



SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

SISSA Digital Library

Attractor-like dynamics in belief updating in schizophrenia

This is the peer reviewed version of the following article:

*Original*

Attractor-like dynamics in belief updating in schizophrenia / A Adams, Rick; Napier, Gary; P Roiser, Jonathan; Mathys, Christoph Daniel; Gilleen, James. - In: THE JOURNAL OF NEUROSCIENCE. - ISSN 0270-6474. - 38:44(2018), pp. 9471-9485.

*Availability:*

This version is available at: 20.500.11767/83361 since: 2018-10-02T12:17:32Z

*Publisher:*

*Published*

DOI:10.1523/JNEUROSCI.3163-17.2018

*Terms of use:*

openAccess

Testo definito dall'ateneo relativo alle clausole di concessione d'uso

*Publisher copyright*

(Article begins on next page)

---

**Research Articles: Behavioral/Cognitive**

## **Attractor-like dynamics in belief updating in schizophrenia**

**Rick A Adams<sup>1,2</sup>, Gary Napier<sup>1</sup>, Jonathan P Roiser<sup>1</sup>, Christoph Mathys<sup>3,4,5</sup> and James Gilleen<sup>6,7</sup>**

<sup>1</sup>*Institute of Cognitive Neuroscience, UCL, 17 Queen Square, London, WC1N 3AZ, UK*

<sup>2</sup>*Division of Psychiatry, UCL, 6th floor, 149 Tottenham Court Road, London, W1T 7NF, UK*

<sup>3</sup>*Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

<sup>4</sup>*Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland*

<sup>5</sup>*Max Planck UCL Centre for Computational Psychiatry and Ageing Research, 10-12 Russell Square, London, WC1B 5EH, UK*

<sup>6</sup>*Department of Psychology, University of Roehampton, London, SE15 4JD.*

<sup>7</sup>*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data; R.A.A. wrote the first draft of the paper; R.A.A., G.N., J.R., C.M., and J.G. wrote the paper; C.M. contributed unpublished reagents/analytic tools.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](mailto:rick.adams@ucl.ac.uk)

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

**Alerts:** Sign up at [www.jneurosci.org/cgi/alerts](http://www.jneurosci.org/cgi/alerts) to receive customized email alerts when the fully formatted version of this article is published.

Accepted manuscripts are peer-reviewed but have not been through the copyediting, formatting, or proofreading process.

Copyright © 2018 the authors

# 1 Attractor-like dynamics in belief 2 updating in schizophrenia

---

3

4

5 Rick A Adams<sup>1,2\*†</sup>, Gary Napier<sup>1†</sup>, Jonathan P Roiser<sup>1</sup>, Christoph Mathys<sup>3,4,5†</sup>,  
6 James Gilleen<sup>6,7†</sup>

7

8 <sup>1</sup>Institute of Cognitive Neuroscience, UCL, 17 Queen Square, London, WC1N 3AZ,  
9 UK

10 <sup>2</sup>Division of Psychiatry, UCL, 6<sup>th</sup> floor, 149 Tottenham Court Road, London, W1T  
11 7NF, UK

12 <sup>3</sup>Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265,  
13 34136 Trieste, Italy

14 <sup>4</sup>Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering,  
15 University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

16 <sup>5</sup>Max Planck UCL Centre for Computational Psychiatry and Ageing Research, 10-  
17 12 Russell Square, London, WC1B 5EH, UK

18 <sup>6</sup>Department of Psychology, University of Roehampton, London, SE15 4JD.

19 <sup>7</sup>Department of Psychosis Studies; Institute of Psychiatry, Psychology and  
20 Neuroscience, Kings College London, London, SE5 8AF.

21 †equal contribution

22 ††joint senior authors

23

24 \*corresponding author:

25 [rick.adams@ucl.ac.uk](mailto:rick.adams@ucl.ac.uk)

26

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

29

30 Abstract word count: 247 words

31 Introduction word count: 610 words

32 Discussion word count: 1500

33 3 Tables, 10 Figures

34 **Abstract**

35

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

44

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

52

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

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

59  
60 These findings are consistent with the hypothesis that attractor network  
61 instability contributes to belief updating abnormalities in Scz, and suggest that  
62 similar changes may exist during acute illness in other psychiatric conditions.

63

#### 64 **Significance Statement**

65  
66

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

78

79 **Introduction**

80

81

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

96         Several belief-updating abnormalities have been found in Scz using the  
97 'probability estimates' task. The most consistent finding is that Scz (or just Scz  
98 with delusions (Moritz and Woodward, 2005)) change their beliefs *more* than  
99 non-psychiatric controls in response to changes in evidence (Langdon et al.,  
100 2010) – particularly 'disconfirmatory' evidence, i.e. evidence contradicting a  
101 current belief (Garety et al., 1991; Fear and Healy, 1997; Young and Bentall,

102 1997; Peters and Garety, 2006). Another is that probability ratings at the start of  
103 the sequence are higher in currently psychotic (but not in recovered) Scz than in  
104 both clinical and healthy controls (Peters and Garety, 2006), similar to the  
105 ‘jumping to conclusions’ bias in the ‘draws to decision’ version of the task. Others  
106 have also found that Scz update *less* than controls to more *consistent* evidence, in  
107 this (Horga, in preparation) and other paradigms (Averbeck et al., 2010).

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

121         In Scz, attractor states in prefrontal cortex are thought to be less stable, so  
122 it is easier for the network to switch between them, but harder to become more  
123 confident about (i.e. increase the stability of) any particular one (Rolls et al.,  
124 2008). This loss of stable neuronal states – recently demonstrated in two animal  
125 models of Scz (Hamm et al., 2017) – is thought to be due to hypofunction of *N*-  
126 methyl-D-aspartate receptors (NMDARs) or cortical dopamine 1 receptors in Scz

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

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



149 **Methods and Materials**

150 **Subject characteristics**

151  
152 Dataset 1 comprised 23 patients with delusions (18 Scz), 22 patients with non-  
153 psychotic mood disorders, and 35 non-clinical controls (overall, 50 male and 30  
154 female – see Table 1 for details of the groups); the first two groups were selected  
155 from inpatient wards at the Maudsley and the Bethlem Royal Hospitals. All  
156 groups were tested twice (with loss of n=25 from the groups – see Table 1); the  
157 clinical groups were tested once when they were unwell ('baseline'), and again  
158 once they had recovered ('follow-up'). The mean time between testing sessions  
159 was 17.4 (range 6 to 41) weeks in the deluded group, 33.4 (range 4 to 68) weeks  
160 in the clinical control group, and 35.6 (range 27 to 46) weeks in the non-clinical  
161 control group. The deluded group's shorter inter-test interval was due to their  
162 shorter admission period and to the prioritization of their follow-up over the  
163 non-clinical control group. Dataset 1 is described in detail elsewhere (Peters and  
164 Garety, 2006).

165         Dataset 2 comprised 56 subjects with a diagnosis of schizophrenia (Scz)  
166 and 111 controls (overall, 83 male and 84 female – see Table 1). All subjects  
167 provided informed, written consent, and ethical permission for the study was  
168 obtained from the local NHS Research Ethics Committee (Reference  
169 14/LO/0532). Given the National Adult Reading Test (Nelson, 1982) was used to  
170 estimate IQ in these participants, a recruitment condition was that English was  
171 their first language.

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

175

## 176 **Experimental design**

177

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

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

190 We used some of the behavioural measures employed in the original  
191 analysis of dataset 1 (Peters and Garety, 2006) to analyse dataset 2. These were  
192 ‘disconfirmatory updating’, the mean change in belief on seeing a bead of a  
193 different colour to the  $\geq 2$  beads preceding it and ‘final certainty’ (the response to  
194 the last bead). We altered their ‘initial certainty’ measure from the mean  
195 response to the first three beads to the response to the first bead, which comes

196 closer to capturing the classic ‘jumping to conclusions’ bias (in which around  
197 50% of Scz decide on the jar colour after seeing only one bead; (Garety et al.,  
198 1991), although the results of both measures are presented below.

199

## 200 **Computational modelling**

201

202 The optimal way to use sensory information to update one’s beliefs under  
203 conditions of uncertainty is to use Bayesian inference. Neural systems are likely  
204 to approximate Bayesian inference using schemes of simple update equations  
205 (Rao and Ballard, 1999; Friston, 2005); one such model is the Hierarchical  
206 Gaussian Filter (HGF). The HGF is a hierarchical Bayesian inference scheme that  
207 gives a principled account of how beliefs are updated on acquiring new data,  
208 using variational Bayes and individual priors. Variational Bayesian schemes (e.g.  
209 (Beal, 2003) use analytic equations to derive an exact solution to an  
210 approximation of the posterior distribution over the latent variables and  
211 parameters (as opposed to sampling methods which approximate a solution to  
212 the exact posterior). The HGF has been used as a generic state model for learning  
213 under uncertainty and has repeatedly been shown to outperform similar  
214 approaches, such as reinforcement learning models with fixed (e.g. Rescorla-  
215 Wagner) or dynamic (e.g. (Sutton, 1992) learning rates (Iglesias et al., 2013;  
216 Diaconescu et al., 2014; Hauser et al., 2014; Vossel et al., 2014). One advantage of  
217 the HGF is that it contains subject-specific parameters (and prior beliefs) that  
218 can account for between-subject differences in learning whilst preserving the  
219 (Bayes) optimality of any individual’s learning (relative to his/her model

220 parameters and prior beliefs). These parameters may be encoded by tonic levels  
221 of neuromodulators such as dopamine (Marshall et al., 2016), or by the intrinsic  
222 properties of neuronal networks (e.g. the ratio of excitatory to inhibitory neural  
223 activity can affect both the speed of evidence accumulation (Lam et al., 2017) –  
224 analogous to the evolution rate in the HGF – and also response stochasticity  
225 (Murray et al., 2014)). Differences in model parameters between Scz and  
226 controls may therefore explain, in computational terms, how pathophysiology  
227 leads to abnormal inference (Adams et al., 2015).

228         In general, when modelling behaviour under Bayesian assumptions, it is  
229 necessary to distinguish between the model of the world used by the subject (the  
230 perceptual model) and a model of how a subject’s beliefs translated into  
231 observed behaviour (the observation or response model). Most of the  
232 parameters pertain to the perceptual model (here, all parameters except  
233 response stochasticity  $v$  – see Table 2) and reflect (inferred) neuronal  
234 processing. In contrast, the parameters of the response model link subjective  
235 states to behavioural outcomes, and thus may reflect stochasticity in neuronal  
236 processing, measurement noise (in some paradigms), or non-random effects that  
237 have not been captured by the perceptual model. This and related learning  
238 models are freely available from  
239 <http://www.translationalneuromodeling.org/tapas/> (version 5.1.0): this  
240 analysis used the perceptual models ‘hgf\_binary’ or ‘hgf\_ar1\_binary’ and the  
241 response model ‘beta\_obs’.

242         At the bottom of the model (Figure 3 shows some simulated responses) is  
243 the bead drawn  $u^{(k)}$  on trial  $k$  and the probability  $x_1^{(k)}$  that draws are coming  
244 from the blue jar. At the level above this is  $x_2$ , the tendency towards the blue jar

245 (a transform of the probability, bounded by  $\pm\infty$ ); by definition,  $x_1 = s(x_2)$ , where  
 246  $s(\bullet)$  is the logistic sigmoid function. As  $x_2$  approaches infinity, the probability of  
 247 the blue jar approaches 1; as it approaches minus infinity, the probability of the  
 248 blue jar approaches 0. For  $x_2 = 0$ , both jars are equally probable. This quantity is  
 249 hidden from the subject and must be inferred: the subject's posterior estimate of  
 250  $x_2$  is  $\mu_2$ , and the subject's posterior estimate of the probability of the jar being  
 251 blue on trial  $k$  is  $s(\mu_2^{(k)})$  – equivalent to the prediction (denoted by  $\hat{\mu}_1$ ) on the next  
 252 trial  $\hat{\mu}_1^{(k+1)}$ .

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

$$256 \quad \hat{\mu}_1^{(k)} \equiv s(\kappa_1 \mu_2^{(k-1)})$$

$$257 \quad \hat{\pi}_1^{(k)} \equiv \frac{1}{\hat{\mu}_1^{(k)}(1 - \hat{\mu}_1^{(k)})}$$

$$258 \quad \hat{\pi}_2^{(k)} \equiv \frac{1}{\sigma_2^{(k-1)} + \exp(\omega)}$$

259 Note that in Models 1-4,  $\kappa_1$  is fixed to 1. A new input  $u^{(k)} \equiv \mu_1^{(k)}$  generates  
 260 a prediction error  $\delta_1^{(k)}$  and the model updates and generates a new prediction as  
 261 follows:

$$262 \quad \delta_1^{(k)} \equiv \mu_1^{(k)} - \hat{\mu}_1^{(k)}$$

$$263 \quad \pi_2^{(k)} = \hat{\pi}_2^{(k)} + \frac{\kappa_1^2}{\hat{\pi}_1^{(k)}}$$

$$264 \quad \mu_2^{(k)} = \mu_2^{(k-1)} + \frac{\kappa_1}{\pi_2^{(k)}} \delta_1^{(k)}$$

$$265 \quad \hat{\mu}_1^{(k+1)} \equiv s(\kappa_1 \mu_2^{(k)})$$

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

269           We parameterize the beta distribution in terms of its mean  $\mu$  and  
270 precision  $\nu$ . These sufficient statistics relate to the conventional  
271 parameterization in terms of the sufficient statistics  $\alpha$  and  $\beta$  by the following  
272 bijection:

$$273 \quad \mu := \frac{\alpha}{\alpha + \beta}$$

$$274 \quad \nu := \alpha + \beta$$

275           Note that updates to  $\mu_2$  are driven by the product of the prediction error  
276 from Bayesian updating explained above and a learning rate which, crucially, can  
277 change over time: this is an important aspect of the HGF in contrast to learning  
278 models such as Rescorla-Wagner that have a fixed learning rate. Parameters  
279 which affect the degree to which  $\mu_2$  can change during the experiment include  $\omega$ ,  
280  $\varphi$ ,  $\kappa_1$  and  $\sigma_2^{(0)}$ . The contributions of  $\varphi$  and  $\kappa_1$  are illustrated in Figure 4 (left  
281 panels).

