doi:10.1093/brain/awz051 BRAIN 2019: 142; 1797–1812 | 1797

BRAIN
A JOURNAL OF NEUROLOGY

A distinct inferential mechanism for delusions in schizophrenia

Seth C. Baker,[1] Anna B. Konova,[2] Nathaniel D. Daw[3] and Guillermo Horga[1]

See Schmack and Sterzer (doi:10.1093/brain/awz092) for a scientific commentary on this article.

Delusions, a core symptom of psychosis, are false beliefs that are rigidly held with strong conviction despite contradictory evidence. Alterations in inferential processes have long been proposed to underlie delusional pathology, but previous attempts to show this have failed to yield compelling evidence for a specific relationship between inferential abnormalities and delusional severity in schizophrenia. Using a novel, incentivized information-sampling task (a modified version of the beads task), alongside well-characterized decision-making tasks, we sought a mechanistic understanding of delusions in a sample of medicated and unmedicated patients with schizophrenia who exhibited a wide range of delusion severity. In this novel task, participants chose whether to draw beads from one of two hidden jars or to guess the identity of the hidden jar, in order to minimize financial loss from a monetary endowment, and concurrently reported their probability estimates for the hidden jar. We found that patients with higher delusion severity exhibited increased information seeking (i.e. increased draws-to-decision behaviour). This increase was highly specific to delusion severity as compared to the severity of other psychotic symptoms, working-memory capacity, and other clinical and socio-demographic characteristics. Delusion-related increases in information seeking were present in unmedicated patients, indicating that they were unlikely due to antipsychotic medication. In addition, after adjusting for delusion severity, patients as a whole exhibited decreased information seeking relative to healthy individuals, a decrease that correlated with lower socioeconomic status. Computational analyses of reported probability estimates further showed that more delusional patients exhibited abnormal belief updating characterized by stronger reliance on prior beliefs formed early in the inferential process, a feature that correlated with increased information seeking in patients. Other decision-making parameters that could have theoretically explained the delusion effects, such as those related to subjective valuation, were uncorrelated with both delusional severity and information seeking among the patients. In turn, we found some preliminary evidence that subjective valuation (rather than belief updating) may explain group differences in information seeking unrelated to delusions. Together, these results suggest that abnormalities in belief updating, characterized by stronger reliance on prior beliefs formed by incorporating information presented earlier in the inferential process, may be a core computational mechanism of delusional ideation in psychosis. Our results thus provide direct empirical support for an inferential mechanism that naturally captures the characteristic rigidity associated with delusional beliefs.

1 Department of Psychiatry, New York State Psychiatric Institute, Columbia University Medical Center, 1051 Riverside Drive, New York, NY, USA
2 Department of Psychiatry, University Behavioral Health Care, and Brain Health Institute, Rutgers University – New Brunswick, 671 Hoes Lane West, Piscataway, NJ, USA
3 Department of Psychology and Princeton Neuroscience Institute, Princeton University, South Drive, Princeton, NJ, USA

Correspondence to: Guillermo Horga
Department of Psychiatry, New York State Psychiatric Institute, Columbia University Medical Center
1051 Riverside Drive, New York, NY, USA
E-mail: HorgaG@nyspi.columbia.edu

Received October 3, 2018. Revised December 27, 2018. Accepted January 16, 2019. Advance Access publication March 21, 2019

Keywords: delusions; schizophrenia; computational psychiatry; Bayesian inference; belief updating

Abbreviations: CAPS = Cardiff Anomalous Perceptions Scale; PANSS = Positive and Negative Syndrome Scale; PDI = Peters Delusion Inventory; POMDP = partially observable Markov decision process; PSYRATS = Psychotic Symptoms Rating Scale

Introduction

Delusions are erroneous, inflexible beliefs that are held with certainty even in light of contradictory evidence (Jaspers, 1913). The most common symptom of psychosis in schizophrenia (Andreasen and Flaum, 1991), delusions can be severely distressing and seriously impair social functioning. Despite this, the cognitive mechanisms underlying this symptom remain elusive. Elucidating these mechanisms, particularly at the computational, algorithmic level, is a critical step in advancing our conceptualization of psychosis.

Delusions have long been proposed to stem from abnormalities in inference, or the process of shaping beliefs through experience (Hemsley and Garety, 1986). This framework assumes that neural systems support inference on hidden states based on available sensory evidence, for instance, inferring someone's intentions (a hidden state) based on their observable actions (sensory evidence). Alterations in this inference process may, in turn, result in the formation of erroneous, delusional beliefs that conflict with the available evidence. From a Bayesian perspective, inference consists of a belief-updating process in which new evidence is optimally integrated with prior beliefs to form an updated (posterior) belief. This Bayesian formalism has been recognized as an appealing quantitative framework for understanding the formation and maintenance of normal and pathological beliefs (Fletcher and Frith, 2009; Adams *et al.*, 2013), but its value in explaining delusions in schizophrenia remains uncorroborated.

The hypothesis that delusions arise from altered inference has motivated a large body of work testing whether delusions indeed correlate with alterations in laboratory measures of inference (Huq *et al.*, 1988; Garety, 1991; Ross *et al.*, 2015; Dudley *et al.*, 2016). To interrogate inference, studies have often used variants of the 'beads' or 'urn' task, where participants draw beads from a hidden jar until they feel confident enough to guess the identity of the hidden jar (e.g. a jar with mostly green beads or one with mostly blue beads) given the observed sequence of beads. Here, the primary measure of inference is 'draws-to-decision', the number of beads drawn by a participant before making a guess. However, decades of work in this area have not produced a satisfactory explanation for delusions (Ross *et al.*, 2015). A diagnosis of schizophrenia has been consistently related to reduced draws-to-decision (Dudley *et al.*, 2016), but so have cognitive impairment (Bentall *et al.*, 2009; Freeman *et al.*, 2014), misunderstanding of task instructions (Balzan *et al.*, 2012), and feeling time-pressured during the task (White and Mansell, 2009), suggesting that performance on the classic beads task conflates a number of cognitive processes apart from inference. More troubling is that studies in schizophrenia have failed to show a relationship between draws-to-decision and delusion severity (Ross *et al.*, 2015). Thus, even assuming that a reduction in draws to decision actually reflected altered inference, this phenotype does not appear to account for clinical delusions. But the assumption that previous approaches can isolate inference is challenged by important methodological limitations. Most classic versions of the beads task used in schizophrenia research do not elicit consequential (incentive-compatible) decisions. They also generally lack sufficient controls for task comprehension, impatience, and general cognitive deficits (Ross *et al.*, 2015) [except for working-memory deficits (Dudley *et al.*, 1997)]. As a result, instead of reflecting inferential abnormalities, the reported reduction in draws-to-decision in schizophrenia could simply reflect miscomprehension of task instructions due to cognitive deficits, reduced task engagement due to motivational deficits, tolerance to uncertainty, or impatience. Furthermore, the lack of a tangible incentive in classic beads tasks, combined with the participants' ability to shorten the total task duration by guessing earlier, renders it possible that reduced draws-to-decision may even reflect an optimal allocation of cognitive effort or time given the subjective costs of both.

Here, we designed a novel approach aimed specifically at testing whether abnormalities in inference, rather than in other cognitive or decision-making processes, are associated with delusions in schizophrenia. Our main task consisted of a controlled, incentive-compatible version of the beads task. During each trial, participants decided whether to draw beads from the hidden jar, at a small cost, or guess the identity of the hidden jar to avoid making incorrect guesses, which incurred a larger penalty, and to keep as much money as possible from an initial endowment. Concurrently, participants gave subjective probability estimates reflecting their beliefs about the identity of the hidden jar on a draw-by-draw basis. This design allowed us to directly capture the dynamics of sequential belief updating within trials so as to interrogate inferential mechanisms. To assess potential effects of antipsychotic medication, another potential confound in the previous literature, we recruited medicated and unmedicated patients. We also capitalized on well-established approaches to evaluate decision-making variables other than inference: risk, ambiguity, and loss aversion. Given the phenomenological definition of delusions as rigid beliefs, we hypothesized *a priori* that delusions—beyond schizophrenia as a diagnostic category and beyond other symptoms of the illness—would be linked specifically to increased reliance on prior beliefs (in the current study, beliefs formed on the basis of

information presented during a trial rather than long-held beliefs formed on the basis of experiences prior to the task). We reasoned that this could lead to reduced draws-to-decision if patients exhibited strong prior beliefs biasing their inference towards one of the jars (at baseline or due to sequential effects). To compare delusion-related behaviours with optimal and normative behaviour on the task, we used an ideal-observer model and collected data from healthy individuals.

Materials and methods

Participants

Twenty-six patients with schizophrenia and 25 healthy volunteers were recruited through procedures approved by the New York State Psychiatric Institute (NYSPI) IRB. Among patients, 12 were taking antipsychotic medication at the time of the experiment and 14 were OFF antipsychotic medication for at least 3 weeks prior to the experiment. Patients met Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM-IV) criteria for schizophrenia, schizoaffective, or schizophreniform disorder [based on the Structured Clinical Interview for DSM-4 (SCID-IV) and consensus diagnosis]. Healthy controls had no DSM-IV Axis I diagnosis or psychotic illness in first-degree relatives. Common exclusion criteria were significant medical illnesses and misuse of substances other than nicotine (Table 1).

Clinical and sociodemographic measures

Our main measurement of delusional severity was the Peters Delusion Inventory (PDI) (Peters *et al.*, 2004), a widely-used scale that provides a fine-grained measure of delusion and delusion-like phenomena. We also measured hallucination and hallucination-like phenomena using the Cardiff Anomalous Perceptions Scale (CAPS) (Bell *et al.*, 2006) to assess the specificity of any behavioural phenotypes to delusions rather than to psychosis more generally. To assess clinical significance, we administered the Positive and Negative Syndrome Scale (PANSS) (Kay *et al.*, 1987). We also collected the Psychotic Symptoms Rating Scale (Haddock *et al.*, 1999) (PSYRATS) to confirm generalizability of our findings across scales. The Hollingshead scale of socioeconomic status (Hollingshead, 1975), the Edinburgh Handedness Inventory (Oldfield, 1971), the Employment/Support Status portion of the Addiction Severity Index (ASI) (McLellan *et al.*, 1980), and a Numeracy module of the 2002 Health and Retirement Study (HRS) were also administered. To assess effects of general cognitive function, we administered the Letter-Number Span (LNS) working-memory task (Nuechterlein *et al.*, 2008).

