0003) but not the clinical controls' (*p(adj)* = 0.25). There wasn't a significant difference between the clinical and non-clinical control groups (*p(adj)* = 0.06). There were no group differences in final certainty ($F(2,77) = 0.7$, p = 0.5, ANOVA).

At follow-up, the difference in disconfirmatory updating between the groups was no longer significant ($F(2,52) = 2.9$, $p = 0.06$, ANOVA); the psychotic group had greater disconfirmatory updating than the non-clinical controls (*p(adj)* = 0.049) but not the clinical controls (*p(adj)* = 0.4). There was no significant difference in initial certainty across the groups ($F(2,52) = 0.9$, $p = 0.4$, ANOVA). Differences in final certainty were no longer significant ($F(2,52) = 2.8$, $p = 0.07$, ANOVA); the biggest difference was the non-clinical controls' final certainty which was numerically higher than the clinical controls' (*p(adj)* = 0.057).

There were negative correlations between initial certainty and disconfirmatory updating at both baseline ($\rho = -0.41$, $p = 0.00015$) and follow-up ($\rho = -0.41$, $p = 0.002$), but not between final certainty and the other two measures ($p > 0.1$ in all four comparisons).

Behavioural results: dataset 2

The mean responses of subjects in each group are plotted in Figure 2B. There was a significant increase in disconfirmatory updating in Scz compared with controls ($t(88.6) = 2.1$, $p = 0.04$, Welch's *t*-test). There was mixed evidence for a difference in initial certainty between Scz and controls: Scz were more certain after the first bead in sequences A and B but not C or D (Figure 2 and Table 2), but the difference in mean initial certainty fell short of statistical significance ($t(110) = -1.9$, $p = 0.059$, Cohen's $d = 0.32$, Welch's *t*-test). Final certainty was only assessed in sequences A and D (B and C contained two changes of colour in the last three beads): in both sequences, Scz were less certain than controls (sequence A: $t(80.1) = 3.0$, $p = 0.004$, sequence D: $t(85.5) = 3.4$, $p = 0.001$, Welch's *t*-tests).

Initial certainty and disconfirmatory updating negatively correlated within both Scz (ρ = -0.46, p = 0.0003) and control (ρ = -0.57, $p = 10^{-11}$) groups. Final certainty did not correlate with either measure in either group (p > 0.4 in four comparisons).

Modelling results: dataset 1

Model selection results for the three groups analysed separately at both baseline and follow-up are plotted in Figure 6 (columns 1, 2, 4 and 5); the probability of each model being best for any given subject is shown in the left panel, and the probability of each model being the best overall is shown in the right panel. Model 6 is the clear winner at each time point, although a minority of psychotic and clinical controls are best fit by Model 4.

Model 6's parameter distributions are shown in Figure 7; they are skewed, hence non-parametric tests were used to determine group differences (full details in Table 3; *p(adj)* refers to the adjusted *p* value of Dunn's *post hoc* test). At baseline there were large group differences in belief instability κ_1 (χ^2(2,n=80) = 9.64, p = 0.008, η^2 = 0.12, Kruskal-Wallis' one-way ANOVA on ranks) and response stochasticity ν (χ^2(2,n=80) = 11.9, p = 0.003, η^2 = 0.15) but not in $\sigma_2^{(0)}$ or ω. There were statistically significant differences in κ_1 between the non-clinical controls and both the psychotic group (*p(adj)* = 0.01, Dunn's test) and the clinical control group (*p(adj)* = 0.01), but not between the latter two groups (*p(adj)* = 0.4). Similarly, there were statistically significant differences in ν between the non-clinical controls and both the psychotic group (*p(adj)* = 0.002, Dunn's test) and the clinical control group (*p(adj)* = 0.01), but not between the latter two groups (*p(adj)* = 0.3).

At follow-up, there were still large group differences in κ_1 ($\chi^2(2,n=55) = 8.0$, $p = 0.02$, $\eta^2 = 0.15$, Kruskal-Wallis' one-way ANOVA on ranks) and ν ($\chi^2(2,n=55) = 8.5$, $p = 0.01$, $\eta^2 = 0.16$) but not in $\sigma_2^{(0)}$ or ω. There was a significant difference in κ_1 between the psychotic and non-clinical control groups (*p(adj)* = 0.007, Dunn's test) but not the clinical and non-clinical control groups (*p(adj)* = 0.1); ν remained significantly different between the non-clinical controls and both the psychotic group (*p(adj)* = 0.01, Dunn's test) and now also between the psychotic and clinical control groups (*p(adj)* = 0.01), but not between the clinical and non-clinical controls (*p(adj)* = 0.5).

We explored whether group differences in κ_1 or ν at baseline and follow up might be ascribable to IQ (Quick Test score (Ammons and Ammons, 1962)), as the groups' IQ scores were not equivalent (Table 1). Including both IQ and group status within one regression model is an unsound method of testing for confounding by IQ because group and IQ are clearly not independent here (Miller and Chapman, 2001), so we tested for relationships between the parameters and IQ separately within each group at each time point. No relationships reached statistical significance (all $p > 0.1$), the closest being a trend between κ_1 and IQ in non-clinical controls only ($r = -0.30$, $p = 0.08$); nevertheless, given the smaller group sizes and larger between- versus within-group variances, it remains plausible that IQ differences contribute to group parameter differences.

We tested whether κ_1 or ν at baseline related to delusion-proneness (Peters Delusion Inventory score) across all groups, after first excluding any interaction between PDI and group; PDI significantly correlated with ν ($F(1,67)$ =

7.1, $p = 0.01$, ANCOVA) but not κ_1 ($F(1,67) = 3.2$, $p = 0.079$, ANCOVA). We tested whether κ_1 or ν at baseline was correlated with any particular subgroup of symptoms (measured using the Manchester Scale (Krawiecka et al., 1977)) in both clinical groups only, using the regression models κ_1 [or ν] ~ const + ν_1*MSaffective + ν_2*MSpositive + ν_3*MSnegative: none of the models were significant, however (all $p > 0.1$).

At baseline, there was no evidence of a correlation between κ_1 and antipsychotic medication dose ($p = 0.3$), but the correlation between ν and medication dose approached significance ($\rho = -0.4$, $p = 0.067$).

We tested for correlations between the Model 6 parameters (Spearman's ρ was used where distributions were not parametric): κ_1 and ν were negatively correlated both at baseline ($\rho = -0.38$, $p = 0.0004$) and at follow up ($\rho = -0.52$, $p = 0.0001$), as were κ_1 and ω at baseline ($\rho = -0.47$, $p = 10^{-5}$) and follow up ($\rho = -0.53$, $p = 10^{-5}$). In estimating the parameters from *simulated* data, the only correlation present in both simulations (indicating some consistent trading-off between these parameters during estimation) was between κ_1 and ω, with $r = -0.5$ in each case. This is not surprising, as both κ_1 and ω affect updating to new information throughout the sequence (unlike $\sigma_2^{(0)}$) in a deterministic way (unlike ν). Nevertheless, κ_1 was estimated very reliably in the first simulation (Figure 5, top row) and with reasonable accuracy in the second (Figure 5, bottom row), so we are confident that the group differences in κ_1 are genuine. The correlations of $\rho \approx -0.5$ between ω and κ_1 in dataset 1 are unlikely to be reliable, however.

Modelling results: dataset 2

We tested the same six models and performed Bayesian model selection as before. As in dataset 1, the winning model was Model 6 overall and in each group separately (Figure 6), although in the Scz group a minority were best captured by Model 4. Model 6's parameter distributions are shown in Figure 8; they are skewed, so non-parametric tests were used (full details in Table 3).

As in dataset 1, belief instability κ_1 was significantly higher in Scz than in controls (Z = -5.6, $p = 10^{-8}$, Mann-Whitney U test) with a medium-to-large effect size (r = 0.43); also response stochasticity ν was lower in Scz than in controls (Z = 3.9, p = 0.0001, r = 0.3, Mann-Whitney U test), as was initial belief variance $\sigma_2^{(0)}$ (Z = 3.1, p = 0.002, r = 0.24, Mann-Whitney U test). There were no statistically significant group differences in evolution rate ω. See Figures 6 and 7 for examples of model fits in subjects with lower κ_1 values (two controls in Figure 9) and higher κ_1 values (two Scz subjects in Figure 10); each figure also illustrates the effects of lower and higher ω values (in the top and bottom rows respectively). We repeated the analysis using a subset of the controls (n=60) that were better matched in age and sex, as the original control group was younger and more female than the patient group (Table 1). The group differences in κ_1 and ν were unchanged in this analysis (Z = -4.1, p = 0.00004; Z = 3.4, p = 0.0007 respectively, Mann-Whitney U tests), but that in $\sigma_2^{(0)}$ was no longer significant (Z = 1.9, p = 0.056, Mann-Whitney U test).

Although IQ (National Adult Reading Test score (Nelson, 1982)) was evenly matched in these groups, working memory (Letter Number Sequencing score (Wechsler, 1997)) was lower in Scz than in controls (see Table 1). We explored whether the group parameter differences might be related to working memory, by testing for correlations between κ_1 or ν and working memory in each group separately (Miller and Chapman, 2001): none were statistically significant (all $p > 0.1$). We also tested for relationships between κ_1 or ν and IQ (NART) in each group: ν and IQ (NART) were correlated in Scz ($r = 0.33$, $p = 0.014$), but no other relationships were significant (all $p > 0.1$).

We tested whether κ_1 or ν related to schizotypy (Schizotypal Personality Questionnaire score) across all groups but neither did so (both $p = 0.4$, ANCOVA). We tested whether κ_1 or ν were predicted by any particular subgroup of symptoms (measured using the Positive and Negative Symptom Scale (Kay et al., 1987)) in the Scz group only, using the regression model κ_1 [or ν] ~ const + ν_1*PANSSgeneral + ν_2*PANSSpositive + ν_3*PANSSnegative: the κ_1 model was not significant ($F = 0.9$, $p = 0.4$), but ν was weakly predicted by negative symptoms (overall $F = 2.76$, $p = 0.051$; for ν_3, $t = -2.1$, $p = 0.04$). We had no record of medication dose in dataset 2.

We tested for correlations between the Model 6 parameters: as in dataset 1, κ_1 and ν were negatively correlated (Figure 8; $\rho = -0.35$, $p = 10^{-6}$), but unlike dataset 1, the only other statistically significant correlation was between κ_1 and $\sigma_2^{(0)}$ ($\rho = -0.54$, $p = 10^{-13}$). There was a correlation of $r = -0.2$ between κ_1 and ν in the data simulated from modal Scz parameter values (Figure 5, bottom row), but no correlation in the first. This implies that the consistent correlations between these parameters of $\rho = -0.38$, $\rho = -0.52$ (dataset 1 baseline and follow-up) and $\rho = -0.35$ (dataset 2) are unlikely to be just estimation artefacts. The only other correlation between parameters in the simulated data was between $\sigma_2^{(0)}$ and κ_1, of r = -0.25, in the first simulation only. These parameters were correlated in dataset 2 but not dataset 1.

Discussion

Scz tend to update their beliefs more to unexpected information and less to consistent information, compared to controls. We have replicated these behavioural effects, and demonstrated a computational basis for them that is informed by the unstable attractor hypothesis of schizophrenia. In computational models of two 'beads task' datasets, Scz had consistently greater belief instability (κ_1) and response stochasticity (ν) than controls, as the unstable attractor hypothesis predicts. Furthermore, ν correlated with κ_1 in all three experiments, supporting the idea that ν is measuring a stochasticity that is related to κ_1 by an underlying neurobiological process, rather than simply an unmodelled effect.

These findings are important because they connect numerous reasoning biases previously found in Scz – e.g. a disconfirmatory bias (Garety et al., 1991; Fear and Healy, 1997; Young and Bentall, 1997; Peters and Garety, 2006), increased initial certainty (Peters and Garety, 2006), and decreased final certainty (Horga, in preparation) – and its associated stochasticity in responding (Moutoussis et al., 2011; Schlagenhauf et al., 2013) to model parameters that describe how belief updating in cortex could be perturbed by unstable attractor states due to NMDA (or dopamine 1) receptor hypofunction (Figure 1).