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

289 In Models 1 and 2, changes in  $x_2$  from trial to trial occur only according to  
 290 the evolution rate  $\omega$ , the variance of the random process at the second level.  
 291 These models were equivalent to the subsequent models with either  $\varphi$  (Models 3  
 292 and 4) fixed to 0 or  $\kappa_1$  (Models 5 and 6) fixed to 1.

293 In Models 3 and 4, changes in  $x_2$  from trial to trial occur according to an  
 294 autoregressive (AR(1)) process that is controlled by three parameters:  $m$ , the  
 295 level to which  $x_2$  is attracted,  $\varphi$ , the rate of change of  $x_2$  towards  $m$ , and  $\omega$ , the  
 296 variance of the random process:

$$297 \quad p(x_2^{(k+1)}) \sim \mathcal{N}(x_2^{(k)} + \varphi(m - x_2^{(k)}), \exp(\omega))$$

298 After inversion, the evolution of  $x_2$  according to this equation is reflected in the  
 299 prediction of  $\mu_2$ :

$$300 \quad \hat{\mu}_2^{(k+1)} = \mu_2^{(k)} + \varphi(m - \mu_2^{(k)})$$

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

305 In Models 5 and 6, changes in  $\mu_2$  from trial to trial occur according to two  
 306 parameters:  $\omega$ , the variance of the random process, and  $\kappa_1$ , a scaling factor that  
 307 changes the size of updates when  $\hat{\mu}_1 = 0.5$ , or maximum uncertainty, relative to  
 308 when  $\hat{\mu}_1$  is closer to 0 or 1, i.e. when the subject is more confident about either  
 309 jar. Figure 4 (lower left panel) illustrates the effect of  $\kappa_1$  on  $\hat{\mu}_1$  over time.

310 Formally, the scaling occurs as:

$$311 \quad \hat{\mu}_1^{(k+1)} \equiv s(\mu_2^{(k)} \kappa_1)$$

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

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

334           As group differences in initial updating had been observed in dataset 1,  
335 we also estimated the standard deviation of  $\mu_2$  before the sequence begins,  $\sigma_2^{(0)}$ ,  
336 in Models 2, 4 and 6.



337 NB for intermediate values of  $\kappa_1$ , Models 5 and 6 produce similar belief  
338 updating trajectories to Models 3 and 4 (containing the disconfirmatory updating  
339 parameter  $\varphi$ ): both make greater updates following disconfirmatory evidence.  
340 For more extreme values of  $\kappa_1$ , however, Models 5 and 6 produce trajectories  
341 that Models 3 and 4 cannot:  $\varphi$  cannot pull beliefs far towards certainty in the  
342 opposite jar (c.f. brown line in Figure 4, lower left panel), and neither can it make  
343 it *more* difficult to update to disconfirmatory evidence (c.f. black line in Figure 4,  
344 lower left panel).

345 The parameters  $\omega$  and  $\nu \pm \sigma_2^{(0)} \pm \varphi$  or  $\kappa_1$  were estimated individually  
346 for each subject. If estimated, the prior probability distributions for their values  
347 are given in Table 2. The means given here refer to the parameters' native space,  
348 but the variances refer not to the parameters' native space, which in many cases  
349 is bounded, but to the unbounded space they were transformed to for estimation  
350 purposes. Otherwise they were fixed as  $\varphi = 0$  (Models 1 and 2) and  $\sigma_2^{(0)} =$   
351 0.006 (Models 1, 3 and 5). The model's prior beliefs about the jars at the start of  
352 the sequence were fixed at  $\mu_2^{(0)} = 0$  (i.e. believing each to be equally likely). The  
353 priors were sufficiently uninformative to be easily updated by the data: all prior  
354 means are standard for the HGF except  $\sigma_2^{(0)}$ , which had to be increased from  
355 0.006 to 0.8 to allow the data to change it. The latter change ensured that group  
356 differences in initial belief updating alone would cause group differences in  $\sigma_2^{(0)}$   
357 rather than  $\kappa_1$ .

358

### 359 **Model fitting and statistical analysis**

360

361 We tested models with different combinations of parameters  $\omega$ ,  $\nu$ ,  $\varphi$  or  $\kappa_1$  and  
362  $\sigma_2^{(0)}$  (see Table 2). In analysing dataset 2, we concatenated all four sequences for  
363 each subject in order to estimate the model parameters as accurately as possible  
364 (resetting the beliefs about the jars at the start of each sequence).

365         After fitting the six models to each subject's data, we performed Bayesian  
366 model selection on all groups separately in both dataset 1 (at baseline and  
367 follow-up) and dataset 2. This procedure weights models according to their  
368 accuracy but penalises them for complexity (i.e. unnecessary extra parameters)  
369 to prevent overfitting (Stephan et al., 2009; Rigoux et al., 2014). The winning  
370 model in all eight groups was Model 6 (Figure 6), although around a third of  
371 psychotic subjects and non-clinical controls in dataset 1 (at baseline) and in  
372 dataset 2 were better fit by Model 4. It is unclear why this change occurs, but  
373 given that Model 6 can produce very similar trajectories to Model 4 for  
374 intermediate values of  $\kappa_1$  (Figure 4), any increase in response stochasticity is  
375 likely to diminish the strength of evidence for one model over a similar one.

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

386

387 **Results**

388

389 **Behavioural results: dataset 1**

390

391 Each group's mean responses are plotted in Figure 2A, and statistical tests  
392 detailed in Table 1 ( $p(adj)$  refers to the adjusted  $p$  value of Tukey's HSD *post hoc*  
393 test). As described previously (Peters and Garety, 2006), at baseline there was a  
394 significant difference in disconfirmatory updating between the groups ( $F(2,77) =$   
395  $6, p = 0.004$ , ANOVA), and the psychotic group had greater disconfirmatory  
396 updating than the non-clinical controls ( $p(adj) = 0.003$ ) but not the clinical  
397 controls ( $p(adj) = 0.4$ ). There was no difference between the clinical and non-  
398 clinical controls ( $p(adj) = 0.13$ ). There were also significant differences in initial  
399 certainty across the three groups ( $F(2,77) = 8.7, p = 0.0004$ , ANOVA); the  
400 psychotic group's initial certainty was higher than the non-clinical controls'  
401 ( $p(adj) = 0.0003$ ) but not the clinical controls' ( $p(adj) = 0.25$ ). There wasn't a  
402 significant difference between the clinical and non-clinical control groups ( $p(adj)$   
403  $= 0.06$ ). There were no group differences in final certainty ( $F(2,77) = 0.7, p = 0.5,$   
404 ANOVA).

405 At follow-up, the difference in disconfirmatory updating between the  
406 groups was no longer significant ( $F(2,52) = 2.9, p = 0.06$ , ANOVA); the psychotic  
407 group had greater disconfirmatory updating than the non-clinical controls  
408 ( $p(adj) = 0.049$ ) but not the clinical controls ( $p(adj) = 0.4$ ). There was no

409 significant difference in initial certainty across the groups ( $F(2,52) = 0.9, p = 0.4,$   
410 ANOVA). Differences in final certainty were no longer significant ( $F(2,52) = 2.8, p$   
411  $= 0.07,$  ANOVA); the biggest difference was the non-clinical controls' final  
412 certainty which was numerically higher than the clinical controls' ( $p(adj) =$   
413  $0.057$ ).

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

418

#### 419 **Behavioural results: dataset 2**

420

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

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

436

#### 437 **Modelling results: dataset 1**

438

439 Model selection results for the three groups analysed separately at both baseline  
440 and follow-up are plotted in Figure 6 (columns 1, 2, 4 and 5); the probability of  
441 each model being best for any given subject is shown in the left panel, and the  
442 probability of each model being the best overall is shown in the right panel.

443 Model 6 is the clear winner at each time point, although a minority of psychotic  
444 and clinical controls are best fit by Model 4.

445 Model 6's parameter distributions are shown in Figure 7; they are  
446 skewed, hence non-parametric tests were used to determine group differences  
447 (full details in Table 3;  $p(adj)$  refers to the adjusted  $p$  value of Dunn's *post hoc*  
448 test). At baseline there were large group differences in belief instability  $\kappa_1$   
449 ( $\chi^2(2, n=80) = 9.64, p = 0.008, \eta^2 = 0.12$ , Kruskal-Wallis' one-way ANOVA on  
450 ranks) and response stochasticity  $\nu$  ( $\chi^2(2, n=80) = 11.9, p = 0.003, \eta^2 = 0.15$ ) but  
451 not in  $\sigma_2^{(0)}$  or  $\omega$ . There were statistically significant differences in  $\kappa_1$  between the  
452 non-clinical controls and both the psychotic group ( $p(adj) = 0.01$ , Dunn's test)  
453 and the clinical control group ( $p(adj) = 0.01$ ), but not between the latter two  
454 groups ( $p(adj) = 0.4$ ). Similarly, there were statistically significant differences in  
455  $\nu$  between the non-clinical controls and both the psychotic group ( $p(adj) = 0.002$ ,

456 Dunn's test) and the clinical control group ( $p(adj) = 0.01$ ), but not between the  
457 latter two groups ( $p(adj) = 0.3$ ).

458 At follow-up, there were still large group differences in  $\kappa_1$  ( $\chi^2(2,n=55) =$   
459  $8.0, p = 0.02, \eta^2 = 0.15$ , Kruskal-Wallis' one-way ANOVA on ranks) and  $\nu$   
460 ( $\chi^2(2,n=55) = 8.5, p = 0.01, \eta^2 = 0.16$ ) but not in  $\sigma_2^{(0)}$  or  $\omega$ . There was a significant  
461 difference in  $\kappa_1$  between the psychotic and non-clinical control groups ( $p(adj) =$   
462  $0.007$ , Dunn's test) but not the clinical and non-clinical control groups ( $p(adj) =$   
463  $0.1$ );  $\nu$  remained significantly different between the non-clinical controls and  
464 both the psychotic group ( $p(adj) = 0.01$ , Dunn's test) and now also between the  
465 psychotic and clinical control groups ( $p(adj) = 0.01$ ), but not between the clinical  
466 and non-clinical controls ( $p(adj) = 0.5$ ).

467 We explored whether group differences in  $\kappa_1$  or  $\nu$  at baseline and follow  
468 up might be ascribable to IQ (Quick Test score (Ammons and Ammons, 1962)),  
469 as the groups' IQ scores were not equivalent (Table 1). Including both IQ and  
470 group status within one regression model is an unsound method of testing for  
471 confounding by IQ because group and IQ are clearly not independent here (Miller  
472 and Chapman, 2001), so we tested for relationships between the parameters and  
473 IQ separately within each group at each time point. No relationships reached  
474 statistical significance (all  $p > 0.1$ ), the closest being a trend between  $\kappa_1$  and IQ in  
475 non-clinical controls only ( $r = -0.30, p = 0.08$ ); nevertheless, given the smaller  
476 group sizes and larger between- versus within-group variances, it remains  
477 plausible that IQ differences contribute to group parameter differences.

478 We tested whether  $\kappa_1$  or  $\nu$  at baseline related to delusion-proneness  
479 (Peters Delusion Inventory score) across all groups, after first excluding any  
480 interaction between PDI and group; PDI significantly correlated with  $\nu$  ( $F(1,67) =$

481 7.1,  $p = 0.01$ , ANCOVA) but not  $\kappa_1$  ( $F(1,67) = 3.2$ ,  $p = 0.079$ , ANCOVA). We tested  
482 whether  $\kappa_1$  or  $\nu$  at baseline was correlated with any particular subgroup of  
483 symptoms (measured using the Manchester Scale (Krawiecka et al., 1977)) in  
484 both clinical groups only, using the regression models  $\kappa_1$  [or  $\nu$ ]  $\sim$  const +  
485  $\nu_1$ \*MSaffective +  $\nu_2$ \*MSpositive +  $\nu_3$ \*MSnegative: none of the models were  
486 significant, however (all  $p > 0.1$ ).

487 At baseline, there was no evidence of a correlation between  $\kappa_1$  and  
488 antipsychotic medication dose ( $p = 0.3$ ), but the correlation between  $\nu$  and  
489 medication dose approached significance ( $\rho = -0.4$ ,  $p = 0.067$ ).

490 We tested for correlations between the Model 6 parameters (Spearman's  
491  $\rho$  was used where distributions were not parametric):  $\kappa_1$  and  $\nu$  were negatively  
492 correlated both at baseline ( $\rho = -0.38$ ,  $p = 0.0004$ ) and at follow up ( $\rho = -0.52$ ,  $p =$   
493  $0.0001$ ), as were  $\kappa_1$  and  $\omega$  at baseline ( $\rho = -0.47$ ,  $p = 10^{-5}$ ) and follow up ( $\rho = -$   
494  $0.53$ ,  $p = 10^{-5}$ ). In estimating the parameters from *simulated* data, the only  
495 correlation present in both simulations (indicating some consistent trading-off  
496 between these parameters during estimation) was between  $\kappa_1$  and  $\omega$ , with  $r = -$   
497  $0.5$  in each case. This is not surprising, as both  $\kappa_1$  and  $\omega$  affect updating to new  
498 information throughout the sequence (unlike  $\sigma_2^{(0)}$ ) in a deterministic way (unlike  
499  $\nu$ ). Nevertheless,  $\kappa_1$  was estimated very reliably in the first simulation (Figure 5,  
500 top row) and with reasonable accuracy in the second (Figure 5, bottom row), so  
501 we are confident that the group differences in  $\kappa_1$  are genuine. The correlations of  
502  $\rho \approx -0.5$  between  $\omega$  and  $\kappa_1$  in dataset 1 are unlikely to be reliable, however.