Procedures

Participants completed three decision-making tasks in fixed order. They first completed the modified beads task, followed by a loss-aversion task (Tom *et al.*, 2007) (Supplementary material), and a risk- and ambiguity-aversion task (Tymula *et al.*, 2013) (Supplementary material). Before the session, participants were given a $30 endowment in cash. They were instructed to put the cash endowment in a metal box that was kept next to them throughout the experimental session, until the payoff was realized at the end of the session. At that time, one of the three tasks was selected at random to determine the payoff based on the participant's choice for one randomly selected trial within the selected task. This was done to engage the participants and to minimize sequential effects, given that the payoff was determined by performance at any given moment rather than by cumulative performance throughout a task or during the entire session.

Modified beads task

We designed a modified beads task that built upon previous designs (Huq *et al.*, 1988; Furl and Averbeck, 2011) (see Supplementary material for details). To ensure task comprehension, participants underwent a detailed set of instructions, practice trials, and a comprehension quiz, in addition to a post-task debriefing (Supplementary material). Participants were instructed that, at the beginning of each trial, one of two jars was randomly selected (with equal probability): one containing mostly blue beads ('the blue jar') or one containing mostly green beads ('the green jar'). Participants were instructed that this selection would not be revealed to them, and that they had to correctly guess the identity of the selected (hidden) jar by drawing beads from it (up to a maximum of eight). The majority-to-minority ratio of coloured beads in the jar ('bead ratio') was shown on each trial, and could be 60:40, 75:25, 90:10, or 100:0. Participants were informed that the $30 endowment was applied to each trial, such that each bead draw would deduct $0.30 from this endowment and an incorrect jar guess would further deduct $15. They were instructed that their goal was to keep as much money as possible from the endowment. A minimum task duration of 30 min was imposed to disincentivize the use of strategies accounting for the cost of time, such as reducing the total duration of the task by guessing hastily to maximize reward rates. The task comprised 70 trials: 10 of the 100:0 bead-ratio ('catch') condition, and 20 of each of the remaining three conditions (60:40, 75:25, and 90:10). Trials of a given bead ratio were arranged in blocks of 10 trials, following a fixed, pseudo-random order (the Supplementary material describes the specific sequences).

Trial structure

Each trial began with a fixation screen (Fig. 1). Participants were then asked for an initial estimate of the probability of the hidden jar before drawing any beads (to elicit 'prior beliefs' before any sensory evidence was presented). Probability estimates generally served to scaffold the decision process and to obtain draw-by-draw reports of participants' beliefs about the hidden jar. Estimates were recorded via a visual analogue scale from 0% to 100%. For simplicity, estimates were always anchored to one of the two jars by asking participants to 'Please estimate the probability that you are currently drawing from the blue jar' or 'Please estimate the probability that you are currently drawing from the green jar', counterbalanced across participants. The next screen, which displayed the current winnings and last probability estimate, prompted a choice between drawing a bead and guessing one of the two jars (blue or green). If participants chose to draw, they were shown the identity of the next bead drawn. If participants chose to

Table 1 Sociodemographic and clinical characteristics of study sample

Characteristic	Patients (*n* = 24)	Controls (*n* = 21)	*P*-value
Sociodemographic characteristics			
Age, mean (SEM), years	36.8 (2.37)	35.9 (2.44)	0.8
Sex, male/female	18/6	15/6	0.8
Race/ethnicity (%)			0.9
African-American	11 (46)	8 (38)	
Caucasian	6 (25)	5 (24)	
Hispanic	3 (12.5)	4 (19)	
Asian	3 (12.5)	2 (9.5)	
Mixed	1 (4)	2 (9.5)	
Handedness, right/left	22/2	18/3	0.67
Socioeconomic status			
Personal SES, mean (SEM)	20.3 (1.6)	32.0 (2.8)	0.001
Parental SES, mean (SEM)	39.5 (2.3)	43.7(2.4)	0.22
ASI 30-day Money, mean (SEM), USD	787.96 (139.9)	1870.5 (158.9)	<0.001
Clinical characteristics			
Duration of illness, mean (SEM), years	12.3 (2.1)	n/a	n/a
Medication status, medicated/unmedicated	13/11	n/a	n/a
PDI global, mean (SEM), [range, 0–336]	67.5 (10.3), [0–170]	13.6 (3.9), [0–51]	<0.001
PDI total, mean (SEM), [range, 0–315]	60.4 (9.4), [0–153]	11.8 (3.4), [0–45]	<0.001
CAPS global (summed), mean (SEM), [range, 0–512]	90.2 (15.1), [0–255]	4.86 (2.6), [0–54]	<0.001
PANSS			
PANSS positive total score, mean (SEM) [range, 7–49]	15.4 (0.9), [7–29]	n/a	n/a
PANSS negative total score, mean (SEM) [range, 7–49]	17.2 (1.0), [9–27]	n/a	n/a
PANSS general total score, mean (SEM) [range, 16–112]	33.2 (1.4), [20–48]	n/a	n/a
Numeracy (2002 HRS)			
Numeracy % correct responses, mean (SEM) [range, 0–100]	77 (5.5) [0–100]	87 (4.0), [50–100]	0.17
Numeracy reported difficulty, mean (SEM) [range, 0–4]	1.5 (0.25) [0–4]	1.43 (0.21), [0–4]	0.83
Letter-Number Span (LNS), mean (SEM) [range, 0–24]	15.7 (0.81) [8–22]	17.6 (0.70), [10–22]	0.1
Nicotine smoking, no/yes	20/4	19/2	0.49

Data are presented as *n* (percentage) of participants unless otherwise indicated. *P*-values correspond to two-sample *t*-tests for continuous variables and χ^2 for categorical variables. Note that here we report both PDI total and global scores for comparison with the previous literature; we use the total score for the main analyses, but the two scores are correlated with each other at r = 0.99.
ASI = addiction severity index; n/a = data are not applicable.

guess, they received feedback on the accuracy of their guess and their final winnings for the trial. To minimize time pressure, responses to the estimate and choice prompts were self-paced and had no time limit. In addition, the estimate and choice screens displayed the sequence of beads drawn from the start of the trial to that point [a memory aid aimed at minimizing the impact of working-memory deficits on performance, following prior work (Dudley *et al.*, 1997)] as well as the bead-ratio condition.

Data analysis

Model-agnostic analyses of the beads task

Our main tests used Pearson correlation and regression to assess the relationships between delusion severity (primarily measured by the PDI) in patients and draws-to-decision behaviour (or model-derived parameters describing belief-updating and other processes underlying behaviour on the beads task; see below). Here and in model-based analyses, only data from correct trials (95.3% of all trials on average) were analysed to avoid artefactual results driven by distractions or extraneous events, but all main results held when all trials were analysed (Supplementary material). To assess specificity, multiple linear regression models including primary (mainly PDI scores) and secondary variables were used to control for the secondary variables. *Post hoc* analyses of single-item clinical scores used non-parametric tests. Although our primary aim was to identify belief-updating correlates of delusion severity in patients, secondary analyses compared patients, divided into high- and low-delusion patients based on a median split of their PDI total scores, to healthy controls. These analyses used ANOVAs and *post hoc* Dunn-Sidak tests. To support the interpretation of our model-based analyses of belief updating, *post hoc*, time-lagged multiple linear regression analyses were also used. Here, time-lagged analyses for each participant consisted of a model predicting the participant's reported probability of the hidden jar on each draw (expressed in reference to the actual hidden jar on each trial) based on whether the presented beads in the current draw (d), one draw back ($d-1$), and two draws back ($d-2$), were of the majority or the minority colour (each coded as 1 or 0, respectively). Statistical significance was set at $P<0.05$, two-sided (Greenhouse-Geisser corrected for ANOVAs). (For comparison with prior work, analyses of responses to 'disconfirmatory evidence' are presented in the Supplementary material.)