The unique features of Model 6 that make attractor dynamics a compelling neurobiological explanation for its dominance are both Scz and controls' non-linearities in belief updating to confirmatory versus disconfirmatory evidence. The Scz group updated its beliefs (sometimes much) more to disconfirmatory than confirmatory evidence – particularly at points of relative certainty about the jar – and the controls were the opposite. Models with uniformly high or low learning rates cannot reproduce these effects; and adding high- or low-level (sensory) uncertainty to a hierarchical model would lead to uniformly high or low learning rates respectively. Although Models 3 and 4 do show differential updating to confirmatory vs disconfirmatory evidence, this results in beliefs in either jar hovering around 0.5 (as in Figure 4, top left) rather than making large updates from belief in one jar to the other (as when κ_1 = exp(1.2): Figure 4, bottom left). Furthermore, degraded neuronal ensemble firing (consistent with unstable attractor states) has recently been shown to be common to two different mouse models of schizophrenia (Hamm et al., 2017).

In dataset 1, belief instability κ_1 and response stochasticity ν were also significantly different between the clinical (mood disorder) and non-clinical control groups when the former were unwell, but not at follow-up, whereas the differences between the psychotic group and non-clinical controls persisted. This indicates that the same computational parameters can be perturbed in either a trait- or state-like manner, perhaps by different mechanisms. It seems unlikely that these parameter changes simply reflect a lack of engagement with the task in clinical groups (especially when unwell), because the consistent changes in κ_1 – with which the changes in ν consistently correlate – reflect specific patterns of belief updating.

Parameter relationships with cognition and symptoms

Neither κ_1 nor ν showed significant relationships with IQ (in dataset 1) or working memory (in dataset 2) within the groups, giving some indication that the group differences in these cognitive measures were unlikely to be the main drivers of group differences in the parameters. Nevertheless, aside from the correlation between response stochasticity ν and IQ in dataset 2, it is perhaps surprising that there weren't more relationships between κ_1 or ν and cognitive measures in Scz, given it is likely that abnormal prefrontal dynamics have profound effects on all these variables. We may have lacked power to detect them – though dataset 2 had 80% power to detect a correlation of 0.33 – or perhaps different prefrontal regions contribute to working memory, IQ and belief updating.

One might also question why there were no strong relationships between κ_1 or ν and positive or negative symptom domains (negative symptoms were weakly associated with ν in dataset 2 only). Again, power may have been an issue, although note that across all subjects in dataset 1, response stochasticity ν was associated with PDI score even after including group in the model, indicating a potential relationship with delusions, but not with the broader concept of schizotypy (assessed in dataset 2). It is also likely that other pathological factors contribute to symptoms, beyond those measured here (e.g. striatal dopamine availability and positive symptoms). Of note, two other computational studies demonstrating clear working memory parameter differences between Scz and controls also failed to detect any relationship between those parameters and symptom domains (Collins et al., 2014, 2017). Both their and our Scz groups were taking antipsychotic medication, which is also likely to weaken correlations of parameters to positive symptoms.

Although replicated numerous times in the beads task, a 'disconfirmatory bias' is perhaps surprising in Scz, given one might expect delusional subjects to show a bias *against* disconfirmatory evidence (as indeed they do in tasks involving scenario interpretation (Woodward et al., 2006)). In fact, the disconfirmatory bias is misleadingly named, as Scz make large shifts in beliefs both away from *and back towards* the current hypothesis (there are numerous examples in both datasets in Figure 2). This pronounced switching behaviour in the beads task is likely to illustrate a more fundamental instability of cognition and prefrontal dynamics in Scz, rather than being related to delusions specifically; indeed, the latter may be an attempt to remedy the former.

It is interesting that non-clinical controls' data were also best fit by Model 6 in both datasets, implying that even healthy subjects show some asymmetry in their belief updating to expected versus unexpected evidence. Most non-clinical control subjects had κ_1<1, i.e. reduced updating to changing evidence.

Related modelling studies

How do these findings relate to other computational modelling work in Scz? A study of unmedicated, mainly first episode Scz performing a reversal learning task (Schlagenhauf et al., 2013) also demonstrated an increased tendency to switch that was not accounted for by reward sensitivity (which would be affected by more stochastic behaviour), and increased switching also occurs in chronic Scz (Waltz et al., 2013), although not always (Pantelis et al., 1999).

Two recent studies of similar tasks in Scz populations have also demonstrated evidence of non-linear belief updating. (Jardri et al., 2017) showed that the Scz group on average "overcount" the likelihood in a single belief update; an effect they attribute to reverberating cortical message-passing, but which could also be due to the belief instability shown by Model 6. (Stuke et al., 2017) showed in a very similar task that all subjects showed evidence of non-linear updating, but the Scz group updated more than controls to "irrelevant information" (i.e. disconfirmatory evidence). Some differences between their model and ours are that they did not estimate response stochasticity in their subjects (neither did (Jardri et al., 2017), and their 'non-linearity' parameter was bounded by linear updating on one side, roughly equivalent to belief instability κ_1 being constrained to being <1 in our model, whereas we have shown (as in (Jardri et al., 2017) that Scz belief updating is often beyond this bound (Figure 7), and more stochastic. Conversely, (Moutoussis et al., 2011) demonstrated increased response stochasticity in acutely psychotic subjects, but did not test for differences in belief updating.

The extent to which a loss of belief stability in Scz is apparent depends critically on the strength (precision) of incoming sensory evidence relative to the current belief (prior): if the former is less precise, no belief switching may occur, and instead the percept may be weighted towards the prior. In the beads task, sensory evidence (i.e., the colour of the bead drawn) is unambiguous, but a task using very imprecise auditory sensory evidence (Powers et al., 2017) demonstrated some interesting heterogeneity in Scz: non-hallucinating Scz showed greater belief updating relative to controls, while in hallucinating Scz, percepts were driven by prior expectations, leading to a reduction in the updating of their beliefs (relative to controls).

Further evidence for heterogeneity in Scz is that those with delusions have greater certainty about the hypothesis that matches the evidence at every stage (Speechley et al., 2010), unlike the reduced final certainty we observed in Scz in dataset 2. On the other hand, Scz with high negative symptoms have difficulty choosing the most rewarding option very consistently (Gold et al., 2012), which may reflect a lack of certainty about its value. We lacked sufficient power to detect differences between Scz with exclusively high positive or negative symptoms, however.

Limitations

Each of our datasets contains some limitations of the beads task that are addressed by the other. Dataset 1 did not include a memory aid or measure working memory, but dataset 2 did both, and dataset 2 also matched IQ across groups much better than dataset 1; dataset 2 used a Likert scale for responding and so could potentially exaggerate small changes in belief updating, but dataset 1 used a continuous measure; dataset 2 only tested stable outpatients, but dataset 1 tested more unwell inpatients and retested them once they were better. The main limitation common to both datasets is that all subjects with psychotic diagnoses were taking antipsychotic medication when tested. Although the correlation between v and medication dose was almost significant in dataset 1, this relationship seems likely to be driven by illness severity rather than medication itself. Dopamine 2 receptor antagonists seem to both reduce overconfidence in probabilistic reasoning (Andreou et al., 2014), and also reduce motor response variability (Galea et al., 2013) and so if anything likely reduce our group differences.

Conclusion

In conclusion, we have shown that Scz subjects in two independent beads task datasets have consistent differences in two parameters of a belief updating model that attempts to reproduce consequences of attractor network instability. Note that this study was designed to link patterns of inferences to model parameters that (do or don’t) mimic the effects of abnormal attractor states on belief updating. The HGF itself does not contain attractor states and no relation between its parameters and NMDAR function has hitherto been tested. More detailed spiking network modelling, pharmacological (or other NMDAR) manipulations and imaging are required in future to understand how neuromodulatory function in both pyramidal cells and inhibitory interneurons contributes to real attractor dynamics and probabilistic inference, and to seek empirical evidence for a correspondence between the stability of network states and the stability of its inferences (especially in schizophrenia). This work underscores the importance of relating psychological biases to their underlying computational mechanisms, and thence (in future) to the constraints – e.g. the hypofunction of NMDARs – that neurobiology imposes on these mechanisms.

Acknowledgements

The authors are very grateful to Dr Emmanuelle Peters for providing them with dataset 1. Dr Rick Adams is funded by the Academy of Medical Sciences (AMS-

SGCL13-Adams) and the National Institute of Health Research (CL-2013-18-003).

JG was supported in his contribution to this project by the British Academy.

Conflicts of Interest

No authors have any biomedical financial interests or potential conflicts of interest.

References

Abi-Saab WM, D'Souza DC, Moghaddam B, Krystal JH (1998) The NMDA antagonist model for schizophrenia: promise and pitfalls. Pharmacopsychiatry 31 Suppl 2:104–109.

Adams RA, Huys QJM, Roiser JP (2015) Computational Psychiatry: towards a mathematically informed understanding of mental illness. J Neurol Neurosurg Psychiatry.

Ammons RB, Ammons CH (1962) The Quick Test (QT): Provisional manual. Psychol Rep 11:111–161.

Andreou C, Moritz S, Veith K, Veckenstedt R, Naber D (2014) Dopaminergic modulation of probabilistic reasoning and overconfidence in errors: a double-blind study. Schizophr Bull 40:558–565.

Averbeck BB, Evans S, Chouhan V, Bristow E, Shergill SS (2010) Probabilistic learning and inference in schizophrenia. Schizophr Res Available at: http://www.ncbi.nlm.nih.gov/pubmed/20810252 [Accessed November 18, 2010].

Beal MJ (2003) Variational algorithms for approximate Bayesian inference. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.131.9951 [Accessed March 26, 2012].

Brunel N, Wang XJ (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11:63–85.

Collins AGE, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017) Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. Biol Psychiatry 82:431–439.

Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ (2014) Working memory contributions to reinforcement learning impairments in schizophrenia. J Neurosci Off J Soc Neurosci 34:13747–13756.

Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, Fehr E, Stephan KE (2014) Inferring on the Intentions of Others by Hierarchical Bayesian Learning. PLoS Comput Biol 10:e1003810.

Dudley R, Taylor P, Wickham S, Hutton P (2016) Psychosis, Delusions and the “Jumping to Conclusions” Reasoning Bias: A Systematic Review and Meta-analysis. Schizophr Bull 42:652–665.

Durstewitz D, Seamans JK (2008) The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. Biol Psychiatry 64:739–749.

Fear CF, Healy D (1997) Probabilistic reasoning in obsessive-compulsive and delusional disorders. Psychol Med 27:199–208.

Foulds GA, Bedford A (1975) Hierarchy of classes of personal illness. Psychol Med 5:181–192.

Friston KJ (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Galea JM, Ruge D, Buijink A, Bestmann S, Rothwell JC (2013) Punishment-induced behavioral and neurophysiological variability reveals dopamine-dependent selection of kinematic movement parameters. J Neurosci Off J Soc Neurosci 33:3981–3988.

Garety PA, Hemsley DR, Wessely S (1991) Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. J Nerv Ment Dis 179:194–201.

Gepperth A, Lefort M (2016) Learning to be attractive: Probabilistic computation with dynamic attractor networks. In: 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), pp 270–277.

Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, Collins AGE, Frank MJ (2012) Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. Arch Gen Psychiatry 69:129–138.

Hamm JP, Peterka DS, Gogos JA, Yuste R (2017) Altered Cortical Ensembles in Mouse Models of Schizophrenia. Neuron 94:153-167.e8.

Hauser TU, Iannaccone R, Ball J, Mathys C, Brandeis D, Walitza S, Brem S (2014) Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. JAMA Psychiatry 71:1165–1173.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 79:2554–2558.

Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HEM, Stephan KE (2013) Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron 80:519–530.

Jardri R, Duverne S, Litvinova AS, Denève S (2017) Experimental evidence for circular inference in schizophrenia. Nat Commun 8:14218.

Javitt DC, Zukin SR, Heresco-Levy U, Umbricht D (2012) Has an angel shown the way? Etiological and therapeutic implications of the PCP/NMDA model of schizophrenia. Schizophr Bull 38:958–966.

