Cortical disinhibition, attractor dynamics and belief updating in schizophrenia

Rick A Adams[1,2]

[1]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, WC1N 3AZ

[2]Division of Psychiatry, University College London, 149 Tottenham Court Road, London, W1T 7NF

Abstract

Genetic and pharmacological evidence implicates *N*-methyl-D-aspartate receptor (NMDAR) dysfunction in the pathophysiology of schizophrenia. Dysfunction of this key receptor – if localised to inhibitory interneurons – could cause a net disinhibition of cortex, and increase in 'noise'. These effects can be computationally modelled in a variety of ways: by reducing the precision in Bayesian models of behaviour, by estimating neuronal excitability changes in schizophrenia from evoked responses, or – as described in detail here – by modelling abnormal belief-updating in a probabilistic inference task. Features of belief updating in schizophrenia include: greater updating to unexpected evidence, lower updating to consistent evidence, and greater stochasticity in responding. All of these features can be explained by a loss of stability of 'attractor states' in cortex and the representations they encode. Indeed, a hierarchical Bayesian model of belief updating indicates that subjects with schizophrenia have a consistently increased 'belief instability' parameter. This instability could be a direct result of cortical disinhibition: this hypothesis should be explored in future studies.

Keywords:

Schizophrenia; Psychosis; Computational; Beads task; Excitation-inhibition balance; Bayesian

A key challenge in schizophrenia-spectrum research is understanding how deficiencies in synaptic function in general and *N*-methyl-D-aspartate receptor (NMDAR) functioning in particular in the disorder – implied by genetic studies[1] and pharmacological models of psychosis that use the NMDAR antagonist ketamine[2] – impact on neural dynamics at the circuit, network, and whole-brain levels. A further challenge is understanding how these changes in neural dynamics then affect brain computations and behaviour.

NMDARs are located on both excitatory pyramidal cells and inhibitory interneurons, although different receptor subtypes are distributed differentially on different populations[3]. NMDAR dysfunction could therefore impact inhibitory interneurons or pyramidal cells to differing degrees: in 'subjects with a diagnosis of schizophrenia' (Scz), there is evidence that the former are more strongly affected[4], resulting in a net loss of inhibitory (relative to excitatory) transmission. This disinhibited state is also known as 'increased E/I balance'.

The consequences of increased E/I balance in Scz can be modelled in a variety of ways. One approach is to assume that this disinhibited state causes a decrease in precision (increase in variance) of the states that neural circuits encode: especially circuits at higher levels of the hierarchy, where there is most evidence for inhibitory dysfunction. One can then model these effects as the loss of precision of prior beliefs within a hierarchical Bayesian model of behaviour, as prior beliefs are most affected by loss of precision at the top of a hierarchical model. This approach has shown that numerous perceptual or behavioural phenomena in Scz can be modelled in this way, e.g. dysfunction of smooth pursuit eye movements, resistance to visual illusions, etc[5].

Another approach is to ignore behaviour altogether and just model neural responses. One of the best-validated electroencephalographic (EEG) findings in Scz is a reduction in the mismatch negativity[6]. The mismatch negativity is the difference in averaged EEG deflection in response to an oddball stimulus (e.g. a high tone following a series of low tones) compared to that following the standard. In Scz, there is less of a difference between EEG responses to oddballs and standards than there is in controls, and E/I balance could contribute to this.

Dynamic causal modelling (DCM) estimates how the activity in neural populations (e.g. pyramidal cells or interneurons) in connected brain areas evolves in response to some input (e.g. a sensory stimulus) according to the parameters of the system (e.g. the degree of disinhibition of pyramidal cells within areas, or the strength of connections between areas). DCM of mismatch negativity responses of Scz, their first degree relatives and healthy controls indicates that both Scz and their relatives have: i) an increase in disinhibition in the (right inferior) prefrontal area involved in the mismatch response, and ii) a reversal of the usual increase in excitability in response to oddballs (seen in controls) in that source[7]. This not only supports the notion of cortical disinhibition in Scz, but also implies that the regulation of neural excitability by stimulus predictability is awry in the disorder: as one might expect if prior beliefs are less precise. Indeed, reducing prior precision in a hierarchical Bayesian (predictive coding) model attenuates the prediction error responses to oddballs[5].