503

504 **Modelling results: dataset 2**

505

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

511 As in dataset 1, belief instability  $\kappa_1$  was significantly higher in Scz than in  
512 controls ( $Z = -5.6$ ,  $p = 10^{-8}$ , Mann-Whitney U test) with a medium-to-large effect  
513 size ( $r = 0.43$ ); also response stochasticity  $\nu$  was lower in Scz than in controls ( $Z$   
514  $= 3.9$ ,  $p = 0.0001$ ,  $r = 0.3$ , Mann-Whitney U test), as was initial belief variance  $\sigma_2^{(0)}$   
515 ( $Z = 3.1$ ,  $p = 0.002$ ,  $r = 0.24$ , Mann-Whitney U test). There were no statistically  
516 significant group differences in evolution rate  $\omega$ . See Figures 6 and 7 for  
517 examples of model fits in subjects with lower  $\kappa_1$  values (two controls in Figure 9)  
518 and higher  $\kappa_1$  values (two Scz subjects in Figure 10); each figure also illustrates  
519 the effects of lower and higher  $\omega$  values (in the top and bottom rows  
520 respectively). We repeated the analysis using a subset of the controls ( $n=60$ ) that  
521 were better matched in age and sex, as the original control group was younger  
522 and more female than the patient group (Table 1). The group differences in  $\kappa_1$   
523 and  $\nu$  were unchanged in this analysis ( $Z = -4.1$ ,  $p = 0.00004$ ;  $Z = 3.4$ ,  $p = 0.0007$   
524 respectively, Mann-Whitney U tests), but that in  $\sigma_2^{(0)}$  was no longer significant ( $Z$   
525  $= 1.9$ ,  $p = 0.056$ , Mann-Whitney U test).

526 Although IQ (National Adult Reading Test score (Nelson, 1982)) was  
527 evenly matched in these groups, working memory (Letter Number Sequencing  
528 score (Wechsler, 1997)) was lower in Scz than in controls (see Table 1). We



529 explored whether the group parameter differences might be related to working  
530 memory, by testing for correlations between  $\kappa_1$  or  $\nu$  and working memory in each  
531 group separately (Miller and Chapman, 2001): none were statistically significant  
532 (all  $p > 0.1$ ). We also tested for relationships between  $\kappa_1$  or  $\nu$  and IQ (NART) in  
533 each group:  $\nu$  and IQ (NART) were correlated in Scz ( $r = 0.33, p = 0.014$ ), but no  
534 other relationships were significant (all  $p > 0.1$ ).

535 We tested whether  $\kappa_1$  or  $\nu$  related to schizotypy (Schizotypal Personality  
536 Questionnaire score) across all groups but neither did so (both  $p = 0.4$ , ANCOVA).  
537 We tested whether  $\kappa_1$  or  $\nu$  were predicted by any particular subgroup of  
538 symptoms (measured using the Positive and Negative Symptom Scale (Kay et al.,  
539 1987)) in the Scz group only, using the regression model  $\kappa_1$  [or  $\nu$ ]  $\sim$  const +  
540  $\nu_1$ \*PANSSgeneral +  $\nu_2$ \*PANSSpositive +  $\nu_3$ \*PANSSnegative: the  $\kappa_1$  model was not  
541 significant ( $F = 0.9, p = 0.4$ ), but  $\nu$  was weakly predicted by negative symptoms  
542 (overall  $F = 2.76, p = 0.051$ ; for  $\nu_3, t = -2.1, p = 0.04$ ). We had no record of  
543 medication dose in dataset 2.

544 We tested for correlations between the Model 6 parameters: as in dataset  
545 1,  $\kappa_1$  and  $\nu$  were negatively correlated (Figure 8;  $\rho = -0.35, p = 10^{-6}$ ), but unlike  
546 dataset 1, the only other statistically significant correlation was between  $\kappa_1$  and  
547  $\sigma_2^{(0)}$  ( $\rho = -0.54, p = 10^{-13}$ ). There was a correlation of  $r = -0.2$  between  $\kappa_1$  and  $\nu$  in  
548 the data simulated from modal Scz parameter values (Figure 5, bottom row), but  
549 no correlation in the first. This implies that the consistent correlations between  
550 these parameters of  $\rho = -0.38, \rho = -0.52$  (dataset 1 baseline and follow-up) and  $\rho$   
551  $= -0.35$  (dataset 2) are unlikely to be just estimation artefacts. The only other  
552 correlation between parameters in the simulated data was between  $\sigma_2^{(0)}$  and  $\kappa_1$ ,

553 of  $r = -0.25$ , in the first simulation only. These parameters were correlated in  
554 dataset 2 but not dataset 1.

## 555 **Discussion**

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

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

574         The unique features of Model 6 that make attractor dynamics a  
575 compelling neurobiological explanation for its dominance are both Scz and  
576 controls' non-linearities in belief updating to confirmatory versus

577 disconfirmatory evidence. The Scz group updated its beliefs (sometimes much)  
578 more to disconfirmatory than confirmatory evidence – particularly at points of  
579 relative certainty about the jar – and the controls were the opposite. Models with  
580 uniformly high or low learning rates cannot reproduce these effects; and adding  
581 high- or low-level (sensory) uncertainty to a hierarchical model would lead to  
582 uniformly high or low learning rates respectively. Although Models 3 and 4 do  
583 show differential updating to confirmatory vs disconfirmatory evidence, this  
584 results in beliefs in either jar hovering around 0.5 (as in Figure 4, top left) rather  
585 than making large updates from belief in one jar to the other (as when  $\kappa_l =$   
586  $\exp(1.2)$ : Figure 4, bottom left). Furthermore, degraded neuronal ensemble firing  
587 (consistent with unstable attractor states) has recently been shown to be  
588 common to two different mouse models of schizophrenia (Hamm et al., 2017).

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

599 **Parameter relationships with cognition and symptoms**

600

601           Neither  $\kappa_1$  nor  $\nu$  showed significant relationships with IQ (in dataset 1) or  
602 working memory (in dataset 2) within the groups, giving some indication that  
603 the group differences in these cognitive measures were unlikely to be the main  
604 drivers of group differences in the parameters. Nevertheless, aside from the  
605 correlation between response stochasticity  $\nu$  and IQ in dataset 2, it is perhaps  
606 surprising that there weren't more relationships between  $\kappa_1$  or  $\nu$  and cognitive  
607 measures in Scz, given it is likely that abnormal prefrontal dynamics have  
608 profound effects on all these variables. We may have lacked power to detect  
609 them – though dataset 2 had 80% power to detect a correlation of 0.33 – or  
610 perhaps different prefrontal regions contribute to working memory, IQ and  
611 belief updating.

612           One might also question why there were no strong relationships between  
613  $\kappa_1$  or  $\nu$  and positive or negative symptom domains (negative symptoms were  
614 weakly associated with  $\nu$  in dataset 2 only). Again, power may have been an  
615 issue, although note that across all subjects in dataset 1, response stochasticity  $\nu$   
616 was associated with PDI score even after including group in the model, indicating  
617 a potential relationship with delusions, but not with the broader concept of  
618 schizotypy (assessed in dataset 2). It is also likely that other pathological factors  
619 contribute to symptoms, beyond those measured here (e.g. striatal dopamine  
620 availability and positive symptoms). Of note, two other computational studies  
621 demonstrating clear working memory parameter differences between Scz and  
622 controls also failed to detect any relationship between those parameters and  
623 symptom domains (Collins et al., 2014, 2017). Both their and our Scz groups

624 were taking antipsychotic medication, which is also likely to weaken correlations  
625 of parameters to positive symptoms.

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

636         It is interesting that non-clinical controls’ data were also best fit by Model  
637 6 in both datasets, implying that even healthy subjects show some asymmetry in  
638 their belief updating to expected versus unexpected evidence. Most non-clinical  
639 control subjects had  $\kappa_1 < 1$ , i.e. reduced updating to changing evidence.

#### 640 **Related modelling studies**

641         How do these findings relate to other computational modelling work in  
642 Scz? A study of unmedicated, mainly first episode Scz performing a reversal  
643 learning task (Schlagenhauf et al., 2013) also demonstrated an increased  
644 tendency to switch that was not accounted for by reward sensitivity (which  
645 would be affected by more stochastic behaviour), and increased switching also  
646 occurs in chronic Scz (Waltz et al., 2013), although not always (Pantelis et al.,  
647 1999).

648           Two recent studies of similar tasks in Scz populations have also  
649 demonstrated evidence of non-linear belief updating. (Jardri et al., 2017) showed  
650 that the Scz group on average “overcount” the likelihood in a single belief update;  
651 an effect they attribute to reverberating cortical message-passing, but which  
652 could also be due to the belief instability shown by Model 6. (Stuke et al., 2017)  
653 showed in a very similar task that all subjects showed evidence of non-linear  
654 updating, but the Scz group updated more than controls to “irrelevant  
655 information” (i.e. disconfirmatory evidence). Some differences between their  
656 model and ours are that they did not estimate response stochasticity in their  
657 subjects (neither did (Jardri et al., 2017), and their ‘non-linearity’ parameter was  
658 bounded by linear updating on one side, roughly equivalent to belief instability  
659  $\kappa_1$  being constrained to being  $<1$  in our model, whereas we have shown (as in  
660 (Jardri et al., 2017) that Scz belief updating is often beyond this bound (Figure 7),  
661 and more stochastic. Conversely, (Moutoussis et al., 2011) demonstrated  
662 increased response stochasticity in acutely psychotic subjects, but did not test  
663 for differences in belief updating.

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

672   percepts were driven by prior expectations, leading to a reduction in the  
673   updating of their beliefs (relative to controls).

674           Further evidence for heterogeneity in Scz is that those with delusions  
675   have greater certainty about the hypothesis that matches the evidence at every  
676   stage (Speechley et al., 2010), unlike the reduced final certainty we observed in  
677   Scz in dataset 2. On the other hand, Scz with high negative symptoms have  
678   difficulty choosing the most rewarding option very consistently (Gold et al.,  
679   2012), which may reflect a lack of certainty about its value. We lacked sufficient  
680   power to detect differences between Scz with exclusively high positive or  
681   negative symptoms, however.

#### 682   **Limitations**

683           Each of our datasets contains some limitations of the beads task that are  
684   addressed by the other. Dataset 1 did not include a memory aid or measure  
685   working memory, but dataset 2 did both, and dataset 2 also matched IQ across  
686   groups much better than dataset 1; dataset 2 used a Likert scale for responding  
687   and so could potentially exaggerate small changes in belief updating, but dataset  
688   1 used a continuous measure; dataset 2 only tested stable outpatients, but  
689   dataset 1 tested more unwell inpatients and retested them once they were  
690   better. The main limitation common to both datasets is that all subjects with  
691   psychotic diagnoses were taking antipsychotic medication when tested. Although  
692   the correlation between  $v$  and medication dose was almost significant in dataset  
693   1, this relationship seems likely to be driven by illness severity rather than  
694   medication itself. Dopamine 2 receptor antagonists seem to both reduce  
695   overconfidence in probabilistic reasoning (Andreou et al., 2014), and also

696 reduce motor response variability (Galea et al., 2013) and so if anything likely  
697 reduce our group differences.

## 698 **Conclusion**

699 In conclusion, we have shown that Scz subjects in two independent beads  
700 task datasets have consistent differences in two parameters of a belief updating  
701 model that attempts to reproduce consequences of attractor network instability.  
702 Note that this study was designed to link patterns of inferences to model  
703 parameters that (do or don't) mimic the effects of abnormal attractor states on  
704 belief updating. The HGF itself does not contain attractor states and no relation  
705 between its parameters and NMDAR function has hitherto been tested. More  
706 detailed spiking network modelling, pharmacological (or other NMDAR)  
707 manipulations and imaging are required in future to understand how  
708 neuromodulatory function in both pyramidal cells and inhibitory interneurons  
709 contributes to real attractor dynamics and probabilistic inference, and to seek  
710 empirical evidence for a correspondence between the stability of network states  
711 and the stability of its inferences (especially in schizophrenia). This work  
712 underscores the importance of relating psychological biases to their underlying  
713 computational mechanisms, and thence (in future) to the constraints – e.g. the  
714 hypofunction of NMDARs – that neurobiology imposes on these mechanisms.  
715

## 716 **Acknowledgements**

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



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

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

721

722

723

## 724 **Conflicts of Interest**

725 No authors have any biomedical financial interests or potential conflicts of

726 interest.

727

## 728 **References**

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

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

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

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

740 Averbeck BB, Evans S, Chouhan V, Bristow E, Shergill SS (2010) Probabilistic  
741 learning and inference in schizophrenia. *Schizophr Res* Available at:  
742 <http://www.ncbi.nlm.nih.gov/pubmed/20810252> [Accessed November  
743 18, 2010].

744 Beal MJ (2003) Variational algorithms for approximate Bayesian inference.  
745 Available at:  
746 <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.131.995>  
747 1 [Accessed March 26, 2012].