Kay SR, Fiszbein A, Opfer LA (1987) The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. Schizophr Bull 13:261–276.

Krawiecka M, Goldberg D, Vaughan M (1977) A standardized psychiatric assessment scale for rating chronic psychotic patients. Acta Psychiatr Scand 55:299–308.

Lam NH, Borduqui T, Hallak J, Roque AC, Anticevic A, Krystal JH, Wang X-J, Murray JD (2017) Effects of altered excitation-inhibition balance on decision making in a cortical circuit model. bioRxiv:100347.

Langdon R, Ward PB, Coltheart M (2010) Reasoning anomalies associated with delusions in schizophrenia. Schizophr Bull 36:321–330.

Marshall L, Mathys C, Ruge D, de Berker AO, Dayan P, Stephan KE, Bestmann S (2016) Pharmacological Fingerprints of Contextual Uncertainty. PLoS Biol 14:e1002575.

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39.

Miller GA, Chapman JP (2001) Misunderstanding analysis of covariance. J Abnorm Psychol 110:40–48.

Moritz S, Woodward TS (2005) Jumping to conclusions in delusional and non-delusional schizophrenic patients. Br J Clin Psychol 44:193–207.

Moutoussis M, Bentall RP, El-Deredy W, Dayan P (2011) Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. Cognit Neuropsychiatry 16:422–447.

Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, Wang X-J (2014) Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cereb Cortex N Y N 1991 24:859–872.

Nelson HE (1982) National Adult Reading Test (NART): For the Assessment of Premorbid Intelligence in Patients with Dementia : Test Manual. NFER-Nelson.

Pantelis C, Barber FZ, Barnes TR, Nelson HE, Owen AM, Robbins TW (1999) Comparison of set-shifting ability in patients with chronic schizophrenia and frontal lobe damage. Schizophr Res 37:251–270.

Peters E, Garety P (2006) Cognitive functioning in delusions: a longitudinal analysis. Behav Res Ther 44:481–514.

Peters ER, Joseph SA, Garety PA (1999) Measurement of delusional ideation in the normal population: introducing the PDI (Peters et al. Delusions Inventory). Schizophr Bull 25:553–576.

Powers AR, Mathys C, Corlett PR (2017) Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science 357:596–600.

Raine A (1991) The SPQ: a scale for the assessment of schizotypal personality based on DSM-III-R criteria. Schizophr Bull 17:555–564.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies - revisited. NeuroImage 84:971–985.

Rolls ET, Loh M, Deco G, Winterer G (2008) Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. Nat Rev Neurosci 9:696–709.

Schlagenhauf F, Huys QJM, Deserno L, Rapp MA, Beck A, Heinze H-J, Dolan R, Heinz A (2013) Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. NeuroImage.

Speechley WJ, Whitman JC, Woodward TS (2010) The contribution of hypersalience to the “jumping to conclusions” bias associated with delusions in schizophrenia. J Psychiatry Neurosci JPN 35:7–17.

Standage D, You H, Wang D-H, Dorris MC (2013) Trading speed and accuracy by coding time: a coupled-circuit cortical model. PLoS Comput Biol 9:e1003021.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. NeuroImage 46:1004–1017.

Stuke H, Stuke H, Weilnhammer VA, Schmack K (2017) Psychotic Experiences and Overhasty Inferences Are Related to Maladaptive Learning. PLoS Comput Biol 13:e1005328.

Sutton R (1992) Gain Adaptation Beats Least Squares? Available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.9218 [Accessed January 26, 2018].

Vinckier F, Gaillard R, Palminteri S, Rigoux L, Salvador A, Fornito A, Adapa R, Krebs MO, Pessiglione M, Fletcher PC (2016) Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. Mol Psychiatry 21:946–955.

Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE (2014) Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. Cereb Cortex N Y N 1991 24:1436–1450.

Waltz JA, Kasanova Z, Ross TJ, Salmeron BJ, McMahon RP, Gold JM, Stein EA (2013) The roles of reward, default, and executive control networks in set-shifting impairments in schizophrenia. PloS One 8:e57257.

Wang X-J (2013) The prefrontal cortex as a quintessential "cognitive-type" neural circuit: Working memory and decision making. Available at: https://nyuscholars.nyu.edu/en/publications/the-prefrontal-cortex-as-a-quintessential-cognitive-type-neural-c [Accessed January 31, 2018].

Wechsler D (1997) WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale. Psychological Corporation.

Woodward TS, Moritz S, Cuttler C, Whitman JC (2006) The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. J Clin Exp Neuropsychol 28:605–617.

Young HF, Bentall RP (1997) Probabilistic reasoning in deluded, depressed and normal subjects: effects of task difficulty and meaningful versus non-meaningful material. Psychol Med 27:455–465.

Figure Legends

Figure 1: Effects of attractor network dynamics on belief updating

This schematic illustrates the energy landscapes of two Hopfield-type networks each with two basins of attraction. The continuous black line depicts a normal network whose basins of attraction are relatively deep. The dotted black line depicts the effect of NMDAR (or cortical dopamine 1 receptor (Durstewitz and Seamans, 2008)) hypofunction (Abi-Saab et al., 1998; Javitt et al., 2012) on the energy landscape: the attractor basins become more shallow. We assume that Basins A and B correspond to different inferences about (hidden) states in the world, e.g. one jar or another being the source of beads in the beads task. The dots correspond to the networks' representations of either control or Scz subjects' beliefs about these hidden states. Such networks are highly reminiscent of Hopfield networks with two stored representations – in this case, the representations correspond to inferences about hidden states, rather than memories. The arrows depict the changes in network states resulting from sensory evidence for (solid arrows) or against (dashed arrows) the current inference. When the attractor basin is shallower, it is harder for supportive evidence to stabilise the current state much further, but it is easier for contradictory evidence – or just stochastic neuronal firing – to shift the current network state towards an alternative state. These changes in network dynamics may also be reflected in the inferences the network computes – i.e. easier switching between attractor basins may correspond to easier switching between beliefs – although this is yet to be demonstrated experimentally. NMDAR hypofunction could contribute to an increased tendency to switch between beliefs and increased stochasticity in responding in several ways (Rolls et al., 2008): i) by reducing inhibitory interneuron activity, via weakened NMDAR synapses from pyramidal cells to interneurons, such that other attractor states are less suppressed when one is active (a spiking network model has shown that this leads to more rapid initial belief updating in perceptual tasks (Lam et al., 2017)), ii) by reducing pyramidal cell activity, via weakened recurrent NMDAR synapses on pyramidal cells, such that attractor states are harder to sustain, and iii) by reducing the NMDAR time constant, making states more vulnerable to random fluctuations in neural activity. See also similar schematics elsewhere (Durstewitz and Seamans, 2008; Rolls et al., 2008).

Figure 2: Beads task schematic and group average confidence ratings in Datasets 1 and 2.

The bottom right panel is an illustrative schematic of the beads task: two jars containing opposite proportions of beads are concealed from view and a subject is asked to rate the probability of either jar being the source of a sequence of beads he/she is viewing (after each bead in turn). The top left panel shows the mean (± standard error) confidence ratings in the blue jar over the 10 bead sequence averaged across each group at baseline in dataset 1. The bottom left panel shows the same quantities at follow-up in dataset 1. The top right panel shows these quantities in four 10 bead sequences concatenated together (they were presented to the subjects separately during testing) in dataset 2.

Figure 3: The structure of the Hierarchical Gaussian Filter (Model 6) and some simulated data

In the upper left panel, the evolution of μ_2, the posterior estimate of tendency x_2 towards the blue (positive) or red (negative) jar, is plotted over two concatenated series of 10 trials (the first two in dataset 2). The estimate of the tendency on trial k+1, $\mu_2^{(k+1)}$, is selected from a Gaussian distribution with mean $\mu_2^{(k)}$ (blue line) and variance $\sigma_2^{(k)} + \exp(\omega)$ (blue shading). ω is a static source of variance at this level. The initial variance $\sigma_2^{(0)}$ (along with ω) affects the size of initial updates, so we estimated this parameter (which is often fixed). The beads seen by the subjects, $u^{(k)}$ (blue and red dots) and the response model are illustrated in the bottom left panel. The response model maps from $\hat{\mu}_1^{(k+1)}$ (purple line) – the prediction of x_1 on the next trial, which is a sigmoid function s of $\mu_2^{(k)}$ (or of $\left(\kappa_1\mu_2^{(k)}\right)$ in Models 5 and 6) – to $y^{(k)}$, the subject's indicated estimate of the probability the jar is blue (green dots). Variation in this mapping is modelled as the precision ν of a beta distribution.

The right panel is a schematic representation of the generative model in Models 5 and 6 (i.e. including κ_1). The black arrows denote the probabilistic network on trial k; the grey arrows denote the network at other points in time. The perceptual model lies above the dotted arrows, and the response model below them. The shaded circles are known quantities, and the parameters and states in unshaded circles are estimated. The dotted line represents the result of an inferential process (the response model builds on a perceptual model inference); the solid lines are generative processes.

Figure 4: Simulated data illustrating the effects of φ (Models 3 and 4) and κ_1 (Model 5 and 6) on inference

This figure illustrates the effects of φ (used in Models 3 and 4) and κ_1 (used in Models 5 and 6) on inference. Both panels show simulated perceptual model predictions in the same format as before, with $\sigma_2^{(0)}$ and ω set to their previous values – hence the purple line in these plots is identical to that in Figure 3. The second level and simulated responses y have been omitted for clarity.

Upper left panel: Simulations of a perceptual model incorporating an autoregressive order (1) process at the second level, using three different values of AR(1) parameter φ: 0, 0.2 and 0.8. The estimate of the tendency on trial k+1, $\mu_2^{(k+1)}$, is selected from a Gaussian distribution with mean $\mu_2^{(k)} + \varphi(m - \mu_2^{(k)})$ and variance $\sigma_2^{(k)} + \exp(\omega)$. Over time, μ_2 is therefore attracted towards level m (fixed to 0, i.e. at $\sigma(\mu_2) = 0.5$) at a rate determined by φ. In effect, this gives the model a 'disconfirmatory bias', such that as φ increases, $\sigma(\mu_2)$ is pulled further away from a belief in either jar, and towards 0.5 (maximum uncertainty about the jars).

Lower left panel: Simulations of a perceptual model using four different values of scaling factor κ_1, which alters the sigmoid transformation: $\hat{\mu}_1^{(k+1)} = s(\kappa_1 \cdot \mu_2^{(k)})$. When $\kappa_1 > \exp(0)$ updating is greater to unexpected evidence and lower to consistent evidence; when $\kappa_1 < \exp(0)$ the reverse is true. The red and brown lines (κ_1 > exp(0)) illustrate the effects of increasingly unstable attractor networks, i.e. switching between states (jars) becomes more likely (a concomitant increase in vulnerability to noise, i.e. response stochasticity, is not shown). The green line (κ_1 = exp(-1)) illustrates slower updating around $\hat{\mu}_1$ = 0.5, as was found in controls. κ_1 permits a greater range of updating patterns than φ (the green and brown trajectories in the lower panel cannot be produced by Model 4) which may be why Model 6 can fit both controls and Scz groups well.

Middle panel: This plot shows the effects of κ_1 on belief updating, as a function of the initial belief $\hat{\mu}_1$ ($\sigma_2^{(0)}$ and ω were set to 1.5 and -1 respectively, as in Figure 5; changing these parameters does not qualitatively alter the effects of κ_1 shown here). For values of κ_1 < exp(0)=1 (bottom three curves) and initial beliefs to the left of these curves' maxima (i.e. that the jar is probably red), relatively small increases in $\hat{\mu}_1$ are made if one blue bead (u = 1) is observed, such that the subject still believes the jar is most likely red. For values of κ_1 > exp(0.5) (top two curves), observing one blue bead causes such a large update for all but the most certain initial beliefs in a red jar that the subject's posterior belief is that the jar is probably blue. These subjects' beliefs are no longer stable, but neither can they reach certainty: only tiny updates towards 1 are possible for $\hat{\mu}_1 > 0.8$.