There are also differences between Scz and controls' resting state functional magnetic resonance imaging (rsfMRI) responses. Scz show greater power and variability of cortical rsfMRI data, especially in association cortices[8], and also greater connectivity (i.e. rsfMRI data correlations) between association areas[9]. Models of interacting cortical areas producing rsfMRI data can reproduce these effects if E/I balance within cortical areas is increased, although increasing coupling between areas also has similar effects[9]: it is hard to distinguish these model perturbations using fMRI data, as it is less temporally precise.

The most complete modelling approach is to relate neural function to behaviour using the same model. This is a complex procedure and there are few examples in Scz research. One successful example used a spiking network model consisting of pyramidal cells and interneurons to predict spatial working memory performance under ketamine or placebo[10]. Increasing E/I balance in this network (as ketamine is thought to do) allows activity to spread laterally through the network over time, making the spatial 'memory' less precise, and predicting increased false alarms to nearby non-target probes in a spatial working

memory task, as is seen under ketamine and also in Scz[11]. The persistent neural activity in the spiking spatial working memory model takes the form of a 'bump attractor', i.e. a subset of neurons which sustain activity from an input over time (the bump) whilst inhibiting local spread of this activity via inhibitory interneurons.

Attractors are essentially quasi-stable states of neural firing that can be implemented in a variety of ways. The first 'attractor networks' were designed to model the storage (and reactivation) of memories in patterns of synaptic weights[12]. In such networks, firing patterns more easily shift towards 'low energy' states, in which strongly connected neurons are active, and other neural activity is low. Once in such a state, the network has to receive a large perturbation to shift its firing pattern into a different state. If the energy of the network is plotted as a function of the neural firing patterns, one can visualise these low energy states as 'basins' in an energy landscape. The deeper the basin, the more difficult it is for the network to be shifted out of it. As well as modelling mnemonic processes, similar networks can also perform decision-making[13] and Bayesian belief updating[14].

For more than a decade, it has been hypothesized that changes in neural function in Scz might reduce the stability of cortical attractor states[15]. In particular, NMDAR dysfunction on both recurrent synapses on to pyramidal cells and on inhibitory interneurons could make firing patterns harder to sustain over time and less able to inhibit other firing patterns (respectively), making attractor basins more shallow. In this case, it would be more easy for the network to shift from one state to another – either due to an input that favours the other state or just to random neuronal spiking – but hard to maintain or 'deepen' any one state (Figure 1).

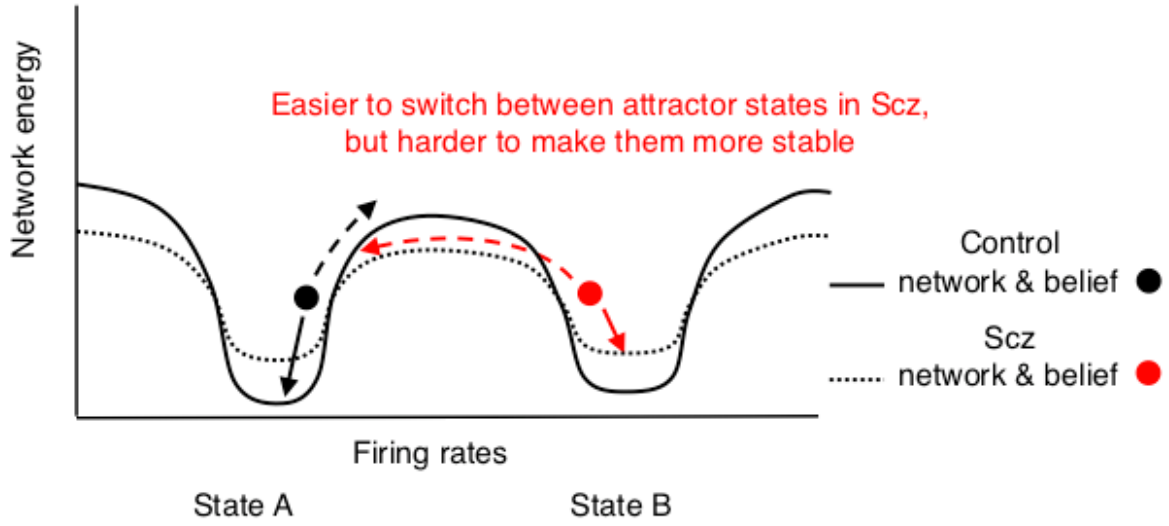


Figure 1: Potential effects of attractor network dynamics on belief updating

A loss of stable neural states was recently demonstrated in visual cortex of two animal models of schizophrenia[16], and interestingly, healthy volunteers given ketamine (which blocks NMDARs and is used as a model of psychosis) show a decrement in updating to