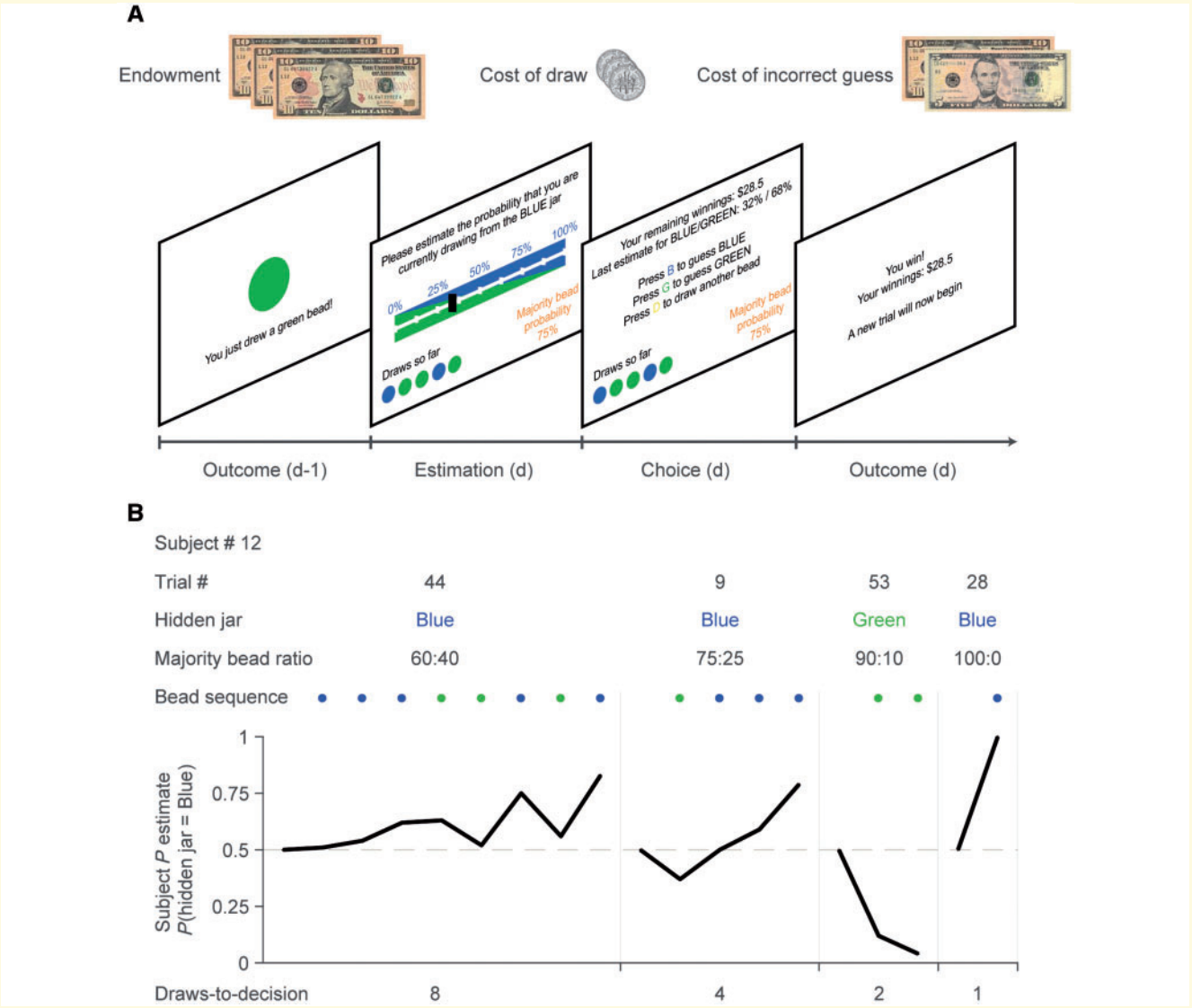

Figure 1 Modified beads task schematic and behaviour for an example subject. (**A**) Schematic depicting trial structure in the beads task. During sequential periods within a trial (estimation, choice, and outcome), participants were first asked for probability estimates about the identity of the hidden jar, and were then prompted to choose between drawing (for $0.30) or guessing at the identity of the hidden jar (incurring a penalty of $15 if their guess was incorrect). Their goal was to keep as much money as possible from an initial endowment of $30. A visual aid indicating the sequence of draws up to the current one was presented throughout (bottom left of screen). The bead-ratio condition (60:40, 75:25, 90:10, or 100:0) was also shown as the percentage of the majority bead probability in the hidden jar (bottom right of screen). The remaining winnings for a given trial were also displayed during the choice period (top centre of choice screen). The outcome period either revealed the drawn bead or provided feedback on the accuracy of a given guess and indicated the final winnings for the trial. (**B**) Behaviour for a representative subject on four different trials, one for each bead-ratio condition (majority bead ratios of 60:40, 75:25, 90:10, and 100:0). Probability estimates given before each draw are presented for each trial's bead sequence (*top*). Below, draws-to-decision are shown for each trial (*bottom*). These two measures represent the two main behaviours of interest in the task.

Model-based analyses of the beads task

Ideal observer model

Following prior work in health (Furl and Averbeck, 2011), we used a finite-horizon, partially observable Markov decision process (POMDP) (Kaelbling *et al.*, 1998) to prescribe an optimal strategy to maximize financial rewards in the task (Averbeck, 2015). In short, this model balances the expected financial costs of information seeking (i.e. drawing more beads) with those associated with actions potentially leading to errors (i.e. guessing the hidden-jar identity too soon). It can be roughly divided into three components: Bayesian belief updating, value comparison, and choice. First, at each step the model estimates the conditional probability of each of the two possible hidden jars. Then, the model estimates the expected value of guessing at the identity of the hidden jar in the current moment [considering the cost of a potential

incorrect guess ($15)] and the expected value of future guesses after drawing a bead [considering the cost of drawing a bead ($0.30)] and of potential incorrect guesses in the future. The intuition here is that, at the beginning of the trial, uncertainty about the identity of the hidden jar is high and so is the associated likelihood of an incorrect guess, which justifies paying to draw more beads (i.e. the expected value of drawing is higher than that of guessing). But after drawing enough beads, guessing becomes more valuable than drawing additional beads (i.e. the expected value of guessing is higher than that of drawing), and so the financially optimal strategy at that point is to guess. We thus constructed an 'ideal observer' that determined the optimal draws-to-decision behaviour for each trial in our task (Supplementary material).

Parameterized belief-updating models

To test whether abnormal belief updating could explain delusion effects on draws-to-decision behaviour ('Results' section), we modelled draw-by-draw changes in the probability estimates (beliefs) reported by the participants about the identity of the hidden jar. In modelling belief updating, consistent with prior work, we chose to fit probability estimates rather than choice behaviour because draw versus guess choices on this task can be driven by value-based processes secondary to the belief-updating process whereas probability estimates are presumably a more direct reflection of the belief-updating process itself.

Bayesian inference (Supplementary material) can be used to frame belief updating as an iterative process in which new sensory evidence (colour of the drawn beads) is integrated with prior beliefs to shift beliefs progressively towards one of two possible states (e.g. towards the belief that the blue jar B is more likely to be the hidden jar than the green jar G). This process can also be thought of as a form of sensory-evidence accumulation similar to that represented by neuronal activity in associative brain regions (Gold and Shadlen, 2002). In our task, consider a participant who has drawn n_b blue beads and n_g green beads, who then chooses to draw a new bead, which is revealed to be blue. At that point, the ratio of the conditional probabilities (updated or 'posterior' beliefs) for the hidden jars is:

$$\frac{P(B|(n_b+1),n_g)}{P(G|(n_b+1),n_g)} = \frac{P(B|n_b,n_g)\cdot q}{P(G|n_b,n_g)\cdot(1-q)} \quad (1)$$

The q and $(1-q)$ terms on the right-hand side of Equation 1 can be thought of as the Bayesian 'likelihoods' (i.e. the likelihood of drawing a blue or a green bead from a given jar), corresponding to the sensory evidence associated with the new bead, and the $P(B|n_b,n_g)$ and $P(G|n_b,n_g)$ terms can be thought of as Bayesian 'prior' beliefs (i.e. the probability that one of the jars is the hidden jar given the observed sequence of beads before the last draw). This expression is often transformed into a simple additive process in logit space, as:

$$log\left(\frac{P(B|(n_b+1),n_g)}{P(G|(n_b+1),n_g)}\right) = log\left(\frac{P(B|n_b,n_g)}{P(G|n_b,n_g)}\right) + log\left(\frac{q}{1-q}\right) \quad (2)$$

The $log\left(\frac{q}{1-q}\right)$ term on the right-hand side now corresponds to the log-likelihood ratio (LLR). Equation 1 can therefore be expressed as:

$$b_{d+1} = b_d + LLR \quad (3)$$

Here, the prior belief b_d (at draw d) is updated by adding the LLR to form a posterior belief b_{d+1} (at draw $[d+1]$). The belief update depends on the magnitude of LLR, the strength of the new sensory evidence, which itself depends on the bead ratio. Moreover, Equation 3 provides a simplified description of the belief-updating process that lends itself to a number of models capturing interindividual variability in this process (Ambuehl and Li, 2018). Our main belief-updating model variants were weighted versions of Equation 3, such as:

$$b_{d+1} = \omega_1 \cdot b_d + \omega_2 \cdot LLR \quad (4)$$

Here, the ω_1 parameter is a multiplicative weight on the prior term (i.e. the prior weight) that describes the proportional contribution to the posterior belief by sensory evidence presented earlier versus later in the trial. Mathematically, ω_1 can be cast in terms of a primacy-recency bias: higher values of ω_1 reflect a primacy bias by which previously incorporated information exerts a greater relative influence on posterior beliefs, and lower values of ω_1 reflect instead a recency bias (Supplementary material). The ω_1 parameter can also be thought as controlling how 'leaky' (given $0<\omega_1<1$) the prior is, similar to leaky integrators in extended drift-diffusion models (Usher and McClelland, 2001; Bogacz *et al.*, 2006; Brunton *et al.*, 2013). In turn, the ω_2 parameter is a multiplicative weight on the likelihood term (i.e. the likelihood weight) that scales the LLR, thereby directly modulating the contribution of new sensory evidence to the formation of posterior beliefs, analogous to the drift-rate parameter in drift-diffusion models (Ratcliff and McKoon, 2008). Models with a single ω_2 parameter systematically underestimated subjects' reported probabilities in conditions with weaker sensory evidence (e.g. 60:40 bead ratio), an effect that was accounted for by incorporating one ω_2 parameter per bead-ratio condition [i.e. $\omega_2(q)$]. We tested 10 belief-updating model variants (including the unweighted, parameter-free Bayesian model in Equation 3). Most models consisted of different combinations of these parameters, as well as different ways of incorporating these parameters by bead-ratio condition (two included additional parameters; Supplementary material).

Model fitting was performed separately for each participant by minimizing the root mean squared error (RMSE) between estimated probabilities and the actual probabilities reported by the participant (Supplementary material). We conducted formal model comparisons taking into account model fit and complexity by calculating the Bayesian Information Criterion (BIC) for each model and individual. We then performed group-level model selection by computing the Bayesian model evidence (Stephan *et al.*, 2009; Rigoux *et al.*, 2014) over BICs. Model fits and parameter recovery for the winning model ('1 ω_1, 4 ω_2') were satisfactory (Supplementary material).

Parameterized variant of the POMDP ideal observer

To rule out a possible contribution of value-based decision making in delusions, we tested a parameterized version of the POMDP model described above (Furl and Averbeck,

2011). Again, this model comprises three components, arranged hierarchically: belief updating, valuation, and choice. Valuation depends on belief updating, as computation of action values depends on the probability of different outcomes, and choices depend on valuation of alternative actions and comparison of their respective values. We assumed that the winning belief-updating model provided the best approximation to the belief-updating process participants engaged in, and we thus built a value-based decision-making model that used the fitted probability estimates derived from the winning belief-updating model (although a model based instead on the objective probabilities yielded similar results; Supplementary material). We fitted the parameterized POMDP model, which included five free parameters for valuation and choice (specifically, one additional subjective-cost parameter C_{sub} for drawing by condition, and one inverse-temperature parameter γ capturing choice stochasticity), to the drawer-by-drawer choice behaviour of each individual (Supplementary material).