- 748 Brunel N, Wang XJ (2001) Effects of neuromodulation in a cortical network  
749 model of object working memory dominated by recurrent inhibition. *J*  
750 *Comput Neurosci* 11:63–85.
- 751 Collins AGE, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017) Interactions  
752 Among Working Memory, Reinforcement Learning, and Effort in Value-  
753 Based Choice: A New Paradigm and Selective Deficits in Schizophrenia.  
754 *Biol Psychiatry* 82:431–439.
- 755 Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ (2014) Working memory  
756 contributions to reinforcement learning impairments in schizophrenia. *J*  
757 *Neurosci Off J Soc Neurosci* 34:13747–13756.
- 758 Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, Fehr  
759 E, Stephan KE (2014) Inferring on the Intentions of Others by Hierarchical  
760 Bayesian Learning. *PLoS Comput Biol* 10:e1003810.
- 761 Dudley R, Taylor P, Wickham S, Hutton P (2016) Psychosis, Delusions and the  
762 “Jumping to Conclusions” Reasoning Bias: A Systematic Review and Meta-  
763 analysis. *Schizophr Bull* 42:652–665.
- 764 Durstewitz D, Seamans JK (2008) The dual-state theory of prefrontal cortex  
765 dopamine function with relevance to catechol-o-methyltransferase  
766 genotypes and schizophrenia. *Biol Psychiatry* 64:739–749.
- 767 Fear CF, Healy D (1997) Probabilistic reasoning in obsessive-compulsive and  
768 delusional disorders. *Psychol Med* 27:199–208.
- 769 Foulds GA, Bedford A (1975) Hierarchy of classes of personal illness. *Psychol*  
770 *Med* 5:181–192.
- 771 Friston KJ (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol*  
772 *Sci* 360:815–836.
- 773 Galea JM, Ruge D, Buijink A, Bestmann S, Rothwell JC (2013) Punishment-  
774 induced behavioral and neurophysiological variability reveals dopamine-  
775 dependent selection of kinematic movement parameters. *J Neurosci Off J*  
776 *Soc Neurosci* 33:3981–3988.
- 777 Garety PA, Hemsley DR, Wessely S (1991) Reasoning in deluded schizophrenic  
778 and paranoid patients. Biases in performance on a probabilistic inference  
779 task. *J Nerv Ment Dis* 179:194–201.
- 780 Gepperth A, Lefort M (2016) Learning to be attractive: Probabilistic computation  
781 with dynamic attractor networks. In: 2016 Joint IEEE International  
782 Conference on Development and Learning and Epigenetic Robotics (ICDL-  
783 EpiRob), pp 270–277.
- 784 Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, Collins  
785 AGE, Frank MJ (2012) Negative symptoms and the failure to represent the

- 786 expected reward value of actions: behavioral and computational modeling  
787 evidence. *Arch Gen Psychiatry* 69:129–138.
- 788 Hamm JP, Peterka DS, Gogos JA, Yuste R (2017) Altered Cortical Ensembles in  
789 Mouse Models of Schizophrenia. *Neuron* 94:153-167.e8.
- 790 Hauser TU, Iannaccone R, Ball J, Mathys C, Brandeis D, Walitza S, Brem S (2014)  
791 Role of the medial prefrontal cortex in impaired decision making in  
792 juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry*  
793 71:1165–1173.
- 794 Hopfield JJ (1982) Neural networks and physical systems with emergent  
795 collective computational abilities. *Proc Natl Acad Sci U S A* 79:2554–2558.
- 796 Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HEM,  
797 Stephan KE (2013) Hierarchical Prediction Errors in Midbrain and Basal  
798 Forebrain during Sensory Learning. *Neuron* 80:519–530.
- 799 Jardri R, Duverne S, Litvinova AS, Denève S (2017) Experimental evidence for  
800 circular inference in schizophrenia. *Nat Commun* 8:14218.
- 801 Javitt DC, Zukin SR, Heresco-Levy U, Umbricht D (2012) Has an angel shown the  
802 way? Etiological and therapeutic implications of the PCP/NMDA model of  
803 schizophrenia. *Schizophr Bull* 38:958–966.
- 804 Kay SR, Fiszbein A, Opfer LA (1987) The Positive and Negative Syndrome Scale  
805 (PANSS) for Schizophrenia. *Schizophr Bull* 13:261–276.
- 806 Krawiecka M, Goldberg D, Vaughan M (1977) A standardized psychiatric  
807 assessment scale for rating chronic psychotic patients. *Acta Psychiatr*  
808 *Scand* 55:299–308.
- 809 Lam NH, Borduqui T, Hallak J, Roque AC, Anticevic A, Krystal JH, Wang X-J,  
810 Murray JD (2017) Effects of altered excitation-inhibition balance on  
811 decision making in a cortical circuit model. *bioRxiv*:100347.
- 812 Langdon R, Ward PB, Coltheart M (2010) Reasoning anomalies associated with  
813 delusions in schizophrenia. *Schizophr Bull* 36:321–330.
- 814 Marshall L, Mathys C, Ruge D, de Berker AO, Dayan P, Stephan KE, Bestmann S  
815 (2016) Pharmacological Fingerprints of Contextual Uncertainty. *PLoS Biol*  
816 14:e1002575.
- 817 Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for  
818 individual learning under uncertainty. *Front Hum Neurosci* 5:39.
- 819 Miller GA, Chapman JP (2001) Misunderstanding analysis of covariance. *J*  
820 *Abnorm Psychol* 110:40–48.
- 821 Moritz S, Woodward TS (2005) Jumping to conclusions in delusional and non-  
822 delusional schizophrenic patients. *Br J Clin Psychol* 44:193–207.

- 823 Moutoussis M, Bentall RP, El-Deredy W, Dayan P (2011) Bayesian modelling of  
824 Jumping-to-Conclusions bias in delusional patients. *Cognit*  
825 *Neuropsychiatry* 16:422–447.
- 826 Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, Wang X-J  
827 (2014) Linking microcircuit dysfunction to cognitive impairment: effects  
828 of disinhibition associated with schizophrenia in a cortical working  
829 memory model. *Cereb Cortex N Y N 1991* 24:859–872.
- 830 Nelson HE (1982) National Adult Reading Test (NART): For the Assessment of  
831 Premorbid Intelligence in Patients with Dementia : Test Manual. NFER-  
832 Nelson.
- 833 Pantelis C, Barber FZ, Barnes TR, Nelson HE, Owen AM, Robbins TW (1999)  
834 Comparison of set-shifting ability in patients with chronic schizophrenia  
835 and frontal lobe damage. *Schizophr Res* 37:251–270.
- 836 Peters E, Garety P (2006) Cognitive functioning in delusions: a longitudinal  
837 analysis. *Behav Res Ther* 44:481–514.
- 838 Peters ER, Joseph SA, Garety PA (1999) Measurement of delusional ideation in  
839 the normal population: introducing the PDI (Peters et al. Delusions  
840 Inventory). *Schizophr Bull* 25:553–576.
- 841 Powers AR, Mathys C, Corlett PR (2017) Pavlovian conditioning-induced  
842 hallucinations result from overweighting of perceptual priors. *Science*  
843 357:596–600.
- 844 Raine A (1991) The SPQ: a scale for the assessment of schizotypal personality  
845 based on DSM-III-R criteria. *Schizophr Bull* 17:555–564.
- 846 Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional  
847 interpretation of some extra-classical receptive-field effects. *Nat Neurosci*  
848 2:79–87.
- 849 Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection  
850 for group studies - revisited. *NeuroImage* 84:971–985.
- 851 Rolls ET, Loh M, Deco G, Winterer G (2008) Computational models of  
852 schizophrenia and dopamine modulation in the prefrontal cortex. *Nat Rev*  
853 *Neurosci* 9:696–709.
- 854 Schlagenhaut F, Huys QJM, Deserno L, Rapp MA, Beck A, Heinze H-J, Dolan R,  
855 Heinz A (2013) Striatal dysfunction during reversal learning in  
856 unmedicated schizophrenia patients. *NeuroImage*.
- 857 Speechley WJ, Whitman JC, Woodward TS (2010) The contribution of  
858 hypersalience to the “jumping to conclusions” bias associated with  
859 delusions in schizophrenia. *J Psychiatry Neurosci JPN* 35:7–17.

- 860 Standage D, You H, Wang D-H, Dorris MC (2013) Trading speed and accuracy by  
861 coding time: a coupled-circuit cortical model. *PLoS Comput Biol*  
862 9:e1003021.
- 863 Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model  
864 selection for group studies. *NeuroImage* 46:1004–1017.
- 865 Stuke H, Stuke H, Weilhhammer VA, Schmack K (2017) Psychotic Experiences  
866 and Overhasty Inferences Are Related to Maladaptive Learning. *PLoS*  
867 *Comput Biol* 13:e1005328.
- 868 Sutton R (1992) Gain Adaptation Beats Least Squares? Available at:  
869 <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.9218>  
870 [Accessed January 26, 2018].
- 871 Vinckier F, Gaillard R, Palminteri S, Rigoux L, Salvador A, Fornito A, Adapa R,  
872 Krebs MO, Pessiglione M, Fletcher PC (2016) Confidence and psychosis: a  
873 neuro-computational account of contingency learning disruption by  
874 NMDA blockade. *Mol Psychiatry* 21:946–955.
- 875 Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE (2014)  
876 Spatial attention, precision, and Bayesian inference: a study of saccadic  
877 response speed. *Cereb Cortex N Y N 1991* 24:1436–1450.
- 878 Waltz JA, Kasanova Z, Ross TJ, Salmeron BJ, McMahon RP, Gold JM, Stein EA  
879 (2013) The roles of reward, default, and executive control networks in  
880 set-shifting impairments in schizophrenia. *PloS One* 8:e57257.
- 881 Wang X-J (2013) The prefrontal cortex as a quintessential “cognitive-type”  
882 neural circuit: Working memory and decision making. Available at:  
883 [https://nyuscholars.nyu.edu/en/publications/the-prefrontal-cortex-as-a-](https://nyuscholars.nyu.edu/en/publications/the-prefrontal-cortex-as-a-quintessential-cognitive-type-neural-c)  
884 [quintessential-cognitive-type-neural-c](https://nyuscholars.nyu.edu/en/publications/the-prefrontal-cortex-as-a-quintessential-cognitive-type-neural-c) [Accessed January 31, 2018].
- 885 Wechsler D (1997) WAIS-III: Administration and scoring manual: Wechsler adult  
886 intelligence scale. Psychological Corporation.
- 887 Woodward TS, Moritz S, Cuttler C, Whitman JC (2006) The contribution of a  
888 cognitive bias against disconfirmatory evidence (BADE) to delusions in  
889 schizophrenia. *J Clin Exp Neuropsychol* 28:605–617.
- 890 Young HF, Bentall RP (1997) Probabilistic reasoning in deluded, depressed and  
891 normal subjects: effects of task difficulty and meaningful versus non-  
892 meaningful material. *Psychol Med* 27:455–465.
- 893
- 894
- 895

896 **Figure Legends**

897

898 **Figure 1: Effects of attractor network dynamics on belief updating**

899

900 This schematic illustrates the energy landscapes of two Hopfield-type networks  
901 each with two basins of attraction. The continuous black line depicts a normal  
902 network whose basins of attraction are relatively deep. The dotted black line  
903 depicts the effect of NMDAR (or cortical dopamine 1 receptor (Durstewitz and  
904 Seamans, 2008)) hypofunction (Abi-Saab et al., 1998; Javitt et al., 2012) on the  
905 energy landscape: the attractor basins become more shallow. We assume that  
906 Basins A and B correspond to different inferences about (hidden) states in the  
907 world, e.g. one jar or another being the source of beads in the beads task. The  
908 dots correspond to the networks' representations of either control or Scz  
909 subjects' beliefs about these hidden states. Such networks are highly reminiscent  
910 of Hopfield networks with two stored representations – in this case, the  
911 representations correspond to inferences about hidden states, rather than  
912 memories. The arrows depict the changes in network states resulting from  
913 sensory evidence for (solid arrows) or against (dashed arrows) the current  
914 inference. When the attractor basin is shallower, it is harder for supportive  
915 evidence to stabilise the current state much further, but it is easier for  
916 contradictory evidence – or just stochastic neuronal firing – to shift the current  
917 network state towards an alternative state. These changes in network dynamics  
918 may also be reflected in the inferences the network computes – i.e. easier  
919 switching between attractor basins may correspond to easier switching between



920 beliefs – although this is yet to be demonstrated experimentally. NMDAR  
921 hypofunction could contribute to an increased tendency to switch between  
922 beliefs and increased stochasticity in responding in several ways (Rolls et al.,  
923 2008): i) by reducing inhibitory interneuron activity, via weakened NMDAR  
924 synapses from pyramidal cells to interneurons, such that other attractor states  
925 are less suppressed when one is active (a spiking network model has shown that  
926 this leads to more rapid initial belief updating in perceptual tasks (Lam et al.,  
927 2017)), ii) by reducing pyramidal cell activity, via weakened recurrent NMDAR  
928 synapses on pyramidal cells, such that attractor states are harder to sustain, and  
929 iii) by reducing the NMDAR time constant, making states more vulnerable to  
930 random fluctuations in neural activity. See also similar schematics elsewhere  
931 (Durstewitz and Seamans, 2008; Rolls et al., 2008).

932

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

935

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

945

946 **Figure 3: The structure of the Hierarchical Gaussian Filter (Model 6) and**947 **some simulated data**

948

949 In the upper left panel, the evolution of  $\mu_2$ , the posterior estimate of tendency  $x_2$ 

950 towards the blue (positive) or red (negative) jar, is plotted over two

951 concatenated series of 10 trials (the first two in dataset 2). The estimate of the

952 tendency on trial  $k+1$ ,  $\mu_2^{(k+1)}$ , is selected from a Gaussian distribution with mean953  $\mu_2^{(k)}$  (blue line) and variance  $\sigma_2^{(k)} + \exp(\omega)$  (blue shading).  $\omega$  is a static source of954 variance at this level. The initial variance  $\sigma_2^{(0)}$  (along with  $\omega$ ) affects the size of

955 initial updates, so we estimated this parameter (which is often fixed). The beads

956 seen by the subjects,  $u^{(k)}$  (blue and red dots) and the response model are957 illustrated in the bottom left panel. The response model maps from  $\hat{\mu}_1^{(k+1)}$ 958 (purple line) – the prediction of  $x_1$  on the next trial, which is a sigmoid function  $s$ 959 of  $\mu_2^{(k)}$  (or of  $(\kappa_1 \mu_2^{(k)})$  in Models 5 and 6) – to  $y^{(k)}$ , the subject's indicated

960 estimate of the probability the jar is blue (green dots). Variation in this mapping

961 is modelled as the precision  $\nu$  of a beta distribution.