consistent stimulus associations and also increased decision stochasticity in this context[17]. In the remainder of this chapter, I shall describe a recent attempt to model alterations in (Bayesian) belief updating in Scz using a computational model designed to mimic the effects on inference of underlying neural attractor states with varying stability.

It has been known for decades that Scz tend to use less evidence than healthy controls to make decisions in belief updating tasks. The paradigm used to demonstrate this effect is often some variation of the 'beads' or 'urn' task[18], in which the participant is shown two jars containing beads of two colours in opposing ratios (e.g. 80:20 and 20:80 ratios of red:blue beads). The jars are then concealed and a sequence of beads drawn (with replacement) from one jar, and the participant has to either stop the sequence when they are sure of the source jar, or give a probability estimate of either jar being the source for the entire sequence. The former version is known as the 'draws to decision' task, and the latter as the 'probability estimates' task.

Well-replicated findings in the beads task include many Scz deciding on the jar identity after seeing only one or two beads[19] – the so-called 'jumping to conclusions' bias – and also Scz adjusting their beliefs more than controls after seeing unexpected evidence, termed a 'disconfirmatory bias'[18,20–23]. Although these biases appear to involve greater belief updating (i.e. higher learning rates) in Scz than in controls, in other tasks Scz seem to update less than controls – especially to longer sequences of more consistent evidence[24], and Scz are often more stochastic in their responding[25,26]. These three effects – greater updating to unexpected evidence, lower updating to consistent evidence, and greater stochasticity – are all consistent with an 'unstable attractor' model of belief updating, in which it is easy to switch from one state into another, but hard to stabilise (increase confidence in) any one state, and in which updates are more vulnerable to stochastic fluctuations in neural firing.