Data availability

The data reported here are available from the corresponding author upon reasonable request.

Results

Task-comprehension checks

Six of the original 51 participants (two patients and four controls) misunderstood the instructions, as determined by predefined task-miscomprehension criteria (Supplementary material), and were thus excluded. All of the remaining 45 participants showed evidence of having properly understood the task (Supplementary material).

Clinical characteristics of patients with schizophrenia

We included 24 patients with a diagnosis of schizophrenia. Eleven (45.83%) were unmedicated. Delusion severity was highly variable [PDI total scores (Peters *et al.*, 2004) ranged from 0 to 153; mean ± SEM: 60.41 ± 9.56] and so was duration of illness (12.25 ± 2.10). Clinical characteristics (Table 1) were comparable to those in previous studies using the classic beads task (Corcoran *et al.*, 2008; So *et al.*, 2008; Langdon *et al.*, 2010; Lincoln *et al.*, 2010; Jacobsen *et al.*, 2012) (except medication status, as most previous studies in schizophrenia examined medicated patients).

Relationship between delusion severity and draws-to-decision behaviour in patients

Contrary to the hypothesized relationship with draws-to-decision, higher delusion severity in patients correlated with increased—more conservative—draws-to-decision behaviour (Fig. 2A): patients with higher PDI scores showed greater increase in draws-to-decision in conditions with smaller majority-to-minority bead ratios (draws-to-decision slope across bead-ratio conditions, $r = 0.51$, $P = 0.01$). This was driven by increased draws-to-decision in the 60:40 condition ($r = 0.49$, $P = 0.014$; note that draws-to-decision in the 60:40 condition and the draws-to-decision slope correlated at $r = 0.9$). The relationship between PDI scores and draws-to-decision slope was significant within unmedicated patients alone ($r = 0.79$, $P = 0.004$). This relationship was also apparent for PDI subscores for specific delusional dimensions across all patients (distress: $P = 0.006$; preoccupation: $P = 0.042$; conviction: $P = 0.008$) and in unmedicated patients alone (distress: $P = 0.002$; preoccupation: $P = 0.025$; conviction: $P = 0.003$). Consistent with a more conservative strategy and inconsistent with the expression of a general cognitive impairment, patients with higher delusion-related PDI scores were more—not less—accurate in the 60:40 condition ($r = 0.55$, $P = 0.005$).

To assess the specificity of the relationship between draws-to-decision behaviour and delusion severity, rather than to psychosis more generally or to other cognitive and socioeconomic variables, we examined whether the effect of delusion severity held after removing the shared variance between this symptom and other relevant variables (Fig. 2A–D). In particular, delusion-related PDI scores and hallucination-related CAPS scores were expected to, and did, share substantial variance among patients ($R^2 = 0.28$). We used a multiple-linear-regression model simultaneously including delusion-related PDI scores, hallucination-related CAPS scores, medication status, working-memory LNS score, numeracy-test accuracy, income, and socio-economic status, each as a separate predictor. The association between delusion severity and draws-to-decision slope held in this model [PDI: $t(16) = 2.63$, $P = 0.018$]. None of the other predictors showed a significant effect, including hallucination severity [CAPS: $t(16) = -0.62$, $P = 0.54$; all other variables: $P > 0.08$]. Furthermore, delusion severity in this model had a stronger effect than hallucination severity [linear contrast of regression coefficients for PDI > CAPS ($\beta_1 > \beta_2$): $t(16) = 2.12$, $P = 0.049$]. Altogether, these results suggest a specific relationship between delusion severity and draws-to-decision behaviour, beyond other psychotic symptoms, medication, and cognitive and socio-economic characteristics. Further supporting this specificity, other than PDI, the only scores that correlated with draws-to-decision slope even marginally were other (secondary) measures of delusions (Fig. 2E). Duration of illness did not correlate with PDI scores or draws-to-decision slopes (both $P > 0.35$). For more detailed analyses of medication effects, see Supplementary material. Control analyses in a large, independent dataset suggested that cognitive variables not measured in the current study were unrelated to delusion severity, and were thus unlikely confounders (Supplementary material).

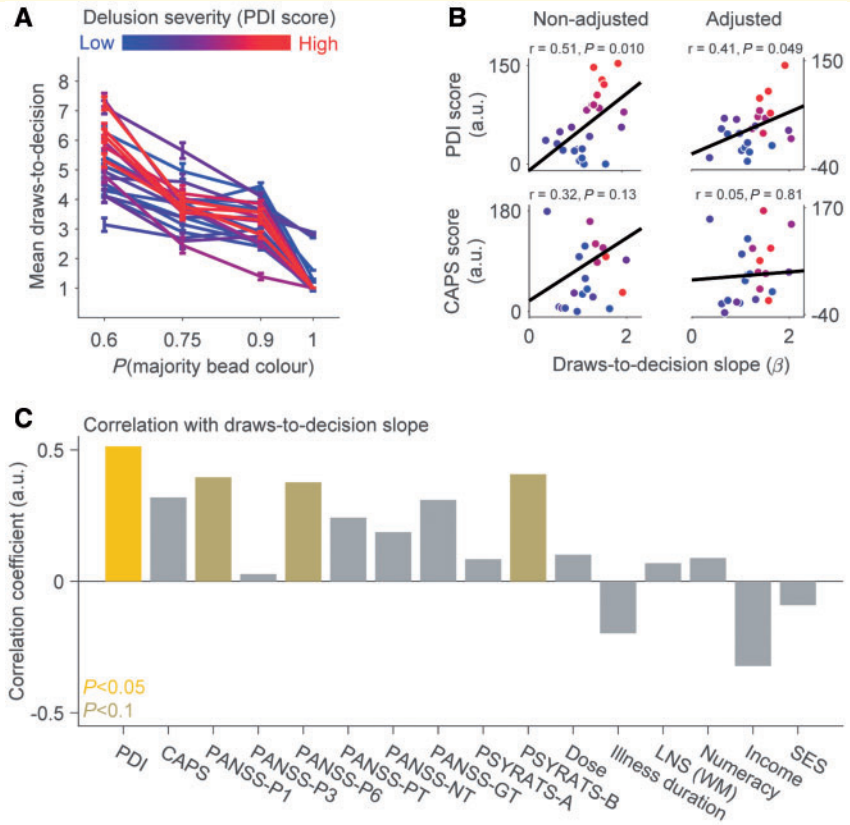

Figure 2 Specific relationship between draws-to-decision behaviour and delusion severity in patients. (**A**) Mean draws-to-decision (*y*-axis) for each bead-ratio condition (*x*-axis, probability of majority bead colour in the hidden jar) are shown in patients with schizophrenia. Error bars represent SEM. Lines are coloured by delusional severity (PDI score), with greater severity indicated in red and lower severity in blue. Patients with more severe delusions show increased draws-to-decision in the 60:40 condition and increased draws-to-decision slope (significant effects held after excluding subjects with draws-to-decision above 2 in the 100:0 condition; all $P < 0.03$). (**B**) *Top*: Scatterplots depicting the relationship between draws-to-decision slope (indicating the change in draws-to-decision as a function of bead-ratio condition) and severity of delusions before (*top left*) and after (*top right*) adjusting for severity of perceptual disturbances (CAPS score) in patients. *Bottom*: Scatterplot showing the relationship between severity of perceptual disturbances (CAPS score) and draws-to-decision slope, before (*bottom left*) and after (*bottom right*) adjusting for delusional severity (PDI score). (**A** and **B**) a.u. = arbitrary units. Dots are coloured by delusional severity (PDI score) as in **A**. (**C**) For *post hoc* assessment of specificity and of generalizability across scales, correlation coefficients (Pearson's r for summed scores and Spearman's ρ for single-item scores) are presented describing the strength of the relationships between the draws-to-decision slope and various clinical, neurocognitive, and socioeconomic variables in patients. Only delusion-related measures show significant ($P < 0.05$; gold) or trend-level ($0.05 < P < 0.1$; darker gold) effects. PDI = PDI total score; CAPS = Cardiff Anomalous Perceptions Scale, global (summed) score; PANSS-P1 refers to the 'delusions' item score, PANSS-P3 to the 'hallucinatory behaviour' item score, PANSS-P6 to the 'persecution/suspiciousness' item score, PANSS-PT to the positive subscale total score, PANSS-NT to the negative subscale total score, and PANSS-GT to the general subscale total score; PSYRATS-A refers to auditory hallucination total scores and PSYRATS-B refers to delusion total scores; Dose = antipsychotic medication dose in chlorpromazine equivalents (mg/day); Illness duration = duration of illness in years as per the SCID-IV; LNS = Letter-Number Span working-memory (WM) task performance score; Numeracy = per cent accuracy on the numeracy module of the 2002 HRS; Income = monthly income ($) measured by the employment section of the ASI support status; SES = personal socioeconomic status measured via the Hollingshead scale.

Post hoc comparisons with ideal-observer model and healthy controls

Splitting patients based on the median PDI score (52.5) for *post hoc* interpretive purposes, there was a group difference between high-delusion patients, low-delusion patients, and healthy controls [one-way ANOVA on draws-to-decision slope: $F(2,42) = 6.61$, $P = 0.0032$]. Numerically, high-delusion patients were more conservative than healthy controls but group differences were not significant (Dunn-Sidak *post hoc* test: $P = 0.36$). Low-delusion patients were significantly more liberal than either controls ($P = 0.045$) or high-delusion patients ($P = 0.003$). Relative to the ideal observer, all three groups exhibited more conservative behaviour, with high-delusion patients being most conservative (Fig. 3A).