962 The right panel is a schematic representation of the generative model in Models

963 5 and 6 (i.e. including  $\kappa_1$ ). The black arrows denote the probabilistic network on964 trial  $k$ ; the grey arrows denote the network at other points in time. The

965 perceptual model lies above the dotted arrows, and the response model below

966 them. The shaded circles are known quantities, and the parameters and states in

967 unshaded circles are estimated. The dotted line represents the result of an



968 inferential process (the response model builds on a perceptual model inference);  
 969 the solid lines are generative processes.

970

971 **Figure 4: Simulated data illustrating the effects of  $\varphi$  (Models 3 and 4)**  
 972 **and  $\kappa_1$  (Model 5 and 6) on inference**

973

974 This figure illustrates the effects of  $\varphi$  (used in Models 3 and 4) and  $\kappa_1$  (used in  
 975 Models 5 and 6) on inference. Both panels show simulated perceptual model  
 976 predictions in the same format as before, with  $\sigma_2^{(0)}$  and  $\omega$  set to their previous  
 977 values – hence the purple line in these plots is identical to that in Figure 3. The  
 978 second level and simulated responses  $y$  have been omitted for clarity.

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

988 Lower left panel: Simulations of a perceptual model using four different values of  
 989 scaling factor  $\kappa_1$ , which alters the sigmoid transformation:  $\hat{\mu}_1^{(k+1)} = s(\kappa_1 \cdot \mu_2^{(k)})$ .  
 990 When  $\kappa_1 > \exp(0)$  updating is greater to unexpected evidence and lower to  
 991 consistent evidence; when  $\kappa_1 < \exp(0)$  the reverse is true. The red and brown

992 lines ( $\kappa_l > \exp(0)$ ) illustrate the effects of increasingly unstable attractor  
993 networks, i.e. switching between states (jars) becomes more likely (a  
994 concomitant increase in vulnerability to noise, i.e. response stochasticity, is not  
995 shown). The green line ( $\kappa_l = \exp(-1)$ ) illustrates slower updating around  $\hat{\mu}_1 = 0.5$ ,  
996 as was found in controls.  $\kappa_l$  permits a greater range of updating patterns than  $\varphi$   
997 (the green and brown trajectories in the lower panel cannot be produced by  
998 Model 4) which may be why Model 6 can fit both controls and Scz groups well.  
999 Middle panel: This plot shows the effects of  $\kappa_l$  on belief updating, as a function of  
1000 the initial belief  $\hat{\mu}_1$  ( $\sigma_2^{(0)}$  and  $\omega$  were set to 1.5 and -1 respectively, as in Figure 5;  
1001 changing these parameters does not qualitatively alter the effects of  $\kappa_l$  shown  
1002 here). For values of  $\kappa_l < \exp(0)=1$  (bottom three curves) and initial beliefs to the  
1003 left of these curves' maxima (i.e. that the jar is probably red), relatively small  
1004 increases in  $\hat{\mu}_1$  are made if one blue bead ( $u = 1$ ) is observed, such that the  
1005 subject still believes the jar is most likely red. For values of  $\kappa_l > \exp(0.5)$  (top two  
1006 curves), observing one blue bead causes such a large update for all but the most  
1007 certain initial beliefs in a red jar that the subject's posterior belief is that the jar  
1008 is probably blue. These subjects' beliefs are no longer stable, but neither can they  
1009 reach certainty: only tiny updates towards 1 are possible for  $\hat{\mu}_1 > 0.8$ .  
1010 Right panel: This plot illustrates the average absolute shifts in beliefs on  
1011 observing beads of either colour. This 'vulnerability to updating' is highly  
1012 reminiscent of the 'energy state' of a neural network model (schematically  
1013 illustrated in Figure 1) - i.e. in low energy states, less updating is expected. The  
1014 effect of increasing  $\kappa_l$  is to convert confident beliefs about the jar (near 0 and 1)  
1015 from low to high 'energy states', i.e. to make them much more unstable.  
1016

1017 **Figure 5: Recovery of model parameters from simulated data**

1018

1019 200 datasets were simulated using Model 6; 100 using modal parameter values  
1020 for the control group (dataset 2) and 100 using modal values for the Scz group  
1021 (also dataset 2) – the values are indicated using red lines. Both used settings of  
1022  $\sigma_2^{(0)} = 1.5$ ,  $\omega = -1$ . The control group used  $\kappa_1 = 0.37$  (i.e.  $\exp(-1)$ ) and  $\nu = \exp(3)$ .  
1023 The Scz group used  $\kappa_1 = 2.7$  (i.e.  $\exp(1)$ ) and  $\nu = \exp(2)$ . Histograms depicting the  
1024 parameter estimates from model inversion using the same priors as were  
1025 employed in the main analysis are shown above: the modal control and Scz  
1026 simulation results are in the upper and lower rows respectively.

1027

1028 **Figure 6: Bayesian model selection results for both datasets.**

1029

1030 The left panel depicts the protected exceedance probabilities for the six models in  
1031 each group in each dataset. The protected exceedance probability is the  
1032 probability a particular model is more likely than any other tested model, above  
1033 and beyond chance, given the group data (Rigoux et al., 2014). Model 6 wins in all  
1034 groups in both datasets (upper row: controls, middle row: Scz, bottom row:  
1035 clinical controls).

1036 The right panel depicts the model likelihoods for the six models in each group in  
1037 each dataset. The model likelihood is the probability of that model being the best  
1038 for any randomly selected subject (Stephan et al., 2009). Model 4 is a clear runner-  
1039 up in the psychotic (Scz) and clinical control groups at baseline in dataset 1, and  
1040 in the Scz group in dataset 2.

1041

1042 **Figure 7: Probability density plots for Model 6 parameters in dataset 1.**

1043

1044 The distributions of parameter values for  $\sigma_2^{(0)}$ ,  $\omega$ ,  $\log(\nu)$  and  $\log(\kappa_1)$  are plotted  
1045 for dataset 1 at baseline (upper row) and dataset 1 at follow-up (lower row). The  
1046 symbols denote significant group differences: § between non-clinical controls  
1047 and clinical controls, \* between non-clinical controls and Scz, † between Scz and  
1048 clinical controls. Please see the text for the details of all statistical comparisons.

1049

1050 **Figure 8: Model 6 parameters in dataset 2 – distributions and correlation**

1051

1052 Upper panel: The distributions of parameter values for  $\sigma_2^{(0)}$ ,  $\omega$ ,  $\log(\nu)$  and  $\log(\kappa_1)$   
1053 are plotted for dataset 2. The \* symbol denotes significant group differences  
1054 between the Scz group and non-clinical control subgroup (well-matched in age  
1055 and sex); the group difference in  $\sigma_2^{(0)}$  is not indicated because it was non-  
1056 significant ( $p=0.056$ ) in the well-matched comparison. Please see the text for the  
1057 details of all statistical comparisons.

1058 Lower panel: The significant correlation between  $\log(\nu)$  and  $\log(\kappa_1)$  in dataset 2  
1059 is plotted, with controls' parameters in black and Scz in red. Similar correlations  
1060 were also found in dataset 1 at both time points (see text).

1061

1062 **Figure 9: Responses and model fits for two control subjects**

1063

1064 These plots show two control subjects' responses to four ten-bead sequences  
1065 concatenated together, in the same format as Figure 3 (but without the second  
1066 level, due to space constraints); in the latter two sequences blue and red were

1067 swapped around for model-fitting purposes. Each plot shows  $u^{(k)}$  – the beads  
1068 seen by the subjects on trials  $k = 1, \dots, 10$  (blue and red dots),  $y$  – the subject’s  
1069 (Likert scale) response about the probability the jar is blue (green dots), and  
1070  $\hat{\mu}_1^{(k+1)}$  – the model’s estimate of the subject’s prediction the jar is blue (purple  
1071 line). The parameter estimates for each subject are shown above their graphs.  
1072 These subjects have fairly similar initial variance  $\sigma_2^{(0)}$ , (inverse) response  
1073 stochasticity  $\nu$ , and instability factor  $\kappa_1$ . Subject 18 in the upper panel has a much  
1074 lower overall evolution rate  $\omega$  than Subject 67 in the lower panel, therefore  
1075 Subject 18 never reaches certainty about either jar, and makes relatively small  
1076 changes to her beliefs in response to beads of varying colours. Both subjects have  
1077 a low  $\kappa_1$ , and so they make relatively small adjustments to their beliefs following  
1078 unexpected evidence (this behaviour can best be captured by the models  
1079 containing  $\kappa_1$  – see Figure 4). Subject 18’s responses are very close to those  
1080 predicted by the model, and this is reflected in her relatively high value of  $\nu$ .

1081

1082 **Figure 10: Responses and model fits for two Scz subjects**

1083

1084 These plots show two Scz subjects’ responses to four ten-bead sequences in the  
1085 same format as Figure 9. These subjects have similar evolution rate  $\omega$  to the  
1086 control subjects in Figure 9, but they both have a much higher  $\kappa_1$ , meaning that  
1087 they make much greater changes to their beliefs when presented with  
1088 unexpected evidence, but do not reach certainty when faced with consistent  
1089 evidence. Subject 122 (lower panel) has a slightly higher evolution rate  $\omega$  than  
1090 Subject 145 (upper panel), and so his switching between jars is even more  
1091 pronounced. These subjects also have slightly lower (inverse) response

1092 stochasticity  $v$  than the control subjects in Figure 9, and so their responses tend

1093 to be further from the model predictions.

1094

1095

1096

| 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 <sup>a</sup>          | 27.77<br>(6.74)          | 27.9<br>(6.37)           | 40.91<br>(13.57)     | 40.1<br>(13)         | 31.22<br>(7.28) | 29.9 (7.83)    | Age                               | 32.8<br>(11.5) | 45.3<br>(8.8) | 39.5<br>(11.4)    |
| Gender                    | 18 M,<br>17 F            | 12 M,<br>8 F             | 11 M,<br>11 F        | 8 M,<br>10 F         | 21 M,<br>2 F    | 17 M,<br>0 F   | Gender                            | 45 M,<br>66 F  | 38 M,<br>18 F | 40 M,<br>20 F     |
| <b>Cognitive measures</b> |                          |                          |                      |                      |                 |                |                                   |                |               |                   |
| IQ <sup>b</sup>           | 107.5<br>(11.6)          | 108.6<br>(10.3)          | 97.4<br>(13.8)       | 99.8<br>(10.2)       | 88.1<br>(12.7)  | 87.8<br>(14.2) | NART <sup>a</sup>                 | 112<br>(6.9)   | 109<br>(8.2)  | 112<br>(7.5)      |
|                           |                          |                          |                      |                      |                 |                | Working memory (LNS) <sup>b</sup> | 16.2<br>(2.8)  | 10.3<br>(4.2) | 16.4 (2.7)        |
| <b>Delusion proneness</b> |                          |                          |                      |                      |                 |                | <b>Schizotypy</b>                 |                |               |                   |
| PDI (total) <sup>c</sup>  | 54.6<br>(43.1)           | 43.6<br>(42.5)           | 87.1<br>(55.2)       | 64.3<br>(57.3)       | 138.1<br>(74.2) | 96.7<br>(42.6) | SPQ, cognitive                    | 2.8<br>(1.9)   | 4.0<br>(2.6)  | 3.1(2)            |
| DSSI <sup>d</sup>         | 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<br>(2.6)  | 3.2 (2.2)         |

|  |                |                |  |  |   |   |   |                |                |                |
|--|----------------|----------------|--|--|---|---|---|----------------|----------------|----------------|
|  |                |                |  |  |   |   | SPQ, disorg                                     | 2.1 (1.7)      | 2.7 (1.9)      | 1.9 (1.8)      |
|  |                |                |  |  |   |   | SPQ, total <sup>c</sup>                         | 8.2 (1.3)      | 12 (5.3)       | 8.2 (4.4)      |
| <b>Diagnosis/<br/>Symptoms</b>             |                |                |  |  |   |   |   |                |                |                |
| Diagnoses                                  | -              | -              | 16<br>Depression,<br>3 anxiety &<br>depression,<br>3 SAD | 12<br>Depression,<br>3 anxiety &<br>depression,<br>3 SAD | 18 Scz,<br>5 bipolar/<br>schizo-<br>affective | 13 Scz,<br>4 bipolar/<br>schizo-<br>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 <sup>e</sup>                      | -              | -              | 5.5 (2.6)  | 2.8 (3.39)   | 9.1 (3.76)                                    | 3.7 (3.9)                                     | PANSS, total                                    | -              | 64.4 (17.3)    | -              |
| <b>Beads task</b>                          |                |                |  |  |   |   |   |                |                |                |
| Initial certainty<br>(1 bead) <sup>f</sup> | 0.58<br>(0.15) | 0.59<br>(0.12) | 0.68<br>(0.19)   | 0.63<br>(0.16)   | 0.76<br>(0.17)                                | 0.68<br>(0.29)                                | Initial certainty<br>(all, 1 bead) <sup>d</sup> | 0.67<br>(0.13) | 0.71<br>(0.14) | 0.68<br>(0.14) |



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

1097

1098

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

1099

1100 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)  
 1101 and delusion proneness using the Peters Delusion Inventory, PDI (Peters et al., 1999) and Delusions-Symptoms-States Inventory, DSSI (Foulds and Bedford,  
 1102 1975). Symptoms were assessed using the Manchester Scale, MS (Krawiecka et al., 1977). In the tests below, 'Scz' refers to the whole Psychotic group.

1103 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  
 1104 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-  
 1105 3 beads') - and using the response to the first bead ('1 bead').