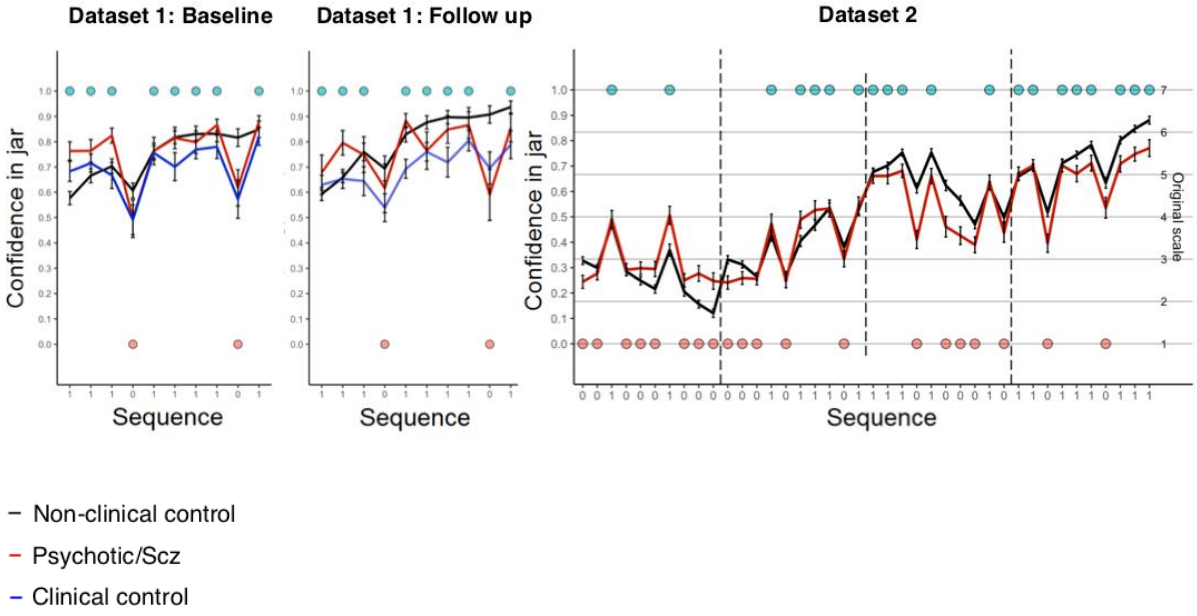


Figure 2: Two beads tasks datasets. Dataset 1[23]: n=80, including Scz and both clinical and non-clinical controls, tested both when unwell and in recovery; and Dataset 2: n=167, including Scz and non-clinical controls, tested as stable outpatients. In Dataset 2, subjects were each tested on four separate sequences, which are shown concatenated together.

Adams et al[27] used a hierarchical Bayesian model (the Hierarchical Gaussian Filter[28] – a variational Bayesian model with individual priors) to model belief updating in two independent 'probability estimates' beads task datasets (Figure 2). Models with a standard learning rate ω and response stochasticity v, or including a parameter increasing updating to 'disconfirmatory evidence' φ, or a parameter encoding belief instability κ_1 (Figure 3) were formally compared.

In these models, the belief about the jar on trial k+1, $x_2^{(k+1)}$, evolved according a Gaussian random walk of variance exp(ω):

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

In the response model, stochasticity v determined the width of the beta distribution centred on the current estimate of the jar probability (i.e. the prediction for the next trial), $\hat{\mu}_1^{(k+1)} \equiv s\left(\mu_2^{(k)}\right)$; here μ denotes the current estimate of x, and s is the sigmoid function.

In the 'disconfimatory bias' model, changes in x_2 from trial to trial occured according to an autoregressive (AR(1)) process controlled by three parameters: m, the level to which x_2 is attracted, φ, the rate of change of x_2 towards m, and ω, the variance of the random process:

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

Given there was no bias towards one jar or the other, m was fixed to 0, so φ always acted to shift the model's beliefs back towards maximum uncertainty (i.e. disconfirm the current belief) about the jar.

In the 'belief instability' model, changes in μ_2 from trial to trial occur according to two parameters: ω, the variance of the random process, and κ_1, a scaling factor that changes the size of updates when $\hat{\mu}_1 = 0.5$, or maximum uncertainty, relative to when $\hat{\mu}_1$ is closer to 0 or 1: $\hat{\mu}_1^{(k+1)} \equiv s\left(\mu_2^{(k)}\kappa_1\right)$. The effect of increasing κ_1 was to increase updating to unexpected evidence, but decrease updating to consistent evidence (Figure 4), as might be seen in a more unstable attractor network (although note that this model is merely simulating attractor network properties: it does not contain attractor states).