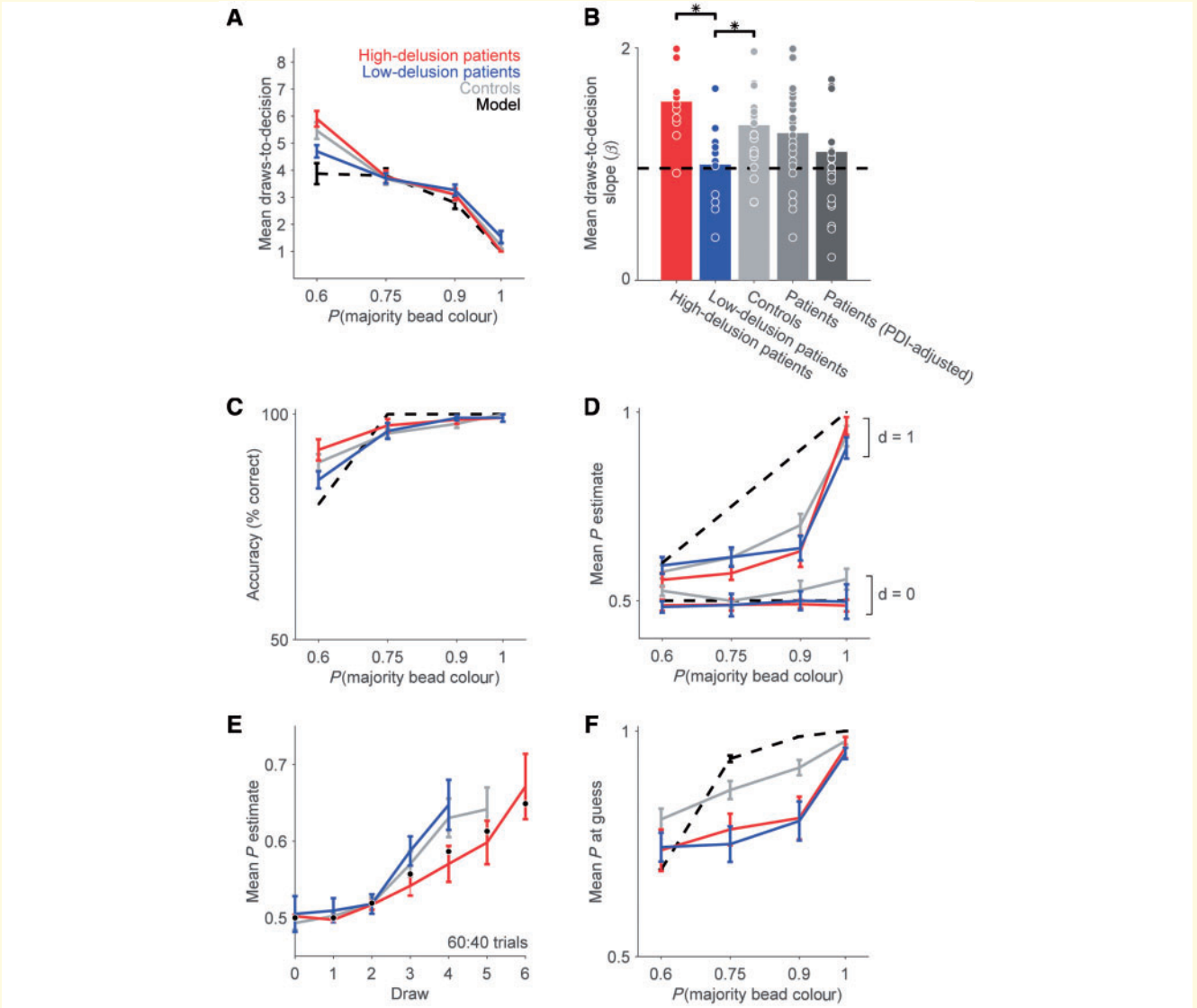

Figure 3 Task behaviour for patient subgroups, healthy controls, and ideal observer. (**A**) Mean draws-to-decision by group, with schizophrenia patients median-split into high-delusion (red) and low-delusion (blue) subgroups, for each bead-ratio condition. The socio-demographically matched healthy control group is shown in grey. The behaviour of the (parameter-free POMDP) ideal-observer model is indicated by the dashed black line. (**B**) Mean draws-to-decision slope for each group. Shown in shades of grey is this slope for controls and for all the schizophrenia patients grouped together (with and without adjustment by PDI score). Dots represent data for individual subjects. Asterisks indicate statistically significant effects at $P < 0.05$. (**C**) Accuracy by group for each bead-ratio condition. (**D**) Mean probability estimate (for the actual hidden jar) before (draw 0) and after the first bead draw (draw 1) by group and condition. (**E**) Mean probability estimates (for the actual hidden jar) before each bead draw, by group, across correct trials in the 60:40 condition. For ease of visualization, the corresponding probability estimates for the ideal-observer model are shown as black dots. (**F**) Mean probability estimate (for the actual hidden jar) at guess (i.e. probability estimate immediately preceding a guess choice) by group, in each bead-ratio condition. (**A**–**F**) The colour scheme in **A** applies to all panels. Error bars represent SEM.

Effects of delusion severity versus effects of diagnosis across patients and controls

To explore the effects of schizophrenia as a diagnostic category, separate from the effects of delusion severity, we also compared all patients versus controls in draws-to-decision slopes after controlling for PDI scores (Fig. 3B): we found support for an additive model whereby a diagnosis of schizophrenia correlates with more liberal behaviour (in accord with prior work), while delusion severity instead correlates specifically with more conservative behaviour [diagnosis: $t(42) = -2.07$, $P = 0.044$; PDI: $t(42) = 2.80$, $P = 0.008$]. General dysfunctions in social and cognitive functioning typically associated with schizophrenia could, however, explain the diagnosis effects in this model. Under

an extended regression model (including diagnosis, PDI scores, LNS working-memory score, numeracy-test accuracy, income, personal socio-economic status, and parental socio-economic status), diagnosis effects were indeed explained away [$t(35) = 1.59$, $P = 0.12$], particularly by personal socio-economic status [$t(35) = 2.51$, $P = 0.016$; all other $P > 0.24$], while delusion-severity effects remained unchanged. PDI scores: [$t(35) = 3.01$, $P = 0.005$; see Supplementary material for analyses by group].

Model-based analyses of belief updating in patients

Having found that delusion severity related specifically to more conservative draws-to-decision behaviour, we sought to dissect the underlying computational mechanisms for this effect by examining participants' probability estimates about the hidden jar, as reported on a draw-by-draw basis. Our *a priori* hypothesis was that delusions would relate to increased reliance on prior beliefs. This could have induced more liberal behaviour on the task if delusional patients tended to exhibit stronger baseline or sequential biases. However, our data were clearly at odds with this scenario. First, patients with more severe delusions tended to exhibit more conservative behaviour. Second, the reported probability estimates at the outset (baseline) of a trial—before any draws—were not significantly biased away from 0.5 nor different between the three groups (Fig. 3D, all within-group one-sample *t*-tests for differences from 0.5: $P > 0.06$; all between-group two-sample *t*-tests: $P > 0.05$ for signed or absolute probability estimates at draw 0). Given this, we reasoned that the hypothesized over-reliance on prior beliefs could instead induce slower updating and explain the conservative draws-to-decision behaviour we observed in more delusional patients. Indeed, inspection of the reported (raw) probability estimates in the 60:40 condition revealed a pattern consistent with slower updating in high-delusion patients (Fig. 3E). Conservative draws-to-decision behaviour could alternatively be due to lower decision thresholds, but the lack of differences in the final (pre-guess) probability estimates in the 60:40 condition revealed between the patient groups ($P = 0.89$; Fig. 3F) rendered this possibility unlikely.

To formally test impairments in belief updating, we first identified which belief-updating model best captured participants' behaviour using Bayesian model comparison ('Materials and methods' section and Supplementary material). This winning model (Fig. 4A–D) included a prior-weight parameter ω_1 representing primacy versus recency biases in prior beliefs and assumed that belief updating depended on the weighted strength of sensory evidence (modelled with one likelihood-weight parameter ω_2 per bead-ratio condition; see 'Materials and methods' section for definitions of the relevant terms and Supplementary material for a detailed mathematical description).

To test how alterations in belief updating may explain more conservative draws-to-decision behaviour in relation to delusions, we analysed whether either of the belief-updating parameters (ω_1 or ω_2, the latter specific to the 60:40 condition as this condition exhibited the strongest delusion-related effects on draws-to-decision) related to both draws-to-decision slopes and PDI delusion-severity scores in patients (Fig. 4E–G). Only the prior-weight ω_1 correlated with both draws-to-decision slopes ($r = 0.50$, $P = 0.013$) and PDI scores ($r = 0.46$, $P = 0.023$); the likelihood-weight $\omega_2(0.6)$ correlated only with draws-to-decision slopes ($r = -0.50$, $P = 0.013$; PDI: $r = -0.30$, $P = 0.16$). This suggests that deficient belief updating characterized by a primacy bias in prior beliefs leads to increased draws-to-decision in delusional patients. Neither ω_1 nor $\omega_2(0.6)$ correlated with the control variables in the specificity analyses above (all $P > 0.09$). Controlling for model goodness-of-fit (RMSE) did not meaningfully affect the delusion effects (see Supplementary material for more comprehensive control analyses). The PDI effects also held in analyses accounting for differences in draws-to-decision between high- and low-delusion patients (Supplementary material).

Post hoc analyses further showed that greater prior-weight ω_1 specifically related to clinical measures of delusion severity (beyond PDI scores), and to specific dimensions that differentiate delusions from non-pathological beliefs (Jones and Watson, 1997; Peters *et al.*, 1999) (Supplementary material).