Right panel: This plot illustrates the average absolute shifts in beliefs on observing beads of either colour. This 'vulnerability to updating' is highly reminiscent of the 'energy state' of a neural network model (schematically illustrated in Figure 1) – i.e. in low energy states, less updating is expected. The effect of increasing κ_1 is to convert confident beliefs about the jar (near 0 and 1) from low to high 'energy states', i.e. to make them much more unstable.

Figure 5: Recovery of model parameters from simulated data

200 datasets were simulated using Model 6; 100 using modal parameter values for the control group (dataset 2) and 100 using modal values for the Scz group (also dataset 2) – the values are indicated using red lines. Both used settings of $\sigma_2^{(0)} = 1.5$, $\omega = -1$. The control group used $\kappa_1 = 0.37$ (i.e. exp(-1)) and $\nu = \exp(3)$. The Scz group used $\kappa_1 = 2.7$ (i.e. exp(1)) and $\nu = \exp(2)$. Histograms depicting the parameter estimates from model inversion using the same priors as were employed in the main analysis are shown above: the modal control and Scz simulation results are in the upper and lower rows respectively.

Figure 6: Bayesian model selection results for both datasets.

The left panel depicts the protected exceedance probabilities for the six models in each group in each dataset. The protected exceedance probability is the probability a particular model is more likely than any other tested model, above and beyond chance, given the group data (Rigoux et al., 2014). Model 6 wins in all groups in both datasets (upper row: controls, middle row: Scz, bottom row: clinical controls).

The right panel depicts the model likelihoods for the six models in each group in each dataset. The model likelihood is the probability of that model being the best for any randomly selected subject (Stephan et al., 2009). Model 4 is a clear runner-up in the psychotic (Scz) and clinical control groups at baseline in dataset 1, and in the Scz group in dataset 2.

Figure 7: Probability density plots for Model 6 parameters in dataset 1.

The distributions of parameter values for $\sigma_2^{(0)}$, ω, log(ν) and log(κ_1) are plotted for dataset 1 at baseline (upper row) and dataset 1 at follow-up (lower row). The symbols denote significant group differences: § between non-clinical controls and clinical controls, * between non-clinical controls and Scz, † between Scz and clinical controls. Please see the text for the details of all statistical comparisons.

Figure 8: Model 6 parameters in dataset 2 – distributions and correlation

Upper panel: The distributions of parameter values for $\sigma_2^{(0)}$, ω, log(ν) and log(κ_1) are plotted for dataset 2. The * symbol denotes significant group differences between the Scz group and non-clinical control subgroup (well-matched in age and sex); the group difference in $\sigma_2^{(0)}$ is not indicated because it was non-significant (p=0.056) in the well-matched comparison. Please see the text for the details of all statistical comparisons.

Lower panel: The significant correlation between log(ν) and log(κ_1) in dataset 2 is plotted, with controls' parameters in black and Scz in red. Similar correlations were also found in dataset 1 at both time points (see text).

Figure 9: Responses and model fits for two control subjects

These plots show two control subjects' responses to four ten-bead sequences concatenated together, in the same format as Figure 3 (but without the second level, due to space constraints); in the latter two sequences blue and red were swapped around for model-fitting purposes. Each plot shows $u^{(k)}$ – the beads seen by the subjects on trials k =1,…,10 (blue and red dots), y – the subject's (Likert scale) response about the probability the jar is blue (green dots), and $\hat{\mu}_1^{(k+1)}$ – the model's estimate of the subject's prediction the jar is blue (purple line). The parameter estimates for each subject are shown above their graphs. These subjects have fairly similar initial variance $\sigma_2^{(0)}$, (inverse) response stochasticity ν, and instability factor κ_1. Subject 18 in the upper panel has a much lower overall evolution rate ω than Subject 67 in the lower panel, therefore Subject 18 never reaches certainty about either jar, and makes relatively small changes to her beliefs in response to beads of varying colours. Both subjects have a low κ_1, and so they make relatively small adjustments to their beliefs following unexpected evidence (this behaviour can best be captured by the models containing κ_1 – see Figure 4). Subject 18's responses are very close to those predicted by the model, and this is reflected in her relatively high value of ν.

Figure 10: Responses and model fits for two Scz subjects

These plots show two Scz subjects' responses to four ten-bead sequences in the same format as Figure 9. These subjects have similar evolution rate ω to the control subjects in Figure 9, but they both have a much higher κ_1, meaning that they make much greater changes to their beliefs when presented with unexpected evidence, but do not reach certainty when faced with consistent evidence. Subject 122 (lower panel) has a slightly higher evolution rate ω than Subject 145 (upper panel), and so his switching between jars is even more pronounced. These subjects also have slightly lower (inverse) response stochasticity ν than the control subjects in Figure 9, and so their responses tend to be further from the model predictions.

Dataset 1							Dataset 2			
	Non-clinical controls t1	**Non-clinical controls t2**	**Clinical controls t1**	**Clinical controls t2**	**Psychotic t1**	**Psychotic t2**		**Controls (all)**	**Scz**	**Controls (subset)**
N	35	20	22	18	23	17	N	111	56	60
Age[a]	27.77 (6.74)	27.9 (6.37)	40.91 (13.57)	40.1 (13)	31.22 (7.28)	29.9 (7.83)	Age	32.8 (11.5)	45.3 (8.8)	39.5 (11.4)
Gender	18 M, 17 F	12 M, 8 F	11 M, 11 F	8 M, 10 F	21 M, 2 F	17 M, 0 F	Gender	45 M, 66 F	38 M, 18 F	40 M, 20 F
Cognitive measures										
IQ[b]	107.5 (11.6)	108.6 (10.3)	97.4 (13.8)	99.8 (10.2)	88.1 (12.7)	87.8 (14.2)	NART[a]	112 (6.9)	109 (8.2)	112 (7.5)
							Working memory (LNS)[b]	16.2 (2.8)	10.3 (4.2)	16.4 (2.7)
Delusion proneness							**Schizotypy**			
PDI (total)[c]	54.6 (43.1)	43.6 (42.5)	87.1 (55.2)	64.3 (57.3)	138.1 (74.2)	96.7 (42.6)	SPQ, cognitive	2.8 (1.9)	4.0 (2.6)	3.1(2)
DSSI[d]	2.3 (4.9)	2.9 (5.3)	4.8 (4.5)	4.5 (5.6)	15.2 (6.3)	8.1 (6.6)	SPQ, interpers	3.2 (2.2)	5.3 (2.6)	3.2 (2.2)

							SPQ, disorg	2.1 (1.7)	2.7 (1.9)	1.9 (1.8)
							SPQ, total[c]	8.2 (1.3)	12 (5.3)	8.2 (4.4)
Diagnosis/ Symptoms										
Diagnoses	-	-	16 Depression, 3 anxiety & depression, 3 SAD	12 Depression, 3 anxiety & depression, 3 SAD	18 Scz, 5 bipolar/ schizo-affective	13 Scz, 4 bipolar/ schizo-affective	Diagnoses	-	56 Scz	-
MS affective	-	-	4.6 (1.7)	1.0 (1.2)	1.8 (1.5)	1.5 (1.3)	PANSS, gen	-	32.6 (9.2)	-
MS positive	-	-	0.3 (0.8)	0 (0)	6.0 (2.4)	1.4 (1.7)	PANSS, pos	-	15.9 (5.8)	-
MS negative	-	-	0.7 (1.6)	1.8 (3.19)	1.3 (2.0)	0.9 (1.6)	PANSS, neg	-	15.9 (6.2)	-
MS total[e]	-	-	5.5 (2.6)	2.8 (3.39)	9.1 (3.76)	3.7 (3.9)	PANSS, total	-	64.4 (17.3)	-
Beads task										
Initial certainty (1 bead)[f]	0.58 (0.15)	0.59 (0.12)	0.68 (0.19)	0.63 (0.16)	0.76 (0.17)	0.68 (0.29)	Initial certainty (all, 1 bead)[d]	0.67 (0.13)	0.71 (0.14)	0.68 (0.14)

Initial certainty (3 beads)[g]	0.65 (0.14)	0.67 (0.1)	0.69 (0.15)	0.64 (0.16)	0.78 (0.15)	0.74 (0.15)	Initial certainty (all, 2-3 beads)[e]	0.7 (0.12)	0.71 (0.12)	0.71 (0.13)
Disconfirmatory updating[h]	-0.06 (0.14)	-0.03 (0.13)	-0.19 (0.3)	-0.11 (0.22)	-0.29 (0.33)	-0.2 (0.3)	Disconfirmatory updating (all sequences)[f]	-0.16 (0.17)	-0.23 (0.22)	-0.19 (0.2)
Final certainty[i]	0.85 (0.2)	0.94 (0.11)	0.82 (0.16)	0.79 (0.23)	0.88 (0.11)	0.85 (0.23)	Final certainty Sequence A[g]	0.88 (0.16)	0.77 (0.25)	0.86 (0.18)
							Final certainty Sequence D[h]	0.12 (0.18)	0.25 (0.24)	0.16 (0.2)

Table 1: Demographic, psychological and behavioural details of both datasets

Dataset 1 includes measures at both baseline (t1) and follow-up (t2). In dataset 1, verbal IQ was estimated using the Quick Test (Ammons and Ammons, 1962) and delusion proneness using the Peters Delusion Inventory, PDI (Peters et al., 1999) and Delusions-Symptoms-States Inventory, DSSI (Foulds and Bedford, 1975). Symptoms were assessed using the Manchester Scale, MS (Krawiecka et al., 1977). In the tests below, 'Scz' refers to the whole Psychotic group. Results are given for 'Initial certainty' using both the measure in the original analysis of dataset 1 (Peters and Garety, 2006), the mean response to the first three beads ('3 beads') – in dataset 2 this had to be the mean response to the first three beads in sequences B and C and two beads in sequences A and D ('2-3 beads') – and using the response to the first bead ('1 bead').

[a] At t1: One-way ANOVA $F(2,77) = 13.9$, $p = 10^{-5}$. Tukey's HSD: Scz vs Non-clinical controls diff = 3.45, $p(adj) = 0.35$; Clinical vs Non-clinical controls diff = 13.1, $p(adj) = 10^{-5}$; Clinical controls vs Scz diff = 9.69, $p(adj) = 0.002$

At t2: One-way ANOVA $F(2,52) = 8.85$, $p = 0.0005$. Tukey's HSD: Scz vs Non-clinical controls diff = 1.98, $p(adj) = 0.8$; Clinical vs Non-clinical controls diff = 12.2, $p(adj) = 0.0006$; Clinical controls vs Scz diff = 10.2, $p(adj) = 0.007$

[b] At t1: One-way ANOVA $F(2,75) = 16.2$, $p = 10^{-6}$; Tukey's HSD: Scz vs Non-clinical controls diff = -19.5, $p(adj) = 10^{-6}$; Clinical vs Non-clinical controls diff = -10.1, $p(adj) = 0.011$; Clinical controls vs Scz diff = 9.36, $p(adj) = 0.043$

At t2: One-way ANOVA $F(2,51) = 14.5$, $p = 10^{-5}$; Tukey's HSD: Scz vs Non-clinical controls diff = -20.8, $p(adj) = 10^{-5}$; Clinical vs Non-clinical controls diff = -8.8, $p(adj) = 0.057$; Clinical controls vs Scz diff = 12, $p(adj) = 0.01$

[c] At t1: One-way ANOVA $F(2,68) = 12.6$, $p = 0.00002$; Tukey's HSD: Scz vs Non-clinical controls diff = 83.5, $p(adj) = 10^{-5}$; Clinical vs Non-clinical controls diff = -32.5, $p(adj) = 0.094$; Clinical controls vs Scz diff = -51, $p(adj) = 0.016$

At t2: One-way ANOVA $F(2,52) = 4$, $p = 0.024$; Tukey's HSD: Scz vs Non-clinical controls diff = 53.1, $p(adj) = 0.018$; Clinical vs Non-clinical controls diff = -20.7, $p(adj) = 0.5$; Clinical controls vs Scz diff = -32.4, $p(adj) = 0.22$