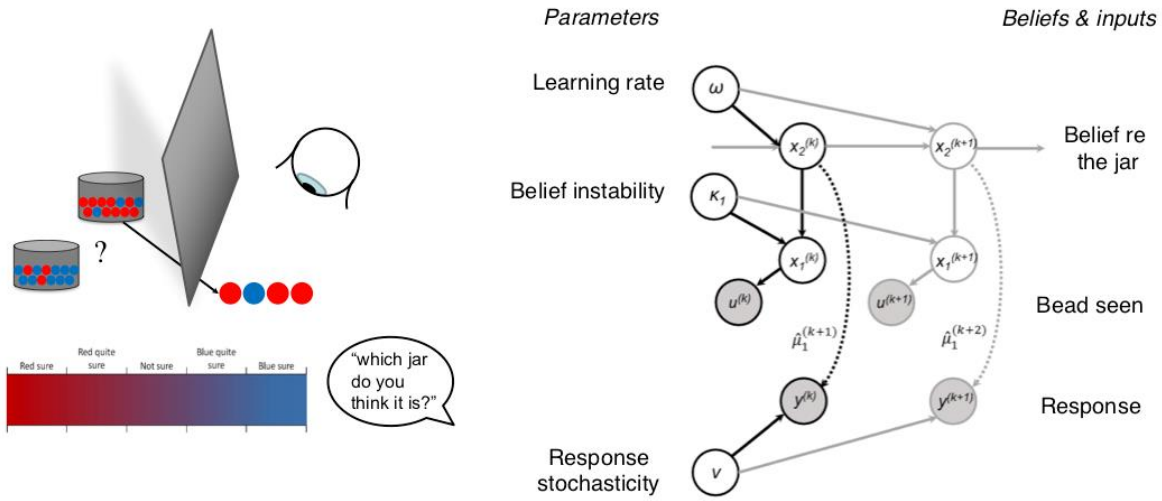


Figure 3: Left – the 'probability estimates' beads task; Right – the winning model. The right panel is a schematic representation of the generative model containing belief instability parameter κ_1. The black arrows denote the probabilistic network on trial k; the grey arrows denote the network at other points in time. The perceptual model lies above the dotted arrows, and the response model below them. The shaded circles are known quantities, and the parameters and states in unshaded circles are estimated. The dotted line represents the result of an inferential process (the response model builds on a perceptual model inference); the solid lines are generative processes. The response model maps from $\hat{\mu}_1^{(k+1)}$ (purple line) – the probability the blue jar is the source (x_1) on the next trial, itself a sigmoid function of the tendency towards the blue jar (x_2) – to $y^{(k)}$, the subject's indicated estimate of the probability the jar is blue. See Adams et al (submitted) for a full description of the model.

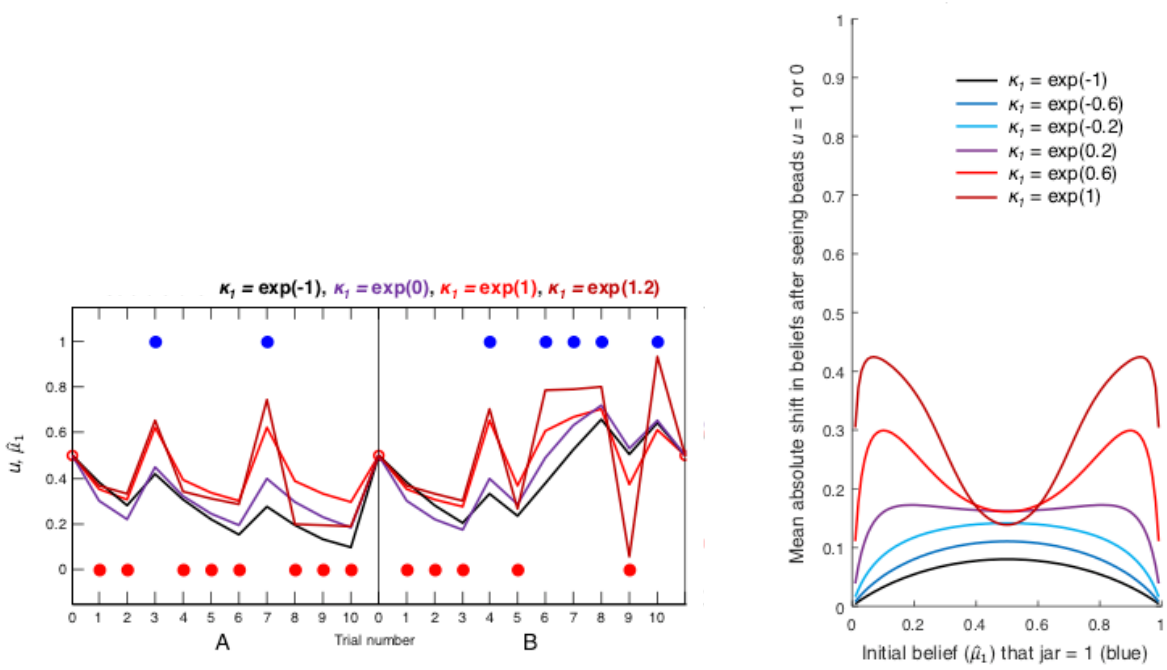
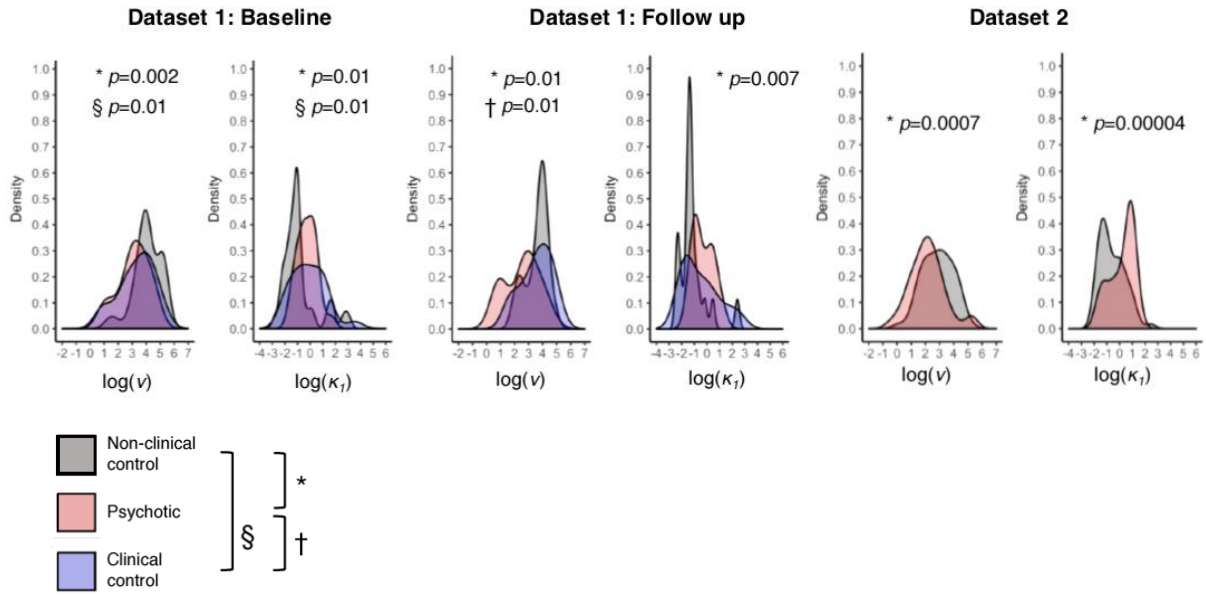