Model-agnostic analyses of belief updating in patients

To provide converging support for our interpretation in terms of a delusion-related primacy bias in belief updating, we conducted a model-agnostic analysis to capture this bias using a time-lagged regression that predicted draw-by-draw probability estimates from the majority status of the previously observed beads ('Materials and methods' section). Unsurprisingly, individuals' beliefs about the identity of the hidden jar overall relied more on the last observed bead than on the previous ones (all $P < 0.0003$; Fig. 5A). Moreover, while this effect was robust in low-delusion patients (and controls), this disproportionate influence of recent versus early draws was less pronounced in high-delusion patients. Indeed, higher delusion severity correlated with less differential influence from the last observed bead relative to the bead presented two draws back (PDI scores and difference in regression coefficients for $[d] - [d-2]$: $r = -0.43$, $P = 0.034$; Fig. 5B) but not with the influence from beads observed in any given past draw (all $P > 0.1$). Overall, these model-agnostic results support our interpretation of the model-based results in terms of a delusion-related primacy bias by showing a relative bias towards information presented earlier in the trial in more delusional patients. Further control analyses were inconsistent with

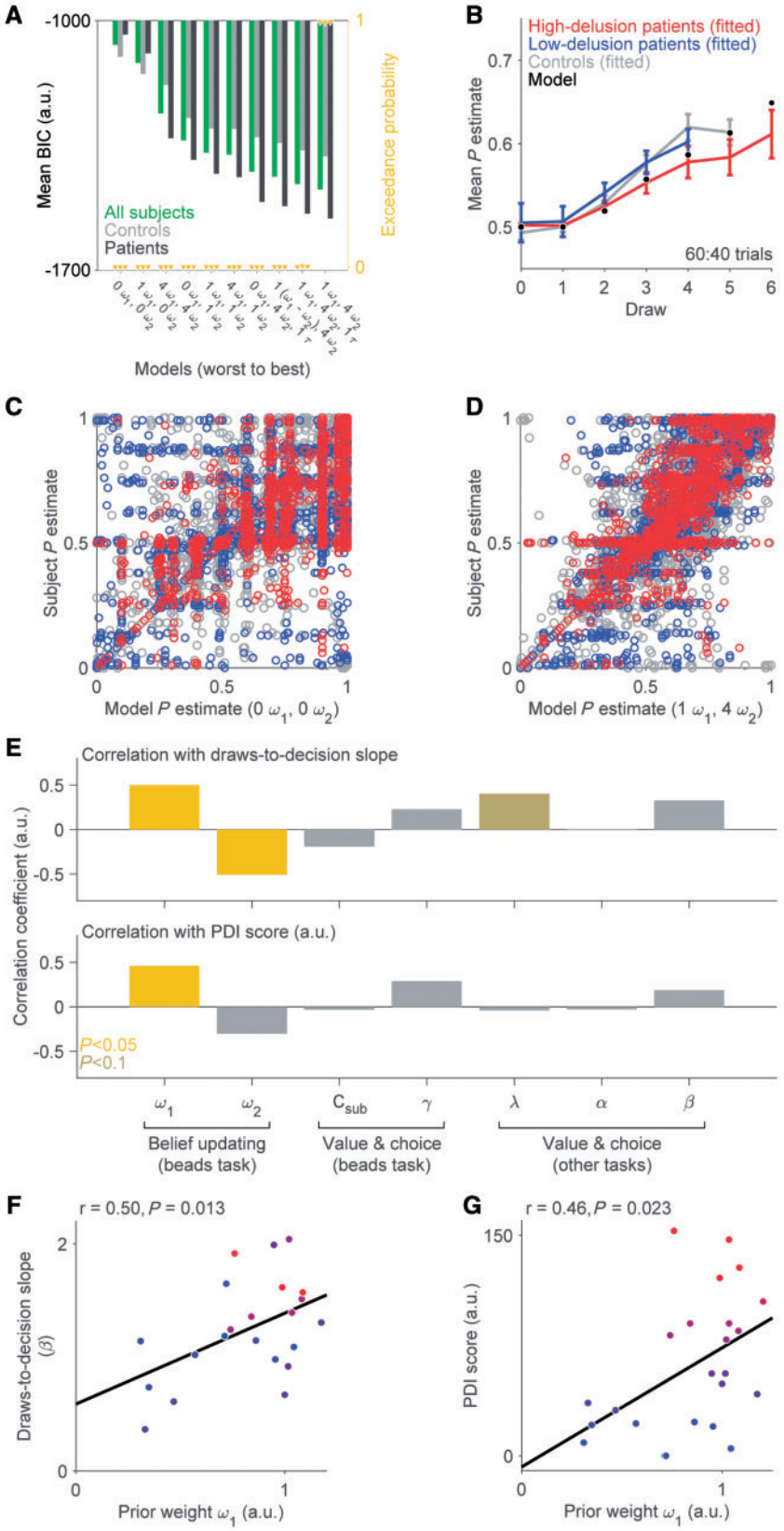

Figure 4 Model-based analyses of belief updating and relationship to delusion severity. (**A**) Mean BIC (left *y*-axis) and exceedance probability (right *y*-axis) values for competing belief-updating models, sorted by BIC (worst to best from left to right): the winning model for all patients, controls, and all groups combined is at the rightmost end of the x axis ('1ω_1, 4ω_2'). a.u. = arbitrary units. (**B**) Mean fitted probability estimates (for the actual hidden jar) before each draw based on individual fits of the winning model across all correct 60:40 trials (note that this represents the mean fits of the data in Fig. 3E). Error bars are SEM. (**C** and **D**) Scatterplots showing subject probability estimates (for the actual

(continued)

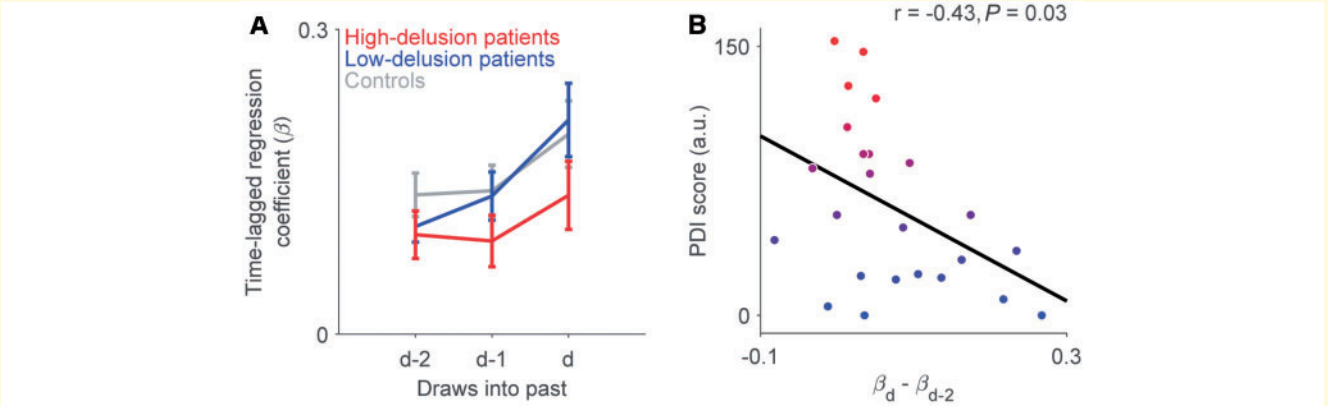

Figure 5 Model-agnostic analysis of belief updating and relationship to delusion severity. (**A**) Regression coefficients by group are shown from the time-lagged regression analysis predicting draw-wise probability estimates (for the actual hidden jar) from the bead colour (majority colour or not) in the current draw and the previous two draws. Error bars represent SEM. (**B**) Scatterplot depicting the relationship between the difference in regression coefficients from the current draw and two draws back ($\beta_d - \beta_{d-2}$) and delusional severity (PDI score). Note that we observed a significant correlation between this model-agnostic measure of primacy bias ($\beta_d - \beta_{d-2}$) and the model-derived measure of primacy bias (ω_1), but not between the model-agnostic measure of primacy bias and the other model-derived parameters from the winning belief-updating model (Supplementary material), indicating convergence between the model-agnostic and model-based analyses. Dots are coloured by delusional severity with greater severity indicated in red and lower severity in blue. a.u. = arbitrary units.

results being driven by group averaging across subsets of participants using grossly distinct strategies (Supplementary material).

Model-based and model-agnostic analyses of decision-making variables other than belief updating in patients

Finally, it is possible that non-inferential mechanisms could explain increased draws-to-decision in high-delusion patients, including abnormalities in valuation (e.g. in the perceived, subjective cost of drawing). To test this, we used a parameterized POMDP model of the beads task that included free parameters for the additional subjective cost of drawing and for choice stochasticity ('Materials and methods' section). Fitted draws-to-decision estimates using this model closely matched the raw behaviour ('Materials and methods' section) and recovered the main effects [Fig. 6A; fitted estimates of draws-to-decision slopes correlated with PDI scores in patients after excluding the three participants with non-significant fits (Supplementary material): r = 0.53, $P = 0.014$]. Critically, none of the fitted parameters related to subjective valuation or choice stochasticity correlated significantly with either delusion severity (all $P > 0.11$) or draws-to-decision slopes in patients (all $P > 0.20$), weighing against an interpretation of delusion-related draws-to-decision in terms of abnormal valuation or choice. Of note, the fitted draws-to-decision behaviour for a parameterized POMDP model using the mean subjective valuation and choice parameters across all subjects (rather than objective values and deterministic choice; Fig. 4E) was closest to behaviour in high-delusion patients, suggesting that departures from optimality in draws-to-decision behaviour in this group (Fig. 3A) can be mostly explained by subjective aspects of valuation that are common to all groups.

We also tested for delusion-related abnormalities in other decision-making processes that might explain increased draws-to-decision in our financially incentivized beads task, namely increased aversion to loss or uncertainty. To that end, we collected data with two

Figure 4 Continued

hidden jar) plotted against the predictions of the unweighted (parameter-free) Bayesian belief-updating model (**C**) and the weighted, winning belief-updating model (**D**). Colour scheme is the same as in **B**. (**E**) Correlations between model parameters (belief-updating [ω_1, ω_2(0.6)], value [C_{sub}(0.6)], and choice parameters [γ] from the beads task on the left, and value parameters from the control decision-making tasks on the right [loss aversion λ, risk aversion α, ambiguity aversion β]) and draws-to-decision slope (*top*) and between model parameters and delusional severity (PDI score; *bottom*). Note that λ, which reflects subjective valuation on the loss-aversion task, correlated with the subjective-valuation parameter of the beads task, C_{sub}(0.6), but not with other parameters of this task (Supplementary material). (**F** and **G**) Scatterplots indicating correlations between the prior weight ω_1 parameter and draws-to-decision slope (**F**) and between ω_1 and delusional severity (PDI score, **G**). Dots are coloured by delusional severity with greater severity indicated in red and lower severity in blue.