[d] At t1: One-way ANOVA $F(2,76) = 43$, $p = 10^{-13}$; Tukey's HSD: Scz vs Non-clinical controls diff = 12.9, $p(adj) = 10^{-10}$; Clinical vs Non-clinical controls diff = 2.52, $p(adj) = 0.19$; Clinical controls vs Scz diff = -10.4, $p(adj) = 10^{-8}$

At t2: One-way ANOVA $F(2,51) = 3.7$, $p = 0.032$; Tukey's HSD: Scz vs Non-clinical controls diff = 5.2, $p(adj) = 0.026$; Clinical vs Non-clinical controls diff = 1.65, $p(adj) = 0.66$; Clinical controls vs Scz diff = -3.56, $p(adj) = 0.18$

[e] At t1: Welch's $t(38.4) = -3.62$, $p = 0.00086$, Cohen's $d = 1.1$

At t2: Welch's $t(17.8) = -2.55$, $p = 0.02$, Cohen's $d = 1.0$

[f] At t1: One-way ANOVA $F(2,77) = 8.7$, $p = 0.0004$; Tukey's HSD: Scz vs Non-clinical controls diff = 0.18, $p(adj) = 0.0003$; Clinical vs Non-clinical controls diff = 0.11, $p = 0.06$; Clinical controls vs Scz diff = -0.08, $p(adj) = 0.25$

At t2: One-way ANOVA $F(2,52) = 0.9$, $p = 0.4$

[g] At t1: One-way ANOVA $F(2,77) = 6.2$, $p = 0.003$; Tukey's HSD: Scz vs Non-clinical controls diff = -0.14, $p(adj) = 0.002$; Clinical vs Non-clinical controls diff = 0.04, $p = 0.57$; Clinical controls vs Scz diff = -0.096, $p(adj) = 0.074$

At t2: One-way ANOVA $F(2,52) = 2.35$, $p = 0.11$; Tukey's HSD: Scz vs Non-clinical controls diff = 0.07, $p(adj) = 0.28$; Clinical vs Non-clinical controls diff = -0.03, $p = 0.8$; Clinical controls vs Scz diff = -0.1, $p(adj) = 0.1$

[h] At t1: One-way ANOVA $F(2,77) = 6$, $p = 0.004$; Tukey's HSD: Scz vs Non-clinical controls diff = -0.23, $p(adj) = 0.003$; Clinical vs Non-clinical controls diff = -0.14, $p = 0.13$; Clinical controls vs Scz diff = 0.097, $p(adj) = 0.41$

At t2: One-way ANOVA $F(2,52) = 2.9$, $p = 0.062$; Tukey's HSD: Scz vs Non-clinical controls diff = -0.18, $p(adj) = 0.049$; Clinical vs Non-clinical controls diff = -0.08, $p = 0.51$; Clinical controls vs Scz diff = 0.098, $p(adj) = 0.4$

[i] At t1: One-way ANOVA $F(2,77) = 0.71$, $p = 0.5$

At t2: One-way ANOVA $F(2,52) = 2.79$, $p = 0.07$; Tukey's HSD: Scz vs Non-clinical controls diff = -0.082, $p(adj) = 0.41$; Clinical vs Non-clinical controls diff = -0.15, $p = 0.057$; Clinical controls vs Scz diff = -0.066, $p(adj) = 0.57$

As reported previously, there were consistent negative correlations between initial certainty (2-3 beads) and disconfirmatory updating in the clinical controls (baseline: $\rho = -0.68$, $p = 0.0005$; follow-up: $\rho = -0.75$, $p = 0.0003$) and the non-clinical controls (baseline: $\rho = -0.52$, $p = 0.001$; follow-up: $\rho = -0.43$, $p = 0.06$), but not in the psychotic group (baseline: $\rho = -0.30$, p = 0.17; follow-up: $\rho = 0.17$, $p = 0.5$). There was no consistent correlation between final certainty and either of the other two measures at either time point ($p \geq 0.1$ in 11 out of 12 comparisons).

In dataset 2, IQ was estimated using the National Adult Reading Test, NART (Nelson, 1982) and working memory using the Letter Number Sequencing task, LNS, from the Wechsler Adult Intelligence Scale-III (Wechsler, 1997). Schizotypy was assessed using the Schizotypal Personality Questionnaire, SPQ (Raine, 1991), and symptoms using the Positive and Negative Syndrome Scale, PANSS (Kay et al., 1987).

As can be seen in Figure 2 (main text), the Scz group showed greater initial certainty (1 bead) in sequences A and B (Welch's $t(94) = 2.8$, $p = 0.007$, Cohen's $d = 0.47$; Welch's $t(97) = 3$, $p = 0.004$, Cohen's $d = 0.5$, respectively) but not C and D (Welch's $t(87) = 0.5$, $p = 0.6$, Cohen's $d = 0.09$; Welch's $t(90) = -0.34$, $p = 0.73$, Cohen's $d = 0.06$, respectively).

[a] Controls (all): Welch's $t(95.1) = 2.27$, $p = 0.026$, Cohen's $d = 0.38$; Controls (subset): Welch's $t(111) = 1.95$, $p = 0.053$, Cohen's $d = 0.36$

[b] Controls (all): Welch's $t(81) = 9.57$, $p = 10^{-14}$, Cohen's $d = 1.66$; Controls (subset): Welch's $t(93.6) = 9.25$, $p = 10^{-15}$, Cohen's $d = 1.73$

[c] Controls (all): Welch's $t(92.4) = -4.64$, $p = 10^{-5}$, Cohen's $d = 0.78$; Controls (subset): Welch's $t(107) = -4.19$, $p = 10^{-5}$, Cohen's $d = 0.78$

[d] Controls (all): Welch's $t(110) = -1.9$, $p = 0.059$, Cohen's $d = 0.32$; Controls (subset): Welch's $t(110) = -1.1$, $p = 0.28$, Cohen's $d = 0.2$

[e] Controls (all): Welch's $t(109.1) = -0.76$, $p = 0.45$, Cohen's $d = 0.12$; Controls (subset): Welch's $t(113.9) = -0.19$, $p = 0.85$, Cohen's $d = 0.03$

[f] Controls (all): Welch's $t(88.2) = 2.09$, $p = 0.04$, Cohen's $d = 0.36$; Controls (subset): Welch's $t(110.4) = -0.94$, $p = 0.35$, Cohen's $d = 0.18$

[g] Controls (all): Welch's $t(80.1) = 2.99$, $p = 0.0038$, Cohen's $d = 0.56$; Controls (subset): Welch's $t(98.7) = 2.18$, $p = 0.032$, Cohen's $d = 0.41$

[h] Controls (all): Welch's $t(85.5) = -3.41$, $p = 0.001$, Cohen's $d = 0.62$; Controls (subset): Welch's $t(106) = -2.21$, $p = 0.029$, Cohen's $d = 0.42$

	Perceptual model parameters (prior mean in native space, prior variance in estimation space)				Response model parameter
Model	Evolution rate	Initial variance of belief re jars	Disconfirmatory bias	Belief instability	Response stochasticity
1	ω (-2, 16)				ν (exp(4.85), 1)
2	ω (-2, 16)	$\sigma_2^{(0)}$ (0.8, 0.5)			ν (exp(4.85), 1)
3	ω (-2, 16)		φ (0.1, 2)		ν (exp(4.85), 1)
4	ω (-2, 16)	$\sigma_2^{(0)}$ (0.8, 0.5)	φ (0.1, 2)		ν (exp(4.85), 1)
5	ω (-2, 16)			κ_1 (1,1)	ν (exp(4.85), 1)
6	ω (-2, 16)	$\sigma_2^{(0)}$ (0.8, 0.5)		κ_1 (1,1)	ν (exp(4.85), 1)

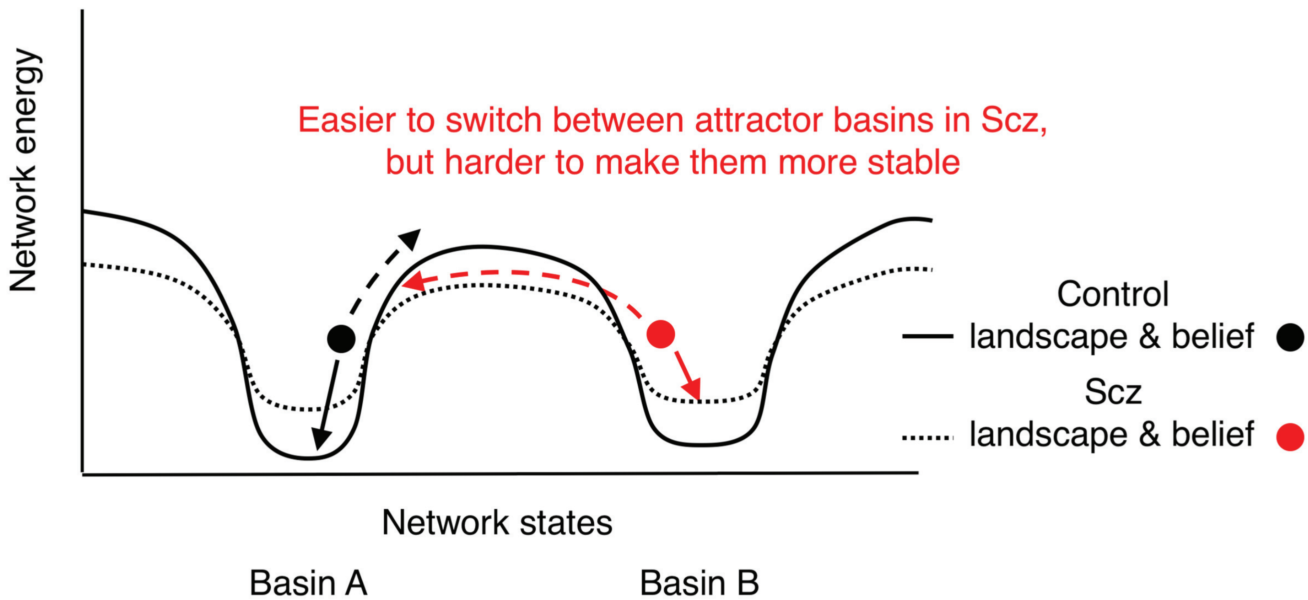
Table 2: Models, parameters and their prior distributions.