Figure 4: The effects of κ_1 on belief updating. Left panel: Simulated belief updating data $\hat{\mu}_1$ (in response to the bead sequence *u* on the plot) shows that higher κ_1 leads to overweighting of unexpected evidence (instability) and underweighting of consistent evidence**.** Right panel: This plot illustrates the average absolute shifts in beliefs on observing beads of either colour. This 'vulnerability to updating' is analogous to the 'energy state' of a neural network model (schematically illustrated in Figure 1) – i.e. in low energy states, less updating is expected. The effect of increasing κ_1 is to convert confident beliefs about the jar (near 0 and 1) from low to high 'energy states', i.e. to make them much more unstable.

The model containing learning rate ω, response stochasticity ν, and belief instability κ_1 won in all subjects in both datasets. Scz had greater belief instability (κ_1) and response stochasticity (ν) than non-clinical controls in both datasets (Figure 5). These parameters correlated in both datasets (Spearman's ρ = -0.38, -0.52 and -0.35; all p < 0.0001). Interestingly, when unwell, clinical controls' parameter distributions resembled those of Scz; but at follow-up, they resembled non-clinical controls.

Figure 5: Parameter differences between Scz, clinical and non-clinical controls. Scz consistently had higher belief instability κ_1 and greater response stochasticity ν than non-clinical controls; clinical controls resembled Scz when unwell and non-clinical controls when better.



Two computational studies of similar tasks in Scz have also demonstrated similar patterns of belief updating. Jardri et al[29] showed that on average, Scz 'overcount' the likelihood (i.e. the sensory evidence, in Bayesian terms) in a single belief update: the authors attributed this effect to disinhibited cortical message-passing, but it could also be due to the belief instability in the model above. Likewise, Stuke et al[30] showed in another beads task variant that Scz updated more than controls to "irrelevant information" (i.e. disconfirmatory evidence).

In conclusion, these results show that Scz subjects in two independent beads task datasets have consistent differences in two parameters of a belief updating model that attempts to

reproduce consequences of attractor network instability. More detailed spiking network modelling, pharmacological (or other NMDAR) manipulations and imaging are required in future to understand how neuromodulatory function in both pyramidal cells and inhibitory interneurons contributes to attractor dynamics and probabilistic inference.