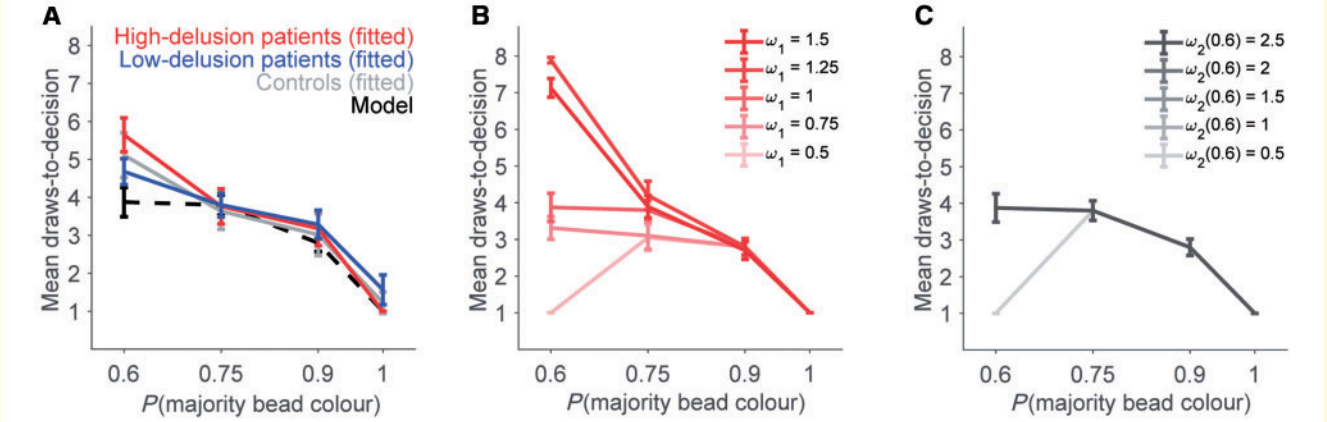

Figure 6 Fitted and simulated draws-to-decision behaviour. (**A**) Mean fitted draws-to-decision by group for each bead-ratio condition using the parameterized variant of the POMDP model (five free parameters for valuation and choice). The behaviour of the ideal-observer model is indicated by the black dashed line. (**B** and **C**) Mean draws-to-decision from a modified ideal-observer model with either varying the ω_1 (**B**) or varying the $\omega_2(0.6)$ (**C**) parameters (other aspects of the model are kept intact to simulate the specific effects of a single change in each of the relevant belief-updating parameters). Note that here, increased ω_1 leads to slower belief updating and consequently smaller estimates of the probability for the hidden jar at a given point within the trial (Fig. 3E); this results in smaller expected values for guessing relative to drawing and an increased tendency to draw. (**B** and **C**) The parameter values cover the approximate range of the individually fitted parameters in the patient data. Greater numerical values for the model parameters are indicated with darker colours. (**A**–**C**) Error bars represent SEM.

additional well-validated, incentive-compatible tasks: a loss-aversion task (Tom *et al.*, 2007) and a risk-and-ambiguity aversion task (Tymula *et al.*, 2013). Behaviour from both tasks was modelled using standard decision-making models (Supplementary material). The relevant loss, risk, and ambiguity aversion parameters derived from these tasks did not correlate with delusion severity (all $P > 0.38$) or draws-to-decision slopes (all $P > 0.06$) in patients (or differ between patients and controls; $P > 0.09$; Fig. 4E and Supplementary material). Furthermore, the relationship between the belief-updating parameter ω_1 and delusion severity held after controlling for subjective-valuation parameters (Supplementary material).

Simulations of selective 'lesions' in belief updating

To provide further support for our interpretation, we simulated draws-to-decision behaviour using a parameterized POMDP model in which we systematically manipulated the parameters of the belief-updating component while leaving the valuation and choice components intact (Fig. 6B and C). Results from these simulations, emulating a selective 'lesion' in belief updating in the context of otherwise optimal value-based decision-making, suggested that increases in the prior weight ω_1, but not in the likelihood-weight ω_2, are theoretically sufficient to explain increased draws-to-decision behaviour and qualitatively recapitulate our findings on delusions.

Discussion

Using a novel variant of the beads task designed to isolate inferential processes, we found abnormalities in information sampling (i.e. abnormal draws-to-decision behaviour) in schizophrenia that can be separated into (i) a specific effect related to delusions; and (ii) a presumably nonspecific effect related to socio-economic status. Patients with more severe delusions exhibited more conservative behaviour characterized by increased information seeking before guessing (increased draws-to-decision). This behaviour was specifically linked to delusions, beyond other symptoms of psychosis, medication, and other general cognitive and socioeconomic measures. By contrast, lower socio-economic status, typically seen in patients with schizophrenia and presumably reflecting an underlying general impairment or resulting from it, correlated with more liberal (reduced) draws-to-decision behaviour. This supports previous findings of reduced draws-to-decision in schizophrenia (Dudley *et al.*, 2016), the extreme of which has been labelled the 'jumping-to-conclusions' bias, while simultaneously suggesting that a specific cognitive mechanism associated with delusions may instead drive patients to behave more conservatively and exhibit increased draws-to-decision behaviour in certain contexts. We aimed to dissect the computational mechanism underlying this delusion-specific behaviour further. Increased draws-to-decision in delusional patients could potentially stem from baseline or sequentially acquired biases in prior beliefs, by abnormal updating of beliefs, or by alterations in value-based decision making. By analysing simultaneously collected draw-wise estimates reflecting how participants' beliefs about the

hidden jar evolved within a trial, we found evidence that specific deficits in belief updating, and not baseline or sequential biases or differences in value-based decision making, underlie delusion-related behaviour, consistent with our *a priori* hypothesis.

A key, novel finding in our study is the specific relationship between increased draws-to-decision behaviour and severity of delusional ideation in patients. Previous studies using the classic beads task have consistently found that patients with schizophrenia tend to exhibit decreased draws-to-decision compared to controls (Dudley *et al.*, 2016) but have failed to consistently find a specific correlation with delusion severity (Ross *et al.*, 2015; Dudley *et al.*, 2016). Using a novel approach to address this question that incentivizes participants to carefully consider each decision to draw or guess, and that minimizes impatience, we found that increased draws to decision correlated specifically with delusion severity, even after controlling for other cognitive and clinical variables. Further indicating the specificity of this effect, more severe delusions correlated with more accurate performance. Thus, the relationship between performance and delusions is unlikely to be explained by a general cognitive impairment in patients and may instead result from a cognitive alteration that benefits performance on this particular task. Also, because approximately half of our sample consisted of unmedicated patients and the delusion-specific effect was present within this subgroup alone, our results strongly suggest that antipsychotic medication was unlikely to induce the observed phenotype.

Despite the correlation between delusions and perceptual disturbances (both of them cardinal symptoms of psychosis), only the former correlated with increased draws-to-decision behaviour. We take this to suggest that, as initially postulated (Hemsley and Garety, 1986), impaired inference of hidden states specifically underlies delusional ideation. A plausible explanation for the common co-occurrence of these symptoms is that hallucination-specific and delusion-specific mechanisms are intertwined within a neural hierarchy of belief updating, in line with previous suggestions (Davies *et al.*, 2017; Sterzer *et al.*, 2018). While hallucination-specific mechanisms may ultimately depend on deficient updating of lower-level beliefs about the presence or absence of sensory stimuli (Horga *et al.*, 2014; Powers *et al.*, 2017), delusion-specific mechanisms may depend primarily on deficient updating of higher-level beliefs in the context of inference about hidden states.

Our modelling results further situate a circumscribed deficit in a belief-updating process, carrying out within-trial inference about hidden states, as a plausible (and theoretically sufficient) candidate for explaining delusion-related behaviours. This deficit was mathematically characterized as increased reliance on prior beliefs formed by incorporating information presented earlier in the trial (Supplementary material): a primacy bias in prior beliefs. Interestingly, such primacy bias in belief updating implies that delusion effects may manifest differently depending on the order in which information is presented, and could explain previously observed reasoning biases (Woodward *et al.*, 2007; McLean *et al.*, 2017). Other relevant decision-making variables, which might have impacted draws-to-decision behaviour, empirically did not relate to delusions. These non-inferential decision-making variables, in particular subjective valuation, could however provide a potential explanation for the group differences related to lower socio-economic status in patients, based on preliminary analyses, and perhaps for jumping-to-conclusions biases (Supplementary material). More generally, our results relating deficient belief updating to delusion severity fit better with the phenomenology of delusional beliefs than prior work suggesting jumping-to-conclusions biases in schizophrenia. This is because delusions are, by definition, rigid beliefs abnormally resistant—rather than susceptible– to updating in light of new information (Jaspers, 1913). Our paradigm thus affords a more specific and direct laboratory measure of belief updating that, at face value, seems better suited to capture clinically relevant features of delusions compared to previous paradigms.

Previous research used a retrospective revaluation procedure to test whether delusions correlated with abnormal neural responses to violation of expectations (Corlett *et al.*, 2007). Delusional individuals exhibited weaker prefrontal responses to violations, which were interpreted as deficient prediction-error signals. Those results suggest a relationship between neural measures of deficient belief updating and delusions, although the underlying causes of the weakened responses could not be fully interrogated. Other previous work in schizophrenia has used an approach similar to ours to test whether patients exhibited abnormal integration of prior knowledge and sensory information (Jardri *et al.*, 2017). This work showed evidence suggesting that psychotic patients 'over-counted' sensory information and under-weighted prior knowledge, in line with other findings showing immunity to low-level visual illusions in schizophrenia (Notredame *et al.*, 2014). However, one potential explanation for this finding could be that patients had a strong initial prior that hampered their ability to acquire a briefly presented visual cue—the experimental prior manipulation—and integrate it into confidence judgments later on. This distinction between acquisition and integration of the prior was a motivation of our current approach, which prompts a report of the baseline prior belief (before presentation of any sensory information in the trial) and tracks the within-trial dynamics of belief updating. Finally, previous investigations into belief updating and reinforcement learning in tasks with uncued reversals of hidden states have generally found increased adjustments in schizophrenia (Waltz and Gold, 2007; Reddy *et al.*, 2016; Adams *et al.*, 2018). However, these have not been linked consistently to delusions or other specific symptom domains.