1106

1107

1108 <sup>a</sup> At t1: One-way ANOVA  $F(2,77) = 13.9, p = 10^{-5}$ . Tukey's HSD: Scz vs Non-clinical controls diff  
1109 = 3.45,  $p(\text{adj}) = 0.35$ ; Clinical vs Non-clinical controls diff = 13.1,  $p(\text{adj}) = 10^{-5}$ ; Clinical controls  
1110 vs Scz diff = 9.69,  $p(\text{adj}) = 0.002$   
1111 At t2: One-way ANOVA  $F(2,52) = 8.85, p = 0.0005$ . Tukey's HSD: Scz vs Non-clinical controls  
1112 diff = 1.98,  $p(\text{adj}) = 0.8$ ; Clinical vs Non-clinical controls diff = 12.2,  $p(\text{adj}) = 0.0006$ ; Clinical  
1113 controls vs Scz diff = 10.2,  $p(\text{adj}) = 0.007$   
1114 <sup>b</sup> At t1: One-way ANOVA  $F(2,75) = 16.2, p = 10^{-6}$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1115 = -19.5,  $p(\text{adj}) = 10^{-6}$ ; Clinical vs Non-clinical controls diff = -10.1,  $p(\text{adj}) = 0.011$ ; Clinical  
1116 controls vs Scz diff = 9.36,  $p(\text{adj}) = 0.043$   
1117 At t2: One-way ANOVA  $F(2,51) = 14.5, p = 10^{-5}$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1118 = -20.8,  $p(\text{adj}) = 10^{-5}$ ; Clinical vs Non-clinical controls diff = -8.8,  $p(\text{adj}) = 0.057$ ; Clinical  
1119 controls vs Scz diff = 12,  $p(\text{adj}) = 0.01$   
1120 <sup>c</sup> At t1: One-way ANOVA  $F(2,68) = 12.6, p = 0.00002$ ; Tukey's HSD: Scz vs Non-clinical controls  
1121 diff = 83.5,  $p(\text{adj}) = 10^{-5}$ ; Clinical vs Non-clinical controls diff = -32.5,  $p(\text{adj}) = 0.094$ ; Clinical  
1122 controls vs Scz diff = -51,  $p(\text{adj}) = 0.016$   
1123 At t2: One-way ANOVA  $F(2,52) = 4, p = 0.024$ ; Tukey's HSD: Scz vs Non-clinical controls diff =  
1124 53.1,  $p(\text{adj}) = 0.018$ ; Clinical vs Non-clinical controls diff = -20.7,  $p(\text{adj}) = 0.5$ ; Clinical controls  
1125 vs Scz diff = -32.4,  $p(\text{adj}) = 0.22$   
1126 <sup>d</sup> At t1: One-way ANOVA  $F(2,76) = 43, p = 10^{-13}$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1127 = 12.9,  $p(\text{adj}) = 10^{-10}$ ; Clinical vs Non-clinical controls diff = 2.52,  $p(\text{adj}) = 0.19$ ; Clinical controls  
1128 vs Scz diff = -10.4,  $p(\text{adj}) = 10^{-8}$   
1129 At t2: One-way ANOVA  $F(2,51) = 3.7, p = 0.032$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1130 = 5.2,  $p(\text{adj}) = 0.026$ ; Clinical vs Non-clinical controls diff = 1.65,  $p(\text{adj}) = 0.66$ ; Clinical controls  
1131 vs Scz diff = -3.56,  $p(\text{adj}) = 0.18$   
1132 <sup>e</sup> At t1: Welch's  $t(38.4) = -3.62, p = 0.00086$ , Cohen's  $d = 1.1$   
1133 At t2: Welch's  $t(17.8) = -2.55, p = 0.02$ , Cohen's  $d = 1.0$   
1134 <sup>f</sup> At t1: One-way ANOVA  $F(2,77) = 8.7, p = 0.0004$ ; Tukey's HSD: Scz vs Non-clinical controls  
1135 diff = 0.18,  $p(\text{adj}) = 0.0003$ ; Clinical vs Non-clinical controls diff = 0.11,  $p = 0.06$ ; Clinical  
1136 controls vs Scz diff = -0.08,  $p(\text{adj}) = 0.25$   
1137 At t2: One-way ANOVA  $F(2,52) = 0.9, p = 0.4$   
1138 <sup>g</sup> At t1: One-way ANOVA  $F(2,77) = 6.2, p = 0.003$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1139 = -0.14,  $p(\text{adj}) = 0.002$ ; Clinical vs Non-clinical controls diff = 0.04,  $p = 0.57$ ; Clinical controls  
1140 vs Scz diff = -0.096,  $p(\text{adj}) = 0.074$   
1141 At t2: One-way ANOVA  $F(2,52) = 2.35, p = 0.11$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1142 = 0.07,  $p(\text{adj}) = 0.28$ ; Clinical vs Non-clinical controls diff = -0.03,  $p = 0.8$ ; Clinical controls vs  
1143 Scz diff = -0.1,  $p(\text{adj}) = 0.1$

1144 <sup>h</sup> At t1: One-way ANOVA  $F(2,77) = 6, p = 0.004$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1145 =  $-0.23, p(adj) = 0.003$ ; Clinical vs Non-clinical controls diff =  $-0.14, p = 0.13$ ; Clinical controls  
1146 vs Scz diff =  $0.097, p(adj) = 0.41$   
1147 At t2: One-way ANOVA  $F(2,52) = 2.9, p = 0.062$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1148 =  $-0.18, p(adj) = 0.049$ ; Clinical vs Non-clinical controls diff =  $-0.08, p = 0.51$ ; Clinical controls  
1149 vs Scz diff =  $0.098, p(adj) = 0.4$   
1150 <sup>i</sup> At t1: One-way ANOVA  $F(2,77) = 0.71, p = 0.5$   
1151 At t2: One-way ANOVA  $F(2,52) = 2.79, p = 0.07$ ; Tukey's HSD: Scz vs Non-clinical controls diff  
1152 =  $-0.082, p(adj) = 0.41$ ; Clinical vs Non-clinical controls diff =  $-0.15, p = 0.057$ ; Clinical controls  
1153 vs Scz diff =  $-0.066, p(adj) = 0.57$   
1154 As reported previously, there were consistent negative correlations between initial certainty  
1155 (2-3 beads) and disconfirmatory updating in the clinical controls (baseline:  $\rho = -0.68, p =$   
1156  $0.0005$ ; follow-up:  $\rho = -0.75, p = 0.0003$ ) and the non-clinical controls (baseline:  $\rho = -0.52, p =$   
1157  $0.001$ ; follow-up:  $\rho = -0.43, p = 0.06$ ), but not in the psychotic group (baseline:  $\rho = -0.30, p =$   
1158  $0.17$ ; follow-up:  $\rho = 0.17, p = 0.5$ ). There was no consistent correlation between final certainty  
1159 and either of the other two measures at either time point ( $p \geq 0.1$  in 11 out of 12 comparisons).  
1160 In dataset 2, IQ was estimated using the National Adult Reading Test, NART (Nelson, 1982)  
1161 and working memory using the Letter Number Sequencing task, LNS, from the Wechsler Adult  
1162 Intelligence Scale-III (Wechsler, 1997). Schizotypy was assessed using the Schizotypal  
1163 Personality Questionnaire, SPQ (Raine, 1991), and symptoms using the Positive and Negative  
1164 Syndrome Scale, PANSS (Kay et al., 1987).  
1165 As can be seen in Figure 2 (main text), the Scz group showed greater initial certainty (1 bead)  
1166 in sequences A and B (Welch's  $t(94) = 2.8, p = 0.007$ , Cohen's  $d = 0.47$ ; Welch's  $t(97) = 3, p =$   
1167  $0.004$ , Cohen's  $d = 0.5$ , respectively) but not C and D (Welch's  $t(87) = 0.5, p = 0.6$ , Cohen's  $d =$   
1168  $0.09$ ; Welch's  $t(90) = -0.34, p = 0.73$ , Cohen's  $d = 0.06$ , respectively).  
1169 <sup>a</sup> Controls (all): Welch's  $t(95.1) = 2.27, p = 0.026$ , Cohen's  $d = 0.38$ ; Controls (subset): Welch's  
1170  $t(111) = 1.95, p = 0.053$ , Cohen's  $d = 0.36$   
1171 <sup>b</sup> Controls (all): Welch's  $t(81) = 9.57, p = 10^{-14}$ , Cohen's  $d = 1.66$ ; Controls (subset): Welch's  
1172  $t(93.6) = 9.25, p = 10^{-15}$ , Cohen's  $d = 1.73$   
1173 <sup>c</sup> Controls (all): Welch's  $t(92.4) = -4.64, p = 10^{-5}$ , Cohen's  $d = 0.78$ ; Controls (subset): Welch's  
1174  $t(107) = -4.19, p = 10^{-5}$ , Cohen's  $d = 0.78$   
1175 <sup>d</sup> Controls (all): Welch's  $t(110) = -1.9, p = 0.059$ , Cohen's  $d = 0.32$ ; Controls (subset): Welch's  
1176  $t(110) = -1.1, p = 0.28$ , Cohen's  $d = 0.2$   
1177 <sup>e</sup> Controls (all): Welch's  $t(109.1) = -0.76, p = 0.45$ , Cohen's  $d = 0.12$ ; Controls (subset): Welch's  
1178  $t(113.9) = -0.19, p = 0.85$ , Cohen's  $d = 0.03$   
1179 <sup>f</sup> Controls (all): Welch's  $t(88.2) = 2.09, p = 0.04$ , Cohen's  $d = 0.36$ ; Controls (subset): Welch's  
1180  $t(110.4) = -0.94, p = 0.35$ , Cohen's  $d = 0.18$

1181 <sup>g</sup> Controls (all): Welch's  $t(80.1) = 2.99, p = 0.0038$ , Cohen's  $d = 0.56$ ; Controls (subset): Welch's  
 1182  $t(98.7) = 2.18, p = 0.032$ , Cohen's  $d = 0.41$   
 1183 <sup>h</sup> Controls (all): Welch's  $t(85.5) = -3.41, p = 0.001$ , Cohen's  $d = 0.62$ ; Controls (subset): Welch's  
 1184  $t(106) = -2.21, p = 0.029$ , Cohen's  $d = 0.42$   
 1185  
 1186  
 1187  
 1188

| Model | Perceptual model parameters (prior mean in native space, prior variance in estimation space) |                                    |                      |                    | Response model parameter |
|-------|--|------------------------------------|----------------------|--------------------|--------------------------|
|       | Evolution rate   | Initial variance of belief re jars | Disconfirmatory bias | Belief instability | Response stochasticity   |
| 1     | $\omega (-2, 16)$  |                                    |                      |                    | $\nu (\exp(4.85), 1)$    |
| 2     | $\omega (-2, 16)$  | $\sigma_2^{(0)} (0.8, 0.5)$        |                      |                    | $\nu (\exp(4.85), 1)$    |
| 3     | $\omega (-2, 16)$  |                                    | $\varphi (0.1, 2)$   |                    | $\nu (\exp(4.85), 1)$    |
| 4     | $\omega (-2, 16)$  | $\sigma_2^{(0)} (0.8, 0.5)$        | $\varphi (0.1, 2)$   |                    | $\nu (\exp(4.85), 1)$    |
| 5     | $\omega (-2, 16)$  |                                    |                      | $\kappa_1 (1,1)$   | $\nu (\exp(4.85), 1)$    |
| 6     | $\omega (-2, 16)$  | $\sigma_2^{(0)} (0.8, 0.5)$        |                      | $\kappa_1 (1,1)$   | $\nu (\exp(4.85), 1)$    |

1189  
 1190  
 1191  
 1192  
 1193

**Table 2: Models, parameters and their prior distributions.**

|                                   | $\sigma_2^{(0)}$                   | $\omega$                         | $\log(\nu)$                         | $\log(\kappa_1)$                |
|-----------------------------------|------------------------------------|----------------------------------|-------------------------------------|---------------------------------|
| <b>Dataset 1 (baseline, n=80)</b> |                                    |                                  |                                     |                                 |
| 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,<br>$p=0.31$<br>$\eta^2=0.02$ | 0.22,<br>$p=0.9$<br>$\eta^2=0.0$ | 11.9,<br>$p=0.003$<br>$\eta^2=0.15$ | 9.6, $p=0.008$<br>$\eta^2=0.12$ |
| 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     | $p(adj)=0.2$                         | $p(adj)=0.5$                        | $p(adj)=0.3$                         | $p(adj)=0.4$                                     |
| <b>Dataset 1 (follow-up, n=55)</b> |                                      |                                     |                                      |  |
| 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, $p=0.3$<br>$\eta^2=0.04$       | 2.32,<br>$p=0.3$<br>$\eta^2=0.04$   | 8.5, $p=0.01$<br>$\eta^2=0.16$       | 8.0, $p=0.02$<br>$\eta^2=0.15$                   |
| 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     | $p(adj)=0.3$                         | $p(adj)=0.3$                        | $p(adj)=0.01$                        | $p(adj)=0.1$                                     |
| <b>Dataset 2 (n=167)</b>           |                                      |                                     |                                      |  |
| 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                | $Z=3.1$ ,<br>$p=0.002$ ,<br>$r=0.24$ | $Z=-0.6$ ,<br>$p=0.6$ ,<br>$r=0.04$ | $Z=3.9$ ,<br>$p=0.0001$ ,<br>$r=0.3$ | $Z=-5.6$ ,<br>$p=3 \times 10^{-8}$ ,<br>$r=0.43$ |
| <b>Dataset 2</b>                   |                                      |                                     |                                      |  |

|   |                              |                             |                               |                                 |
|---|------------------------------|-----------------------------|-------------------------------|---------------------------------|
| <b>(better-matched controls, n=116)</b> |                              |                             |                               |                                 |
| 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,<br>p=0.056,<br>r=0.18 | Z=0.12,<br>p=0.9,<br>r=0.01 | Z=3.4,<br>p=0.0007,<br>r=0.31 | Z=-4.1,<br>p=0.00004,<br>r=0.38 |