References

1. Harrison, P. J. Recent genetic findings in schizophrenia and their therapeutic relevance. *J. Psychopharmacol. Oxf. Engl.* **29**, 85–96 (2015).
2. Javitt, D. C., Zukin, S. R., Heresco-Levy, U. & Umbricht, D. Has an angel shown the way? Etiological and therapeutic implications of the PCP/NMDA model of schizophrenia. *Schizophr. Bull.* **38**, 958–966 (2012).
3. Paoletti, P., Bellone, C. & Zhou, Q. NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. *Nat. Rev. Neurosci.* **14**, 383–400 (2013).
4. Weickert, C. S. *et al.* Molecular evidence of N-methyl-D-aspartate receptor hypofunction in schizophrenia. *Mol. Psychiatry* (2012). doi:10.1038/mp.2012.137
5. Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. & Friston, K. J. The computational anatomy of psychosis. *Front. Psychiatry* **4**, 47 (2013).
6. Umbricht, D. & Krljes, S. Mismatch negativity in schizophrenia: a meta-analysis. *Schizophr. Res.* **76**, 1–23 (2005).
7. Ranlund, S. *et al.* Impaired prefrontal synaptic gain in people with psychosis and their relatives during the mismatch negativity. *Hum. Brain Mapp.* **37**, 351–365 (2016).
8. Yang, G. J. *et al.* Altered global brain signal in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7438–7443 (2014).
9. Yang, G. J. *et al.* Functional hierarchy underlies preferential connectivity disturbances in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E219-228 (2016).

10. Murray, J. D. *et al.* Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cereb. Cortex N. Y. N 1991* **24**, 859–872 (2014).

11. Mayer, J. S. & Park, S. Working memory encoding and false memory in schizophrenia and bipolar disorder in a spatial delayed response task. *J. Abnorm. Psychol.* **121**, 784–794 (2012).

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

13. Wang, X.-J. The prefrontal cortex as a quintessential 'cognitive-type' neural circuit: Working memory and decision making. (2013).

14. Gepperth, A. & Lefort, M. Learning to be attractive: Probabilistic computation with dynamic attractor networks. in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* 270–277 (2016). doi:10.1109/DEVLRN.2016.7846831

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

16. Hamm, J. P., Peterka, D. S., Gogos, J. A. & Yuste, R. Altered Cortical Ensembles in Mouse Models of Schizophrenia. *Neuron* **94**, 153-167.e8 (2017).

17. Vinckier, F. *et al.* Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Mol. Psychiatry* **21**, 946–955 (2016).

18. Garety, P. A., Hemsley, D. R. & Wessely, S. Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. *J. Nerv. Ment. Dis.* **179**, 194–201 (1991).

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

20. Langdon, R., Ward, P. B. & Coltheart, M. Reasoning anomalies associated with delusions in schizophrenia. *Schizophr. Bull.* **36**, 321–330 (2010).

21. Fear, C. F. & Healy, D. Probabilistic reasoning in obsessive-compulsive and delusional disorders. *Psychol. Med.* **27**, 199–208 (1997).

22. Young, H. F. & Bentall, R. P. Probabilistic reasoning in deluded, depressed and normal subjects: effects of task difficulty and meaningful versus non-meaningful material. *Psychol. Med.* **27**, 455–465 (1997).

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

24. Averbeck, B. B., Evans, S., Chouhan, V., Bristow, E. & Shergill, S. S. Probabilistic learning and inference in schizophrenia. *Schizophr. Res.* (2010). doi:10.1016/j.schres.2010.08.009

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

26. Schlagenhauf, F. *et al.* Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage* (2013). doi:10.1016/j.neuroimage.2013.11.034

27. Adams, R. A., Napier, G., Roiser, J. P., Mathys, C. & Gilleen, J. Attractor-like Dynamics in Belief Updating in Schizophrenia. *J. Neurosci. Off. J. Soc. Neurosci.* **38**, 9471–9485 (2018).

28. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39 (2011).

29. Jardri, R., Duverne, S., Litvinova, A. S. & Denève, S. Experimental evidence for circular inference in schizophrenia. *Nat. Commun.* **8**, 14218 (2017).

30. Stuke, H., Stuke, H., Weilnhammer, V. A. & Schmack, K. Psychotic Experiences and Overhasty Inferences Are Related to Maladaptive Learning. *PLoS Comput. Biol.* **13**, e1005328 (2017).