Some limitations of this study are worth discussing. Whether the mechanisms we report here in relation to the fixity of delusions also relate to their formation is a question that warrants future work in prodromal stages of

psychosis. An alternative interpretation of our data would be that more paranoid patients behaved differently on the task because they were suspicious. Nonetheless, more delusional patients were more responsive to the task conditions, and non-paranoid ideation also correlated with the primacy bias, making this alternative interpretation less tenable (Supplementary material). Also, while our high-delusion patients were moderately severe, only more acute and extreme delusional states of generalized paranoia would typically interfere with behaviour in the research setting (particularly on a non-social, emotionally neutral task). While reconciling the large body of beads-task literature is beyond the scope of this work, future work should further examine the behavioural effects of real-life incentives (van der Leer and McKay, 2014) and state reversals, as well as of sequence ordering, which according to our model may result in opposite draws-to-decision behaviours under a single belief-updating alteration. Along these lines, our work suggests caution in using information-seeking behaviour as a specific proxy for inference (Supplementary material).

In conclusion, our results are consistent with a model in which deficient belief updating, characterized by stronger reliance on prior beliefs formed early in the inferential process, specifically underlies delusional ideation in schizophrenia, and drives increased information seeking in more delusional patients. This may explain the typical form of delusions as inflexible beliefs that remain unchanged upon presentation of new evidence. In contrast, we found some preliminary support for the notion that delusion-unrelated decreases in information seeking in schizophrenia could instead be driven by non-inferential processes. The specificity of the observed inferential abnormalities to delusional ideation suggests that the pathophysiological mechanisms underlying this symptom may be dissociable from those underlying other aspects of schizophrenia, and thus selectively targetable by novel therapies.

Acknowledgements

We would like to extend heartfelt thanks to Stephanie Todd for her assistance with this project.

Funding

This work was supported by grant K23-MH101637 (G.H.) as well as a pilot grant awarded by the Sydney R. Baer Jr. Foundation. Data used in the preparation of this manuscript (Supplementary material) were obtained and analysed from the controlled access datasets distributed from the NIMH-supported National Database for Clinical Trials (NDCT). NDCT is a collaborative informatics system created by the National Institute of Mental Health to provide a national resource to support and accelerate discovery related to clinical trial research in mental health. Dataset identifiers: N01 MH090001–06.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

Adams RA, Napier G, Roiser JP, Mathys C, Gilleen J. Attractor-like dynamics in belief updating in schizophrenia. J Neurosci 2018; 38: 9471–85.

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. Front Psychiatry 2013; 4: 47.

Ambuehl S, Li S. Belief updating and the demand for information. Games Econ Behav 2018; 109: 21–39.

Andreasen NC, Flaum M. Schizophrenia: the characteristic symptoms. Schizophr Bull 1991; 17: 27–49.

Averbeck BB. Theory of choice in bandit, information sampling and foraging tasks. PLoS Comput Biol 2015; 11: e1004164.

Balzan RP, Delfabbro PH, Galletly CA, Woodward TS. Over-adjustment or miscomprehension? A re-examination of the jumping to conclusions bias. Aust N Z J Psychiatry 2012; 46: 532–40.

Bell V, Halligan PW, Ellis HD. The Cardiff Anomalous Perceptions Scale (CAPS): a new validated measure of anomalous perceptual experience. Schizophr Bull 2006; 32: 366–77.

Bentall RP, Rowse G, Shryane N, Kinderman P, Howard R, Blackwood N, et al. The cognitive and affective structure of paranoid delusions: a transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. Arch Gen Psychiatry 2009; 66: 236–47.

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev 2006; 113: 700–65.

Brunton BW, Botvinick MM, Brody CD. Rats and humans can optimally accumulate evidence for decision-making. Science 2013; 340: 95–8.

Corcoran R, Rowse G, Moore R, Blackwood N, Kinderman P, Howard R, et al. A transdiagnostic investigation of 'theory of mind' and 'jumping to conclusions' in patients with persecutory delusions. Psychol Med 2008; 38: 1577–83.

Corlett PR, Murray GK, Honey GD, Aitken MR, Shanks DR, Robbins TW, et al. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. Brain 2007; 130 (Pt 9): 2387–400.

Davies DJ, Teufel C, Fletcher PC. Anomalous perceptions and beliefs are associated with shifts toward different types of prior knowledge in perceptual inference. Schizophr Bull 2017; 44: 1245–53.

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 2016; 42: 652–65.

Dudley RE, John CH, Young AW, Over DE. Normal and abnormal reasoning in people with delusions. Br J Clin Psychol 1997; 36 (Pt 2): 243–58.

Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 2009; 10: 48–58.

Freeman D, Startup H, Dunn G, Cernis E, Wingham G, Pugh K, et al. Understanding jumping to conclusions in patients with persecutory delusions: working memory and intolerance of uncertainty. Psychol Med 2014; 44: 3017–24.

Furl N, Averbeck BB. Parietal cortex and insula relate to evidence seeking relevant to reward-related decisions. J Neurosci 2011; 31: 17572–82.

Garety P. Reasoning and delusions. Br J Psychiatry Suppl 1991; 159: 14–8.

Gold JI, Shadlen MN. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron 2002; 36: 299–308.

Haddock G, McCarron J, Tarrier N, Faragher EB. Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). Psychol Med 1999; 29: 879–89.

Hemsley DR, Garety PA. The formation of maintenance of delusions: a Bayesian analysis. Br J Psychiatry 1986; 149: 51–6.

Hollingshead AB. Four factor index of social status. New Haven, CT: Yale University Department of Sociology; 1975.

Horga G, Schatz KC, Abi-Dargham A, Peterson BS. Deficits in predictive coding underlie hallucinations in schizophrenia. J Neurosci 2014; 34: 8072–82.

Huq SF, Garety PA, Hemsley DR. Probabilistic judgements in deluded and non-deluded subjects. Q J Exp Psychol A 1988; 40: 801–12.

Jacobsen P, Freeman D, Salkovskis P. Reasoning bias and belief conviction in obsessive-compulsive disorder and delusions: jumping to conclusions across disorders? Br J Clin Psychol 2012; 51: 84–99.

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

Jaspers K. Allgemeine psychopathologie. Berlin: J. Springer; 1913.

Jones E, Watson JP. Delusion, the overvalued idea and religious beliefs: a comparative analysis of their characteristics. Br J Psychiatry 1997; 170: 381–6.

Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artif Intell 1998; 101: 99–134.

Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987; 13: 261–76.

Langdon R, Ward PB, Coltheart M. Reasoning anomalies associated with delusions in schizophrenia. Schizophr Bull 2010; 36: 321–30.

Lincoln TM, Ziegler M, Mehl S, Rief W. The jumping to conclusions bias in delusions: specificity and changeability. J Abnorm Psychol 2010; 119: 40–9.

McLean BF, Mattiske JK, Balzan RP. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. Schizophr Bull 2017; 43: 344–54.

McLellan AT, Luborsky L, Woody GE, O'Brien CP. An improved diagnostic evaluation instrument for substance abuse patients. The Addiction Severity Index. J Nerv Ment Dis 1980; 168: 26–33.

Notredame CE, Pins D, Deneve S, Jardri R. What visual illusions teach us about schizophrenia. Front Integr Neurosci 2014; 8: 63.

Nuechterlein KH, Green MF, Kern RS, Baade LE, Barch DM, Cohen JD, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. Am J Psychiatry 2008; 165: 203–13.

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 1971; 9: 97–113.

Peters E, Joseph S, Day S, Garety P. Measuring delusional ideation: the 21-item Peters et al. Delusions Inventory (PDI). Schizophr Bull 2004; 30: 1005–22.

Peters ER, Joseph SA, Garety PA. Measurement of delusional ideation in the normal population: introducing the PDI (Peters et al. Delusions Inventory). Schizophr Bull 1999; 25: 553–76.

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

Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput 2008; 20: 873–922.

Reddy LF, Waltz JA, Green MF, Wynn JK, Horan WP. Probabilistic reversal learning in schizophrenia: stability of deficits and potential causal mechanisms. Schizophr Bull 2016; 42: 942–51.

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—revisited. Neuroimage 2014; 84: 971–85.

Ross RM, McKay R, Coltheart M, Langdon R. Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. Schizophr Bull 2015; 41: 1183–91.

So SH, Freeman D, Garety P. Impact of state anxiety on the jumping to conclusions delusion bias. Aust N Z J Psychiatry 2008; 42: 879–86.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage 2009; 46: 1004–17.

Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The predictive coding account of psychosis. Biol Psychiatry 2018; 84: 634–43.

Tom SM, Fox CR, Trepel C, Poldrack RA. The neural basis of loss aversion in decision-making under risk. Science 2007; 315: 515–8.

Tymula A, Rosenberg Belmaker LA, Ruderman L, Glimcher PW, Levy I. Like cognitive function, decision making across the life span shows profound age-related changes. Proc Natl Acad Sci USA 2013; 110: 17143–8.

Usher M, McClelland JL. The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 2001; 108: 550–92.

van der Leer L, McKay R. "Jumping to conclusions" in delusion-prone participants: an experimental economics approach. Cogn Neuropsychiatry 2014; 19: 257–67.

Waltz JA, Gold JM. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. Schizophr Res 2007; 93: 296–303.

White LO, Mansell W. Failing to ponder? Delusion-prone individuals rush to conclusions. Clin Psychol Psychother 2009; 16: 111–24.

Woodward TS, Buchy L, Moritz S, Liotti M. A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. Schizophr Bull 2007; 33: 1023–8.