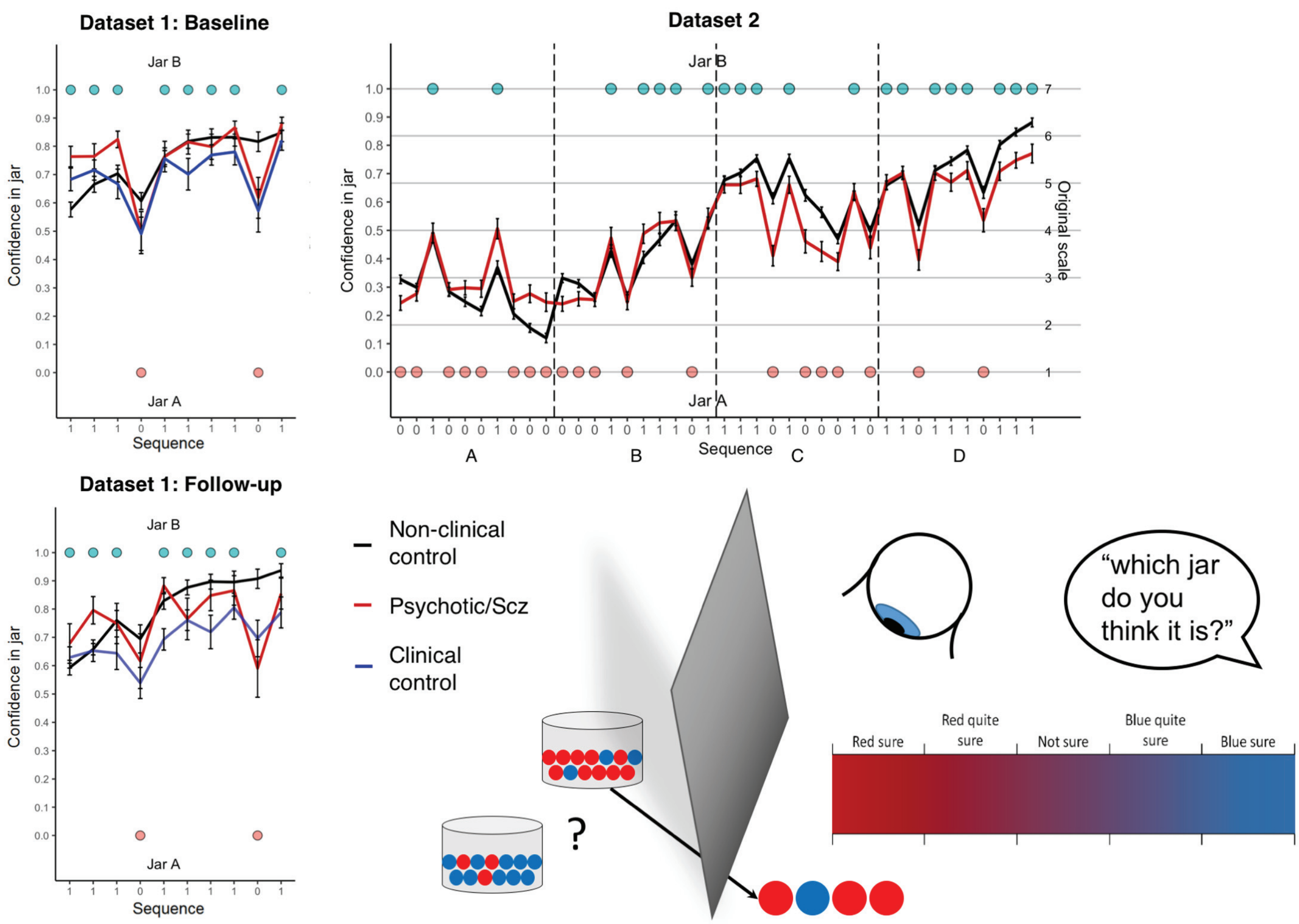
	$\sigma_2^{(0)}$	ω	$\log(\nu)$	$\log(\kappa_1)$
Dataset 1 (baseline, n=80)				
Non-clinical controls: mean(std)	2.5(3.9)	-1.3(2.4)	4.1(1.0)	-0.8(1.4)
Psychotic: mean(std)	3.0(3.9)	-1.4(2.0)	3.1(1.1)	-0.2(0.8)
Clinical controls: mean(std)	1.4(1.9)	-1.2(2.0)	3.3(1.3)	-0.1(1.4)
Kruskal-Wallis Chi Sq (2,80)	2.33, p=0.31 η^2=0.02	0.22, p=0.9 η^2=0.0	11.9, p=0.003 η^2=0.15	9.6, p=0.008 η^2=0.12
Post hoc Dunn tests				

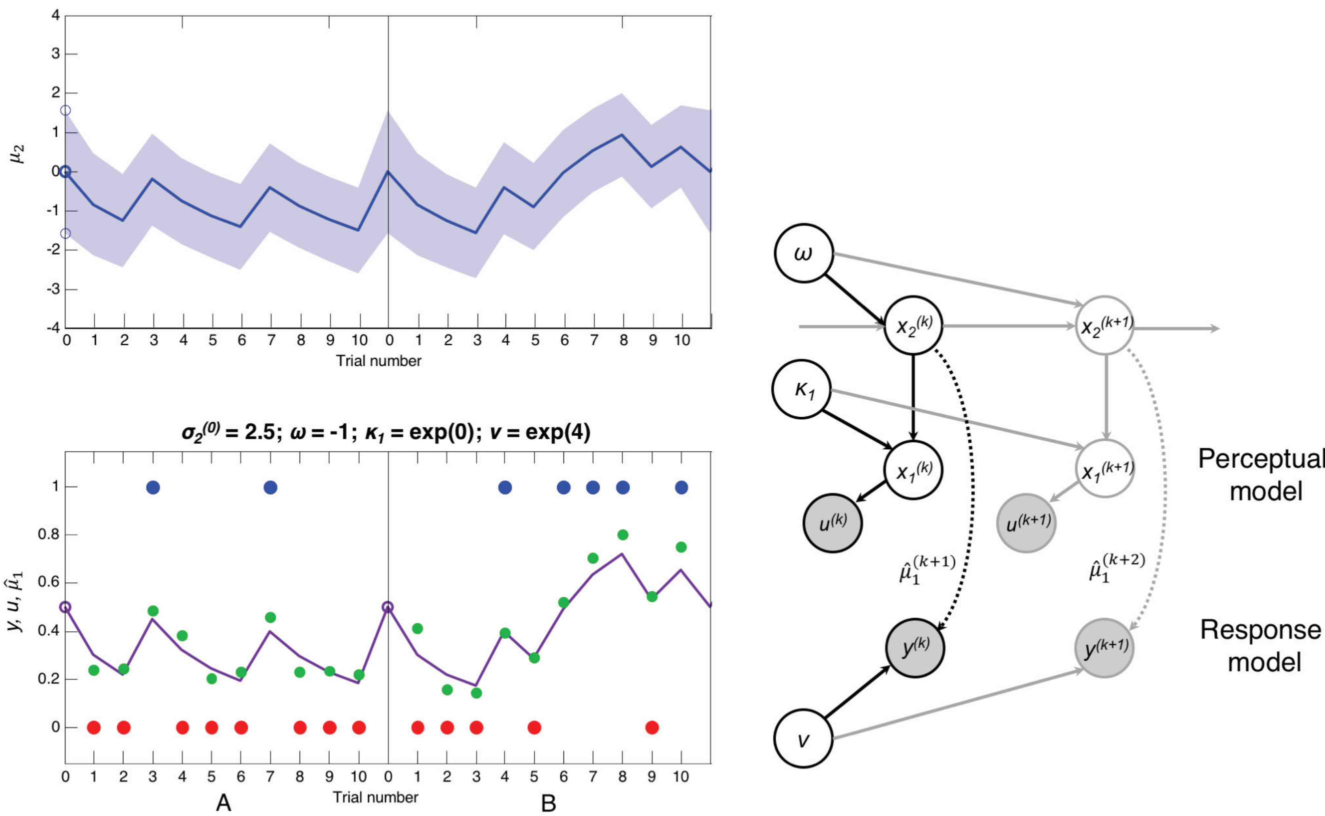
Psychotic vs non-clinical controls	*p(adj)*=0.3	*p(adj)*=1	*p(adj)*=0.002	*p(adj)*=0.01
Clinical vs non-clinical controls	*p(adj)*=0.2	*p(adj)*=0.7	*p(adj)*=0.01	*p(adj)*=0.01
Psychotic vs clinical controls	*p(adj)*=0.2	*p(adj)*=0.5	*p(adj)*=0.3	*p(adj)*=0.4
Dataset 1 (follow-up, n=55)				
Non-clinical controls: mean(std)	2.8(3.4)	-0.9(2.0)	3.6(0.8)	-1.2(1.1)
Psychotic: mean(std)	3.2(3.7)	-1.4(1.5)	2.5(1.2)	-0.3(0.8)
Clinical controls: mean(std)	1.2(0.9)	-1.1(2.0)	3.5(1.1)	-0.5(1.4)
Kruskal-Wallis Chi Sq (2,80)	2.35, p=0.3 η^2=0.04	2.32, p=0.3 η^2=0.04	8.5, p=0.01 η^2=0.16	8.0, p=0.02 η^2=0.15
Post hoc Dunn tests				
Psychotic vs non-clinical controls	*p(adj)*=0.4	*p(adj)*=0.2	*p(adj)*=0.01	*p(adj)*=0.007
Clinical vs non-clinical controls	*p(adj)*=0.2	*p(adj)*=0.3	*p(adj)*=0.5	*p(adj)*=0.1
Psychotic vs clinical controls	*p(adj)*=0.3	*p(adj)*=0.3	*p(adj)*=0.01	*p(adj)*=0.1
Dataset 2 (n=167)				
Non-clinical controls: mean(std)	3.1(2.6)	-2.3(2.0)	2.8(1.0)	-0.8(0.9)
Scz: mean(std)	1.9(1.5)	-2.1(1.8)	2.1(1.2)	0.2(1.0)
Mann-Whitney U test	Z=3.1, p=0.002, r=0.24	Z=-0.6, p=0.6, r=0.04	Z=3.9, p=0.0001, r=0.3	Z=-5.6, p=3x10^{-8}, r=0.43
Dataset 2				

(better-matched controls, n=116)				
Non-clinical controls: mean(std)	2.8(2.7)	-2.2(2.1)	2.9(1.1)	-0.6(1.0)
Scz: mean(std)	1.9(1.5)	-2.1(1.8)	2.1(1.2)	0.2(1.0)
Mann-Whitney U test	Z=1.9, p=0.056, r=0.18	Z=0.12, p=0.9, r=0.01	Z=3.4, p=0.0007, r=0.31	Z=-4.1, p=0.00004, r=0.38

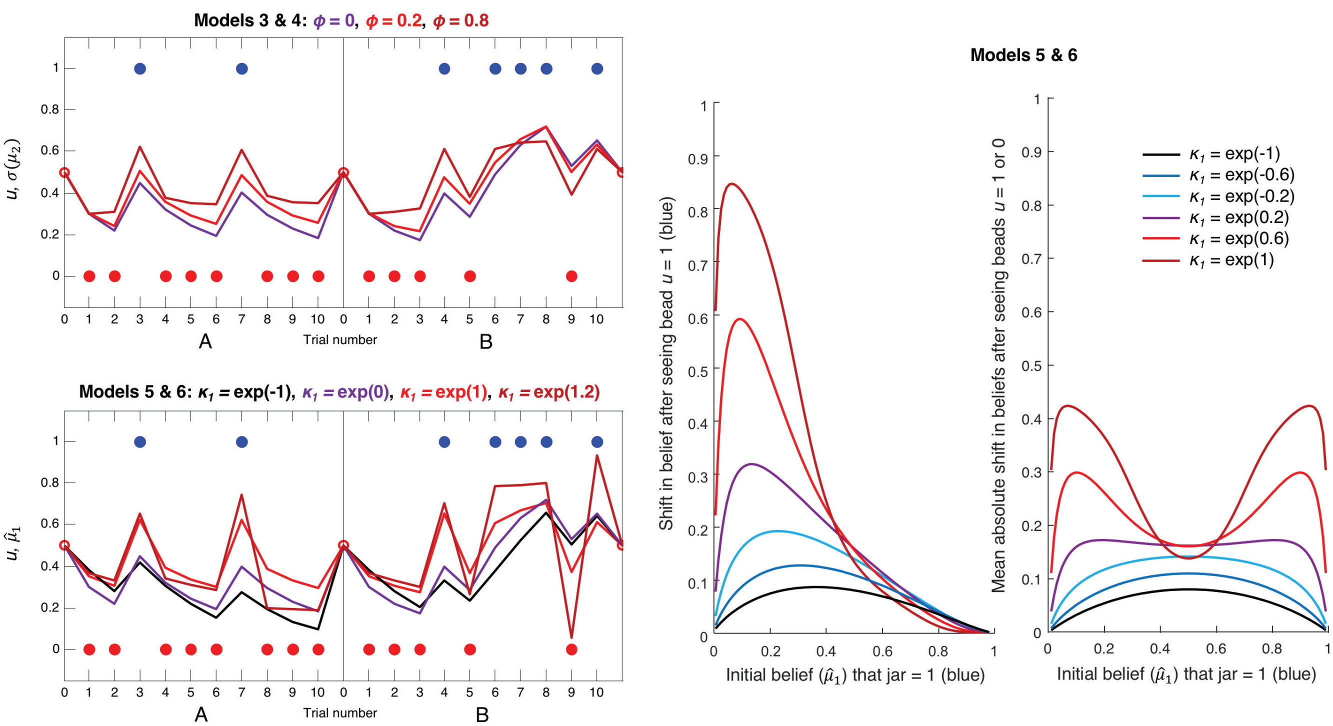
Table 3: Parameter distributions and statistical tests in Datasets 1 and 2