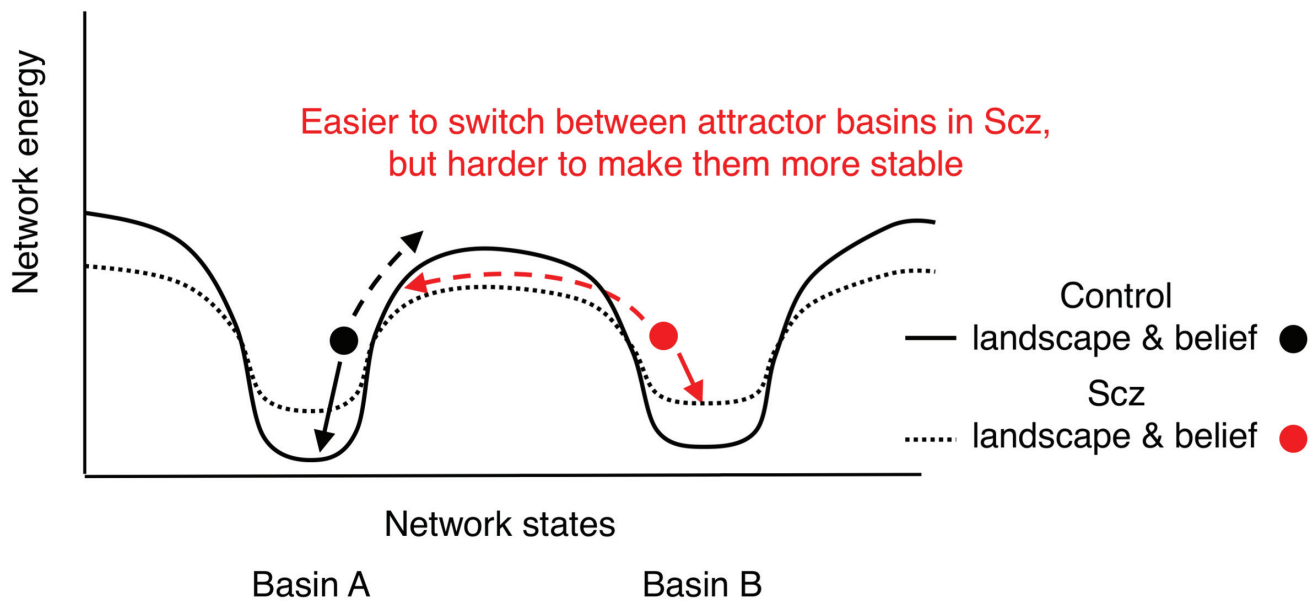
1194

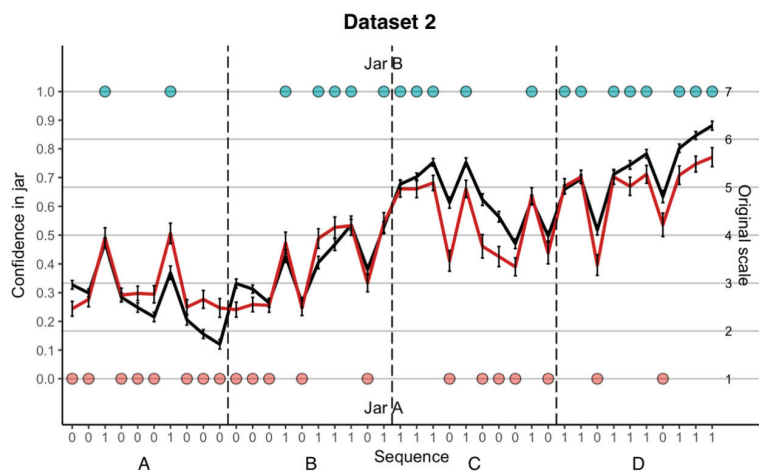
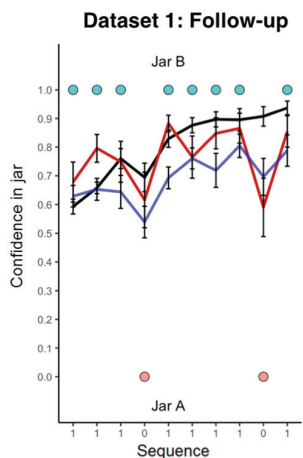
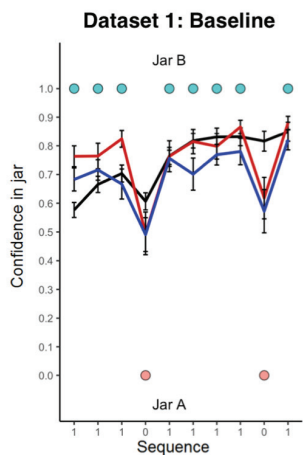
1195

1196

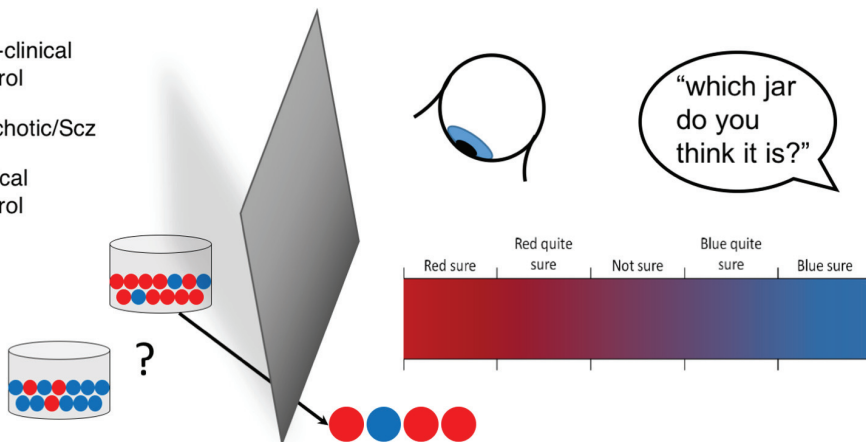
1197

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

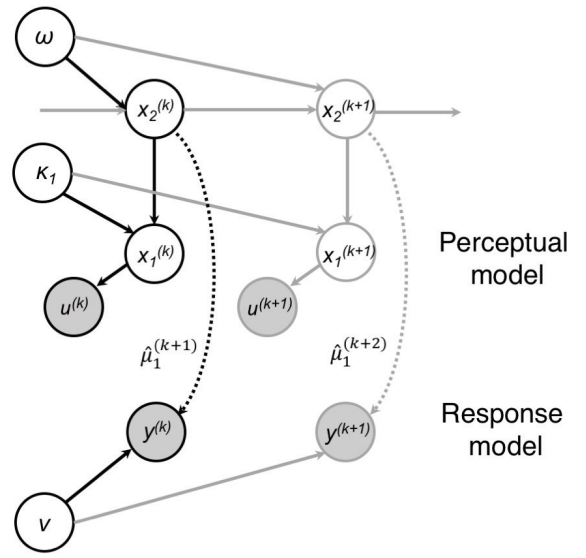
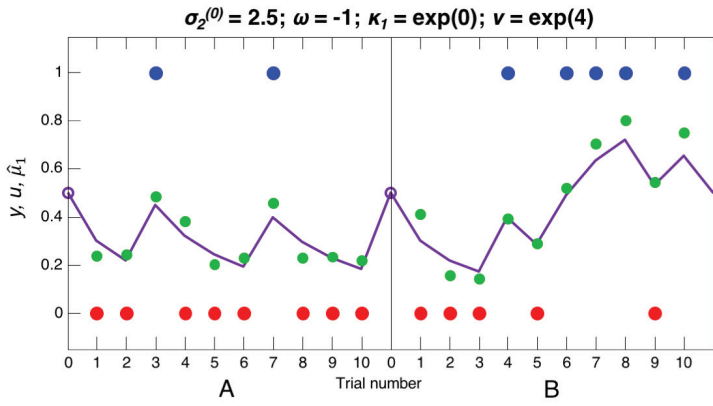
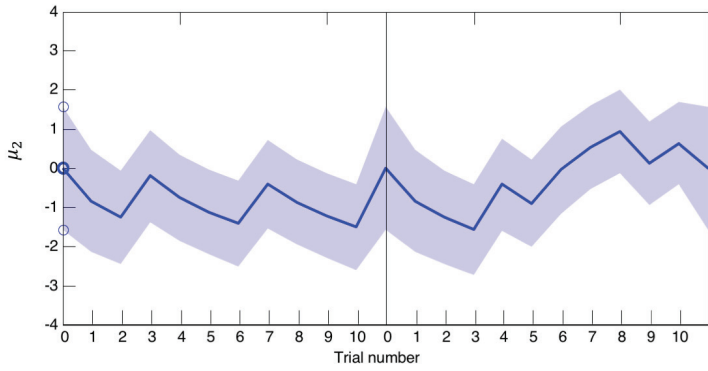


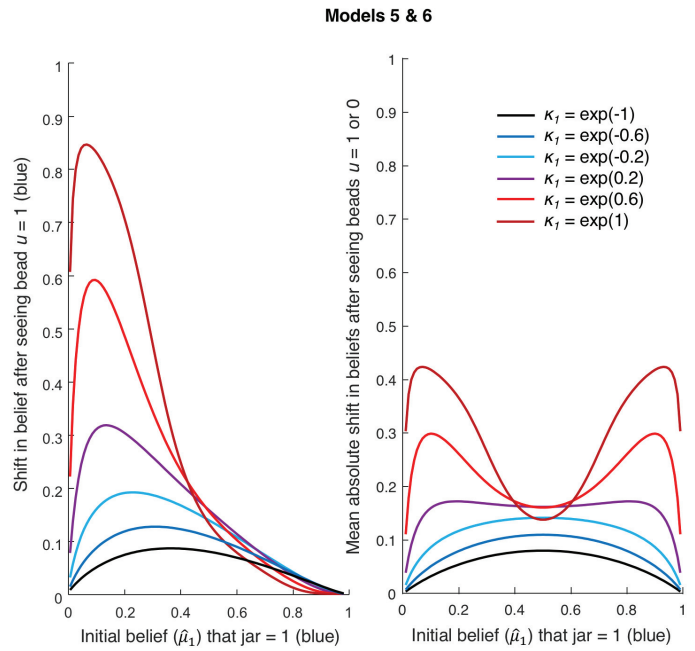
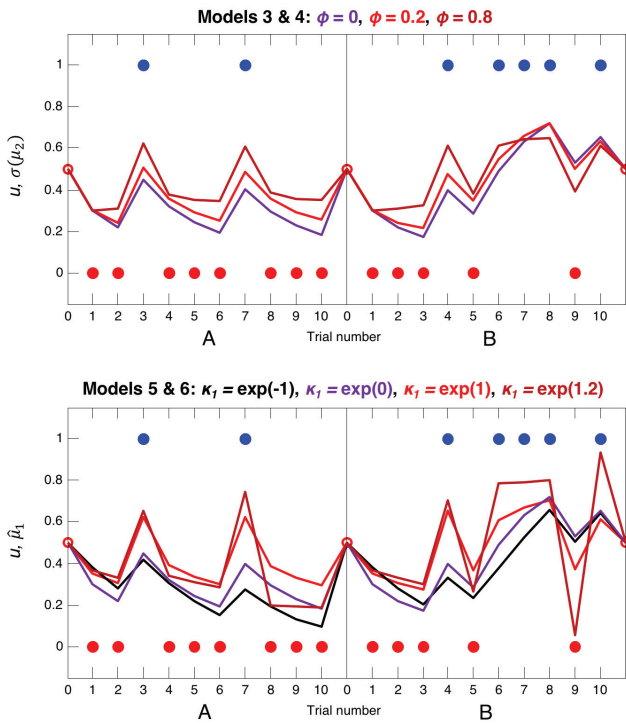


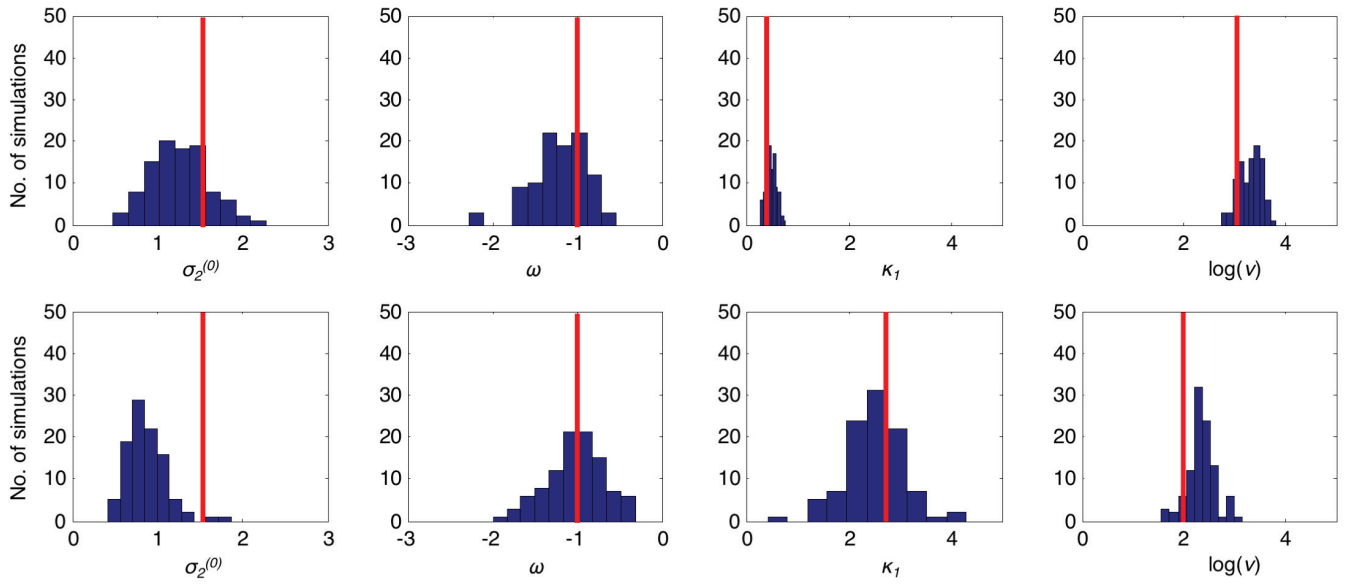
- Non-clinical control
- Psychotic/Scz
- Clinical control