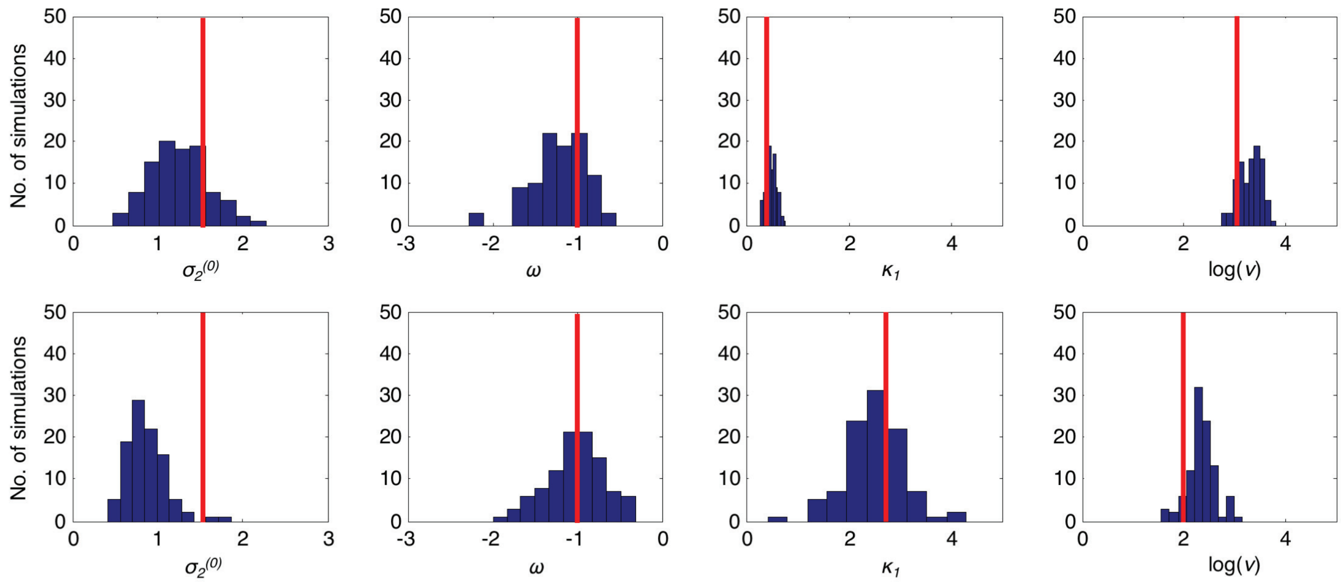

Network energy
Easier to switch between attractor basins in Scz,
but harder to make them more stable
Control
landscape & belief
Scz
landscape & belief
Network states
Basin A
Basin B


Dataset 1: Baseline
Dataset 2
Dataset 1: Follow-up
Jar B
Jar A
Confidence in jar
Sequence
Original scale
A
B
C
D
Non-clinical control
Psychotic/Scz
Clinical control
"which jar do you think it is?"
Red sure
Red quite sure
Not sure
Blue quite sure
Blue sure
?

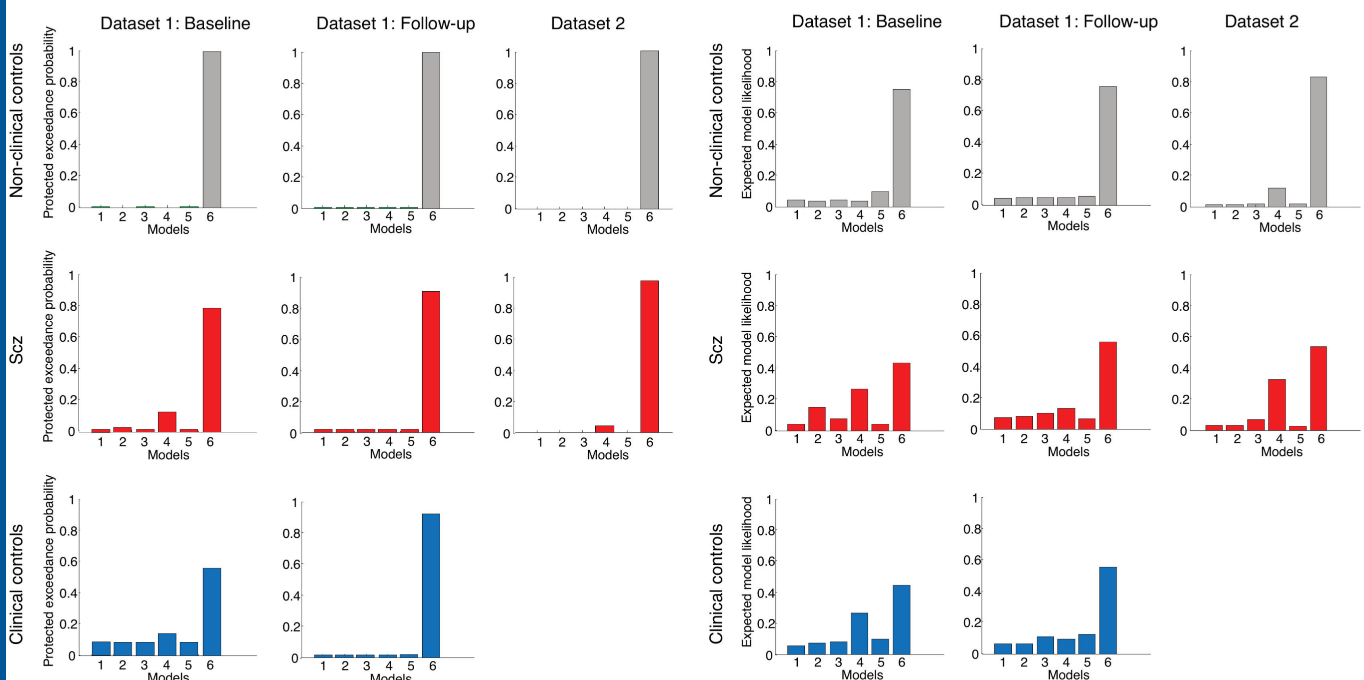


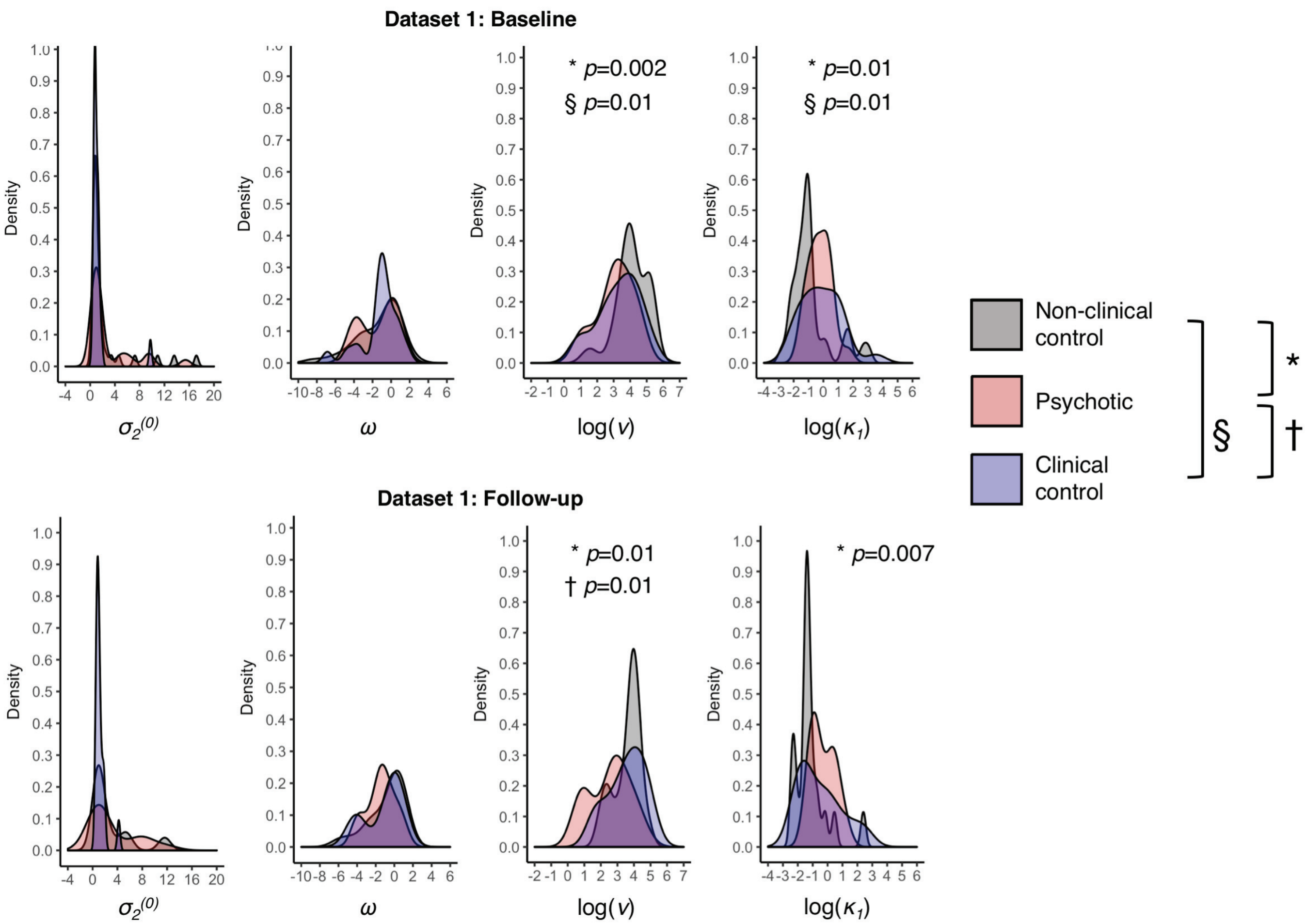
μ_2
Trial number
$\sigma_2^{(0)}$ = 2.5; ω = -1; κ_1 = exp(0); v = exp(4)
$y, u, \hat{\mu}_1$
Trial number
A
B
ω
$x_2^{(k)}$
$x_2^{(k+1)}$
κ_1
$x_1^{(k)}$
$x_1^{(k+1)}$
$u^{(k)}$
$u^{(k+1)}$
$\hat{\mu}_1^{(k+1)}$
$\hat{\mu}_1^{(k+2)}$
$y^{(k)}$
$y^{(k+1)}$
v
Perceptual model
Response model


Models 3 & 4: φ = 0, φ = 0.2, φ = 0.8
u, σ(μ2)
Trial number
A
B
Models 5 & 6: κ1 = exp(-1), κ1 = exp(0), κ1 = exp(1), κ1 = exp(1.2)
u, μ̂1
Models 5 & 6
Shift in belief after seeing bead u = 1 (blue)
Initial belief (μ̂1) that jar = 1 (blue)
Mean absolute shift in beliefs after seeing beads u = 1 or 0
κ1 = exp(-1)
κ1 = exp(-0.6)
κ1 = exp(-0.2)
κ1 = exp(0.2)
κ1 = exp(0.6)
κ1 = exp(1)