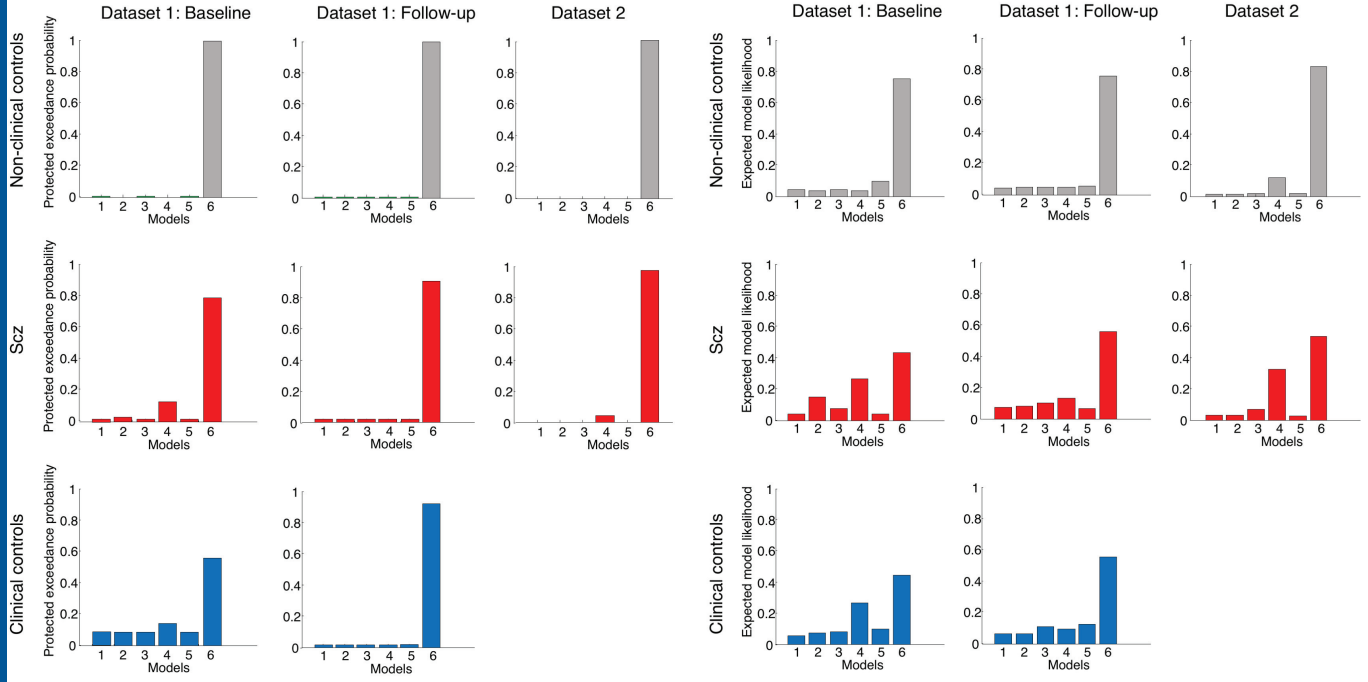


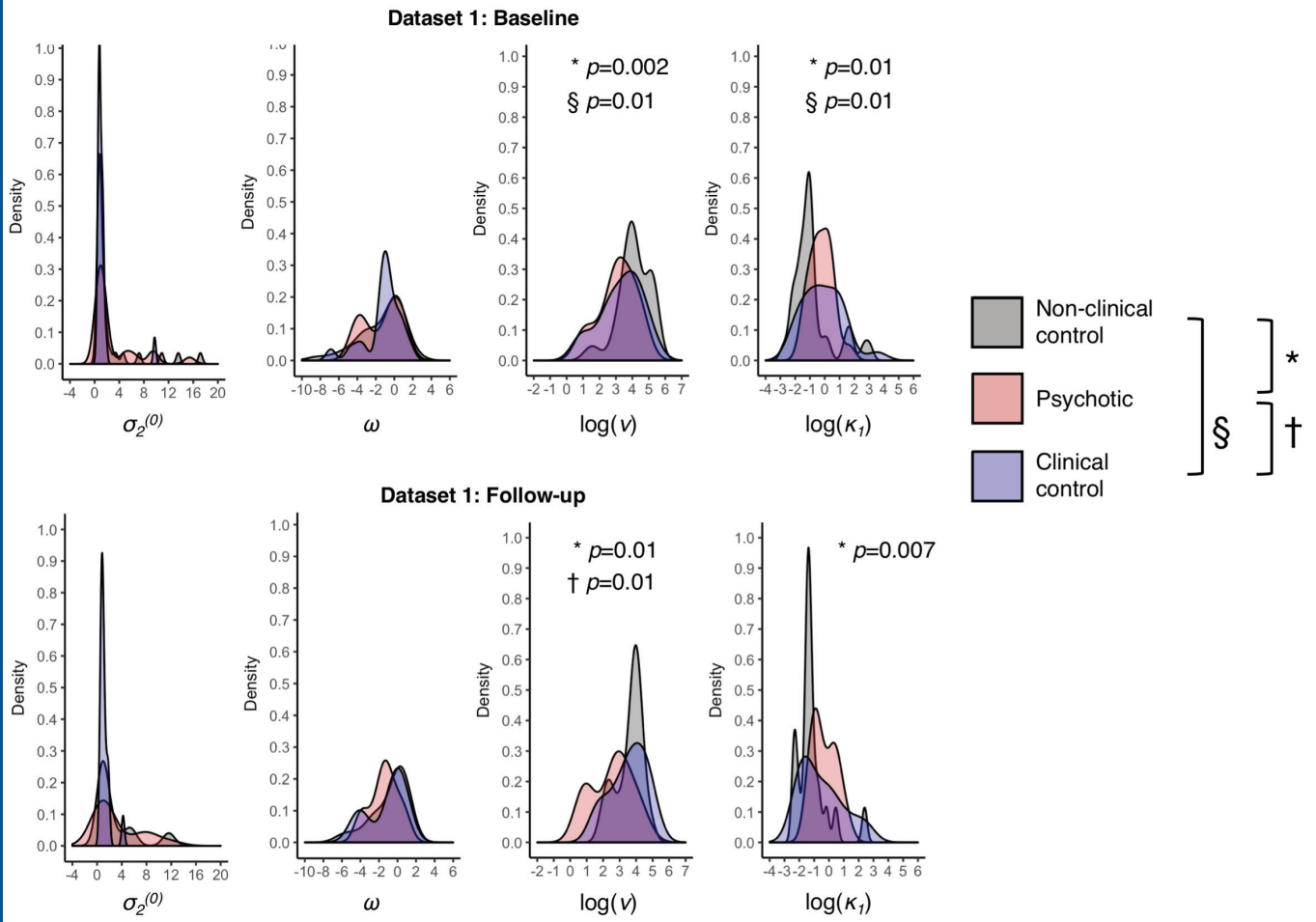




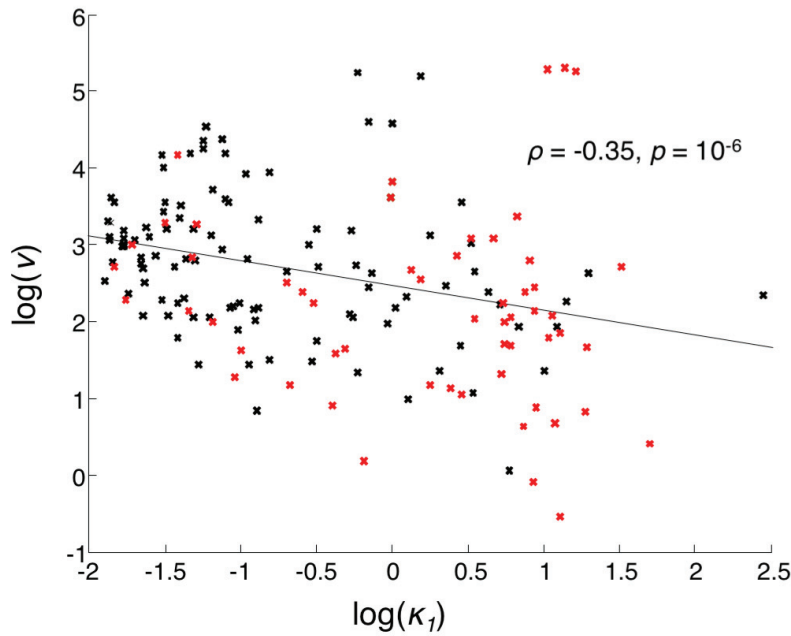
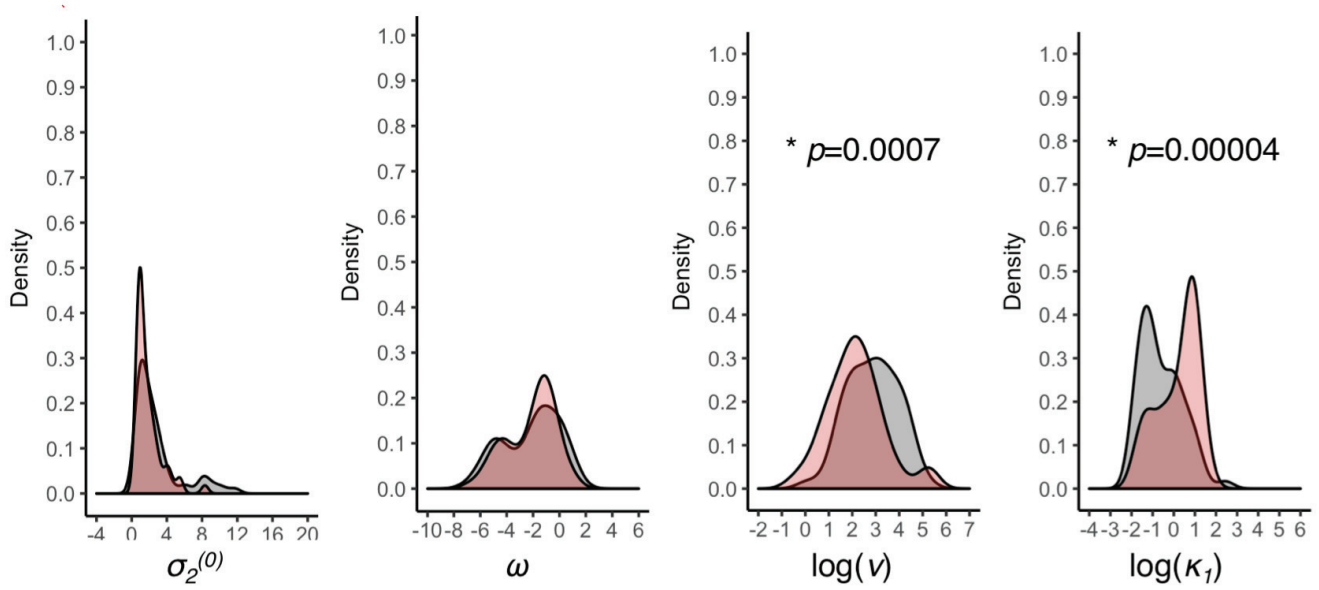




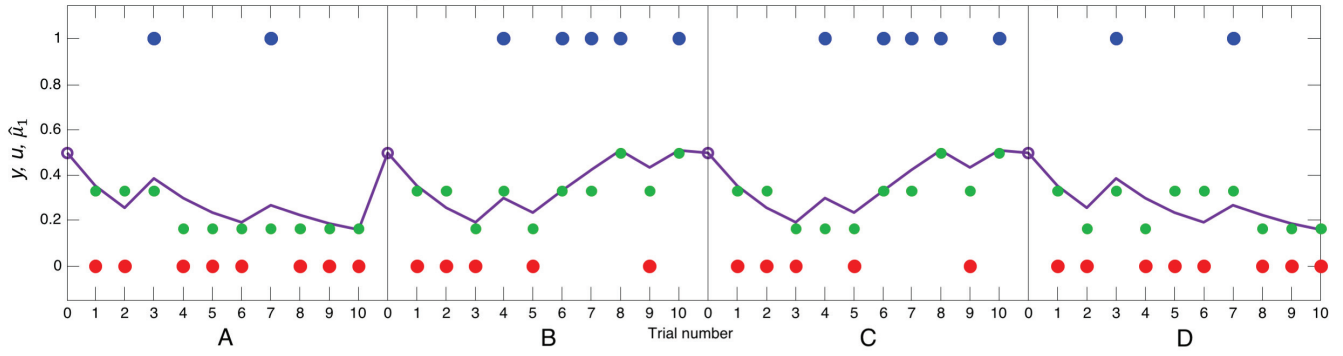




Dataset 2



Subject 18 (control):  $\sigma_2^{(0)} = 3.1$ ;  $\omega = -5.2$ ;  $v = \exp(3.9)$ ;  $\kappa_1 = \exp(-0.8)$



Subject 67 (control):  $\sigma_2^{(0)} = 4.5$ ;  $\omega = -0.87$ ;  $v = \exp(2.9)$ ;  $\kappa_1 = \exp(-1.1)$