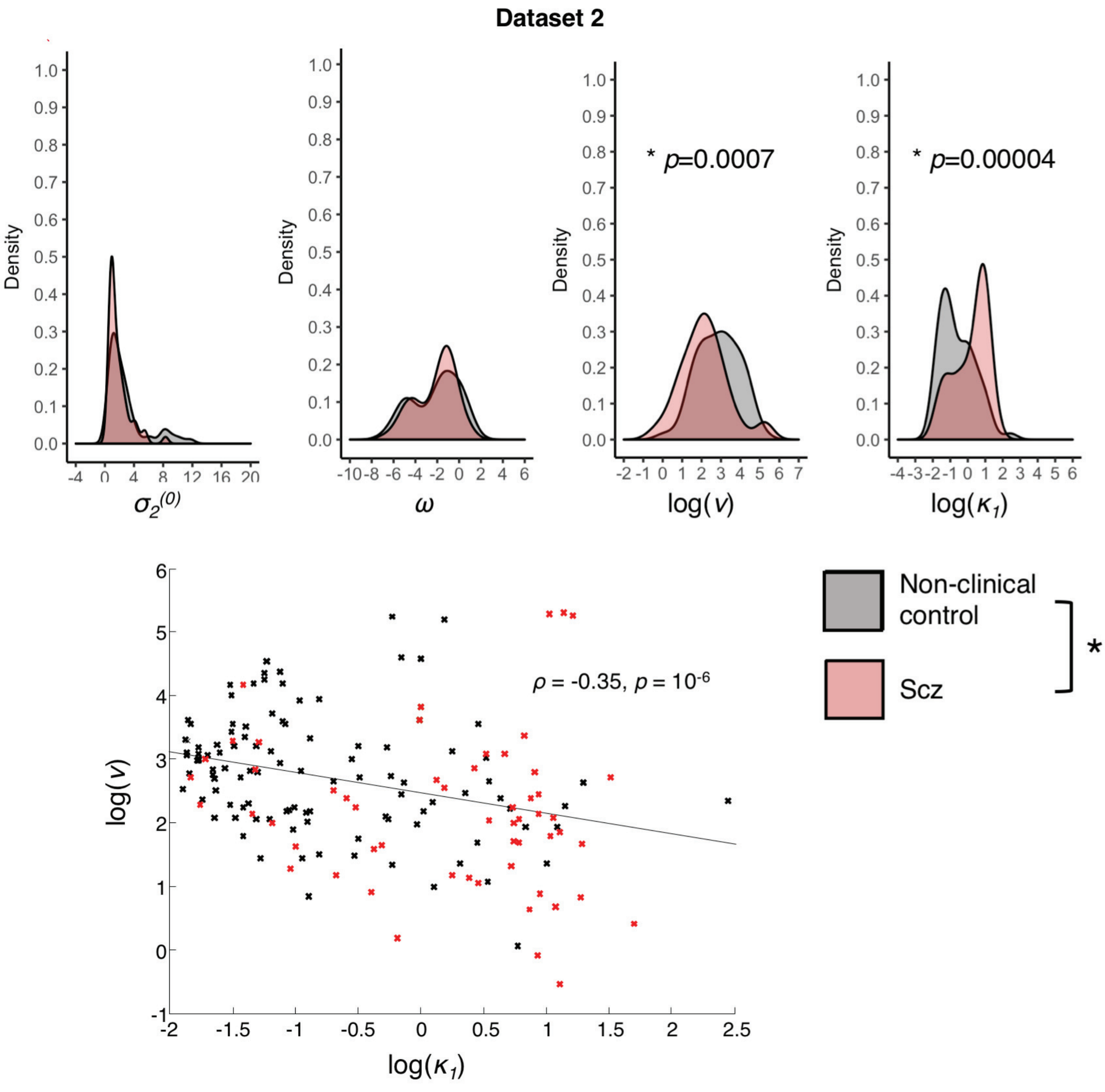


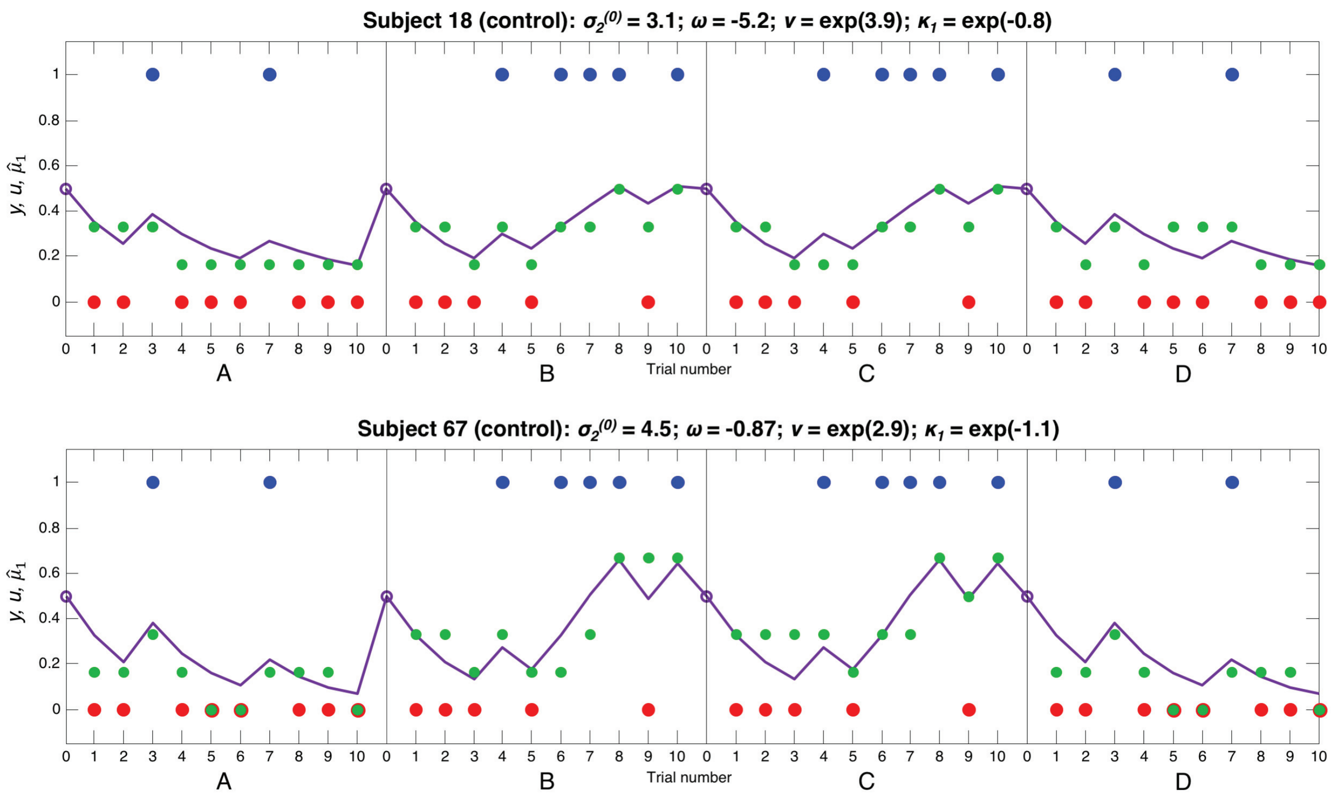
No. of simulations
50
40
30
20
10
0
0
1
2
3
$\sigma_2^{(0)}$
-3
-2
-1
0
ω
2
4
κ_1
log(v)
No. of simulations
$\sigma_2^{(0)}$
ω
κ_1
log(v)


Dataset 1: Baseline
Dataset 1: Follow-up
Dataset 2
Dataset 1: Baseline
Dataset 1: Follow-up
Dataset 2
Non-clinical controls
Scz
Clinical controls
Protected exceedance probability
Expected model likelihood
Models
1 2 3 4 5 6
0 0.2 0.4 0.6 0.8 1

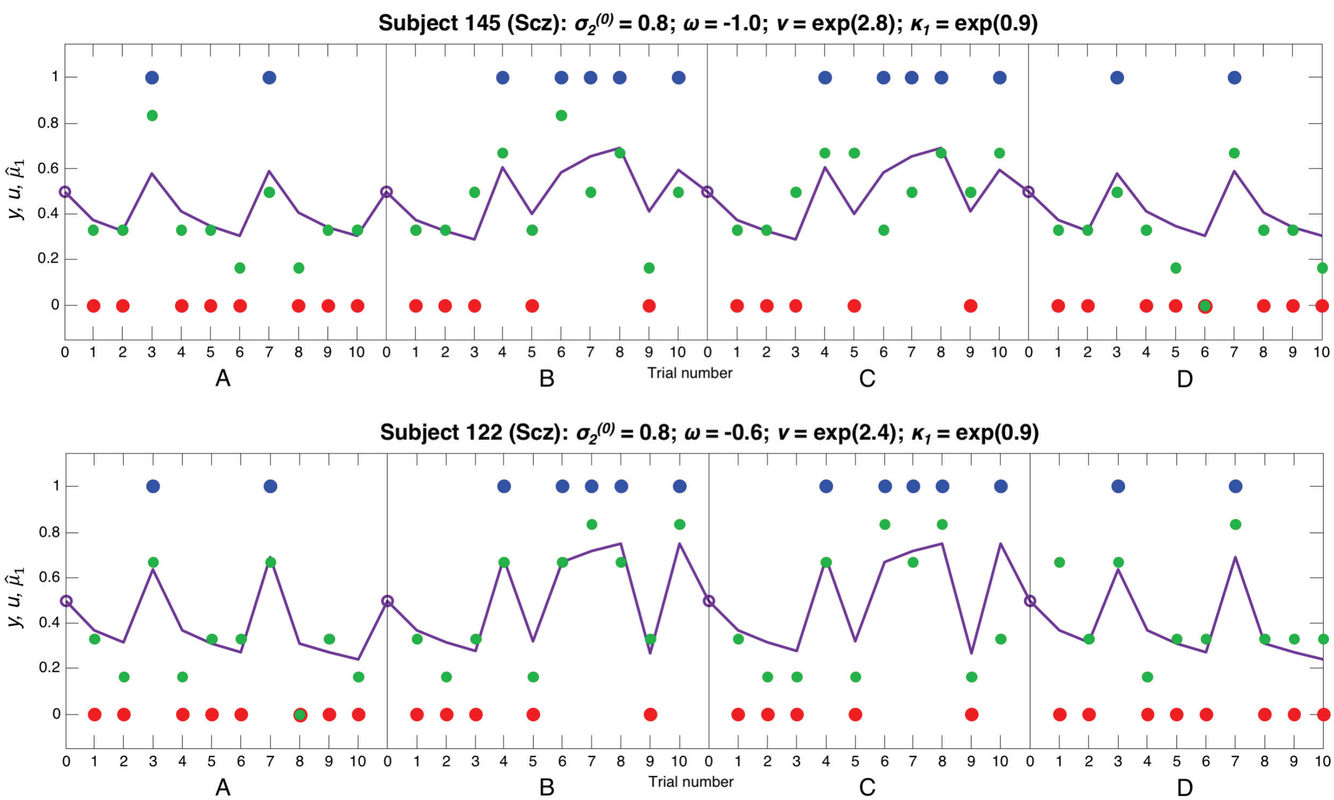


Dataset 1: Baseline
Density
$\sigma_2^{(0)}$
ω
$\log(\nu)$
$\log(\kappa_1)$
* p=0.002
§ p=0.01
* p=0.01
§ p=0.01
Dataset 1: Follow-up
* p=0.01
† p=0.01
* p=0.007
Non-clinical control
Psychotic
Clinical control
*
§
†


Dataset 2
Density
σ2(0)
ω
log(ν)
log(κ1)
* p=0.0007
* p=0.00004
log(ν)
log(κ1)
ρ = -0.35, p = 10-6
Non-clinical control
Scz
*



Subject 18 (control): $\sigma_2^{(0)}$ = 3.1; ω = -5.2; v = exp(3.9); κ_1 = exp(-0.8)
$y, u, \hat{\mu}_1$
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9 10
A
B
C
D
Trial number
Subject 67 (control): $\sigma_2^{(0)}$ = 4.5; ω = -0.87; v = exp(2.9); κ_1 = exp(-1.1)
$y, u, \hat{\mu}_1$
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9 10
A
B
C
D
Trial number



Subject 145 (Scz): $\sigma_2^{(0)}$ = 0.8; ω = -1.0; v = exp(2.8); κ_1 = exp(0.9)
$y, u, \hat{\mu}_1$
Trial number
A
B
C
D
Subject 122 (Scz): $\sigma_2^{(0)}$ = 0.8; ω = -0.6; v = exp(2.4); κ_1 = exp(0.9)
$y, u, \hat{\mu}_1$
Trial number
A
B
C
D