American Psychological Association

Journal of Abnormal Psychology

© 2020 American Psychological Association
ISSN: 0021-843X

2020, Vol. 129, No. 6, 556–569
http://dx.doi.org/10.1037/abn0000500

Hierarchical Bayesian Models of Social Inference for Probing Persecutory Delusional Ideation

Andreea Oliviana Diaconescu
University of Zurich and ETH Zurich and University of Toronto

Katharina V. Wellstein and Lars Kasper
University of Zurich and ETH Zurich

Christoph Mathys
International School of Advanced Studies, Trieste, Italy, and Aarhus University

Klaas Enno Stephan
University of Zurich and ETH Zurich and Max Planck Institute for Metabolism Research, Cologne, Germany

While persecutory delusions (PDs) have been linked to fallacies of reasoning and social inference, computational characterizations of delusional tendencies are rare. Here, we examined 151 individuals from the general population on opposite ends of the PD spectrum (Paranoia Checklist [PCL]). Participants made trial-wise predictions in a probabilistic lottery, guided by advice from a more informed human and a nonsocial cue. Additionally, 2 frames differentially emphasized causes of invalid advice: (a) the adviser's possible intentions (dispositional frame) or (b) the rules of the game (situational frame). We applied computational modeling to examine possible reasons for group differences in behavior. Comparing different models, we found that a hierarchical Bayesian model (hierarchical Gaussian filter) explained participants' responses better than other learning models. Model parameters determining participants' belief updates about the adviser's fidelity and the contribution of prior beliefs about fidelity to trial-wise decisions, respectively, showed significant Group × Frame interactions: High PCL scorers held more rigid beliefs about the adviser's fidelity across both experimental frames and relied less on advice in situational frames than low scorers. These results suggest that PD tendencies are associated with rigid beliefs and prevent adaptive use of social information in "safe" contexts. This supports previous proposals of a link between PD and aberrant social inference.

General Scientific Summary
Persecutory delusions—unfounded beliefs that others deliberately intend to harm—are core psychosis symptoms. This study examines computational alterations in inference about others' changing intentions in relation to subclinical persecutory ideation in the general population. It examines learning in a volatile, social context under two different frames developed to probe rigid beliefs about others' intentions in a large sample of prescreened participants with high or low persecutory delusions.

Keywords: psychosis, hierarchical Bayesian inference, hierarchical Gaussian filter, computational psychiatry, Bayesian model selection

Supplemental materials: http://dx.doi.org/10.1037/abn0000500.supp

Persecutory delusions (PDs) are understood as an agent's seemingly unfounded beliefs that others are acting deliberately to cause harm. These beliefs are by definition persistent despite disconfirming evidence (Freeman, 2007). Holding these beliefs has been

Andreea Oliviana Diaconescu, Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, and Centre for Addiction and Mental Health (CAMH), University of Toronto; Katharina V. Wellstein and Lars Kasper, Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich; Christoph Mathys, International School of Advanced Studies, Trieste, Italy, and Interacting Minds Centre, Aarhus University; Klaas Enno Stephan, Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, and Max Planck Institute for Metabolism Research, Cologne, Germany.

This study was approved by the Ethics Committee of the Canton of Zurich, Switzerland (KEK-ZH-Req-2016-00236).

The authors acknowledge support by the SNF Ambizione PZ00P3_167952 and Krembil Foundation (Andreea Oliviana Diaconescu) and the René and Susanne Braginsky Foundation (Klaas Enno Stephan).

Andreea Oliviana Diaconescu and Katharina V. Wellstein contributed equally to this work.

Correspondence concerning this article should be addressed to Andreea Oliviana Diaconescu, Centre for Addiction and Mental Health (CAMH), University of Toronto, 250 College Street, 12th Floor Toronto, ON, M5T 1R8. E-mail: Andreea.Diaconescu@camh.ca

associated with fallacies of reasoning, such as a propensity to jump to conclusions in probabilistic reasoning tasks (Fine, Gardner, Craigie, & Gold, 2007; So et al., 2012; Young & Bentall, 1997), and theory of mind deficits (Bentall et al., 2009; White, Borgan, Ralley, & Shergill, 2016), including an impoverished ability to predict others' mental states (Corcoran, Mercer, & Frith, 1995; C. D. Frith & Corcoran, 1996; R. C. Frith, 1996). Despite existing cognitive models (Blackwood, Howard, Bentall, & Murray, 2001; Freeman & Garety, 2014; Freeman, Garety, Kuipers, Fowler, & Bebbington, 2002), formal characterizations of persecutory ideation, in terms of computational processes such as inference and belief updating, are rare.

One leading computational account of delusions derives from Bayesian theories of brain function (Adams, Stephan, Brown, Frith, & Friston, 2013; Corlett, Taylor, Wang, Fletcher, & Krystal, 2010; Fletcher & Frith, 2009; Sterzer et al., 2018). This Bayesian view on delusions mainly refers to predictive coding and proposes that the brain infers on the causes of its sensations using a hierarchically structured model of the external world (Friston, 2005; Rao & Ballard, 1999). This generative model, which describes how sensory inputs are probabilistically generated by hidden states of the world, provides top-down predictions about sensory inputs that are updated by experience via prediction errors (PEs; Doya, Ishii, Pouget, & Rao, 2011; Lee & Mumford, 2003).

In this hierarchical framework, beliefs are formalized as probability distributions, and the influence of PEs on higher-level beliefs depends on their weight or, more specifically, on their precision (inverse variance). The critical aspect of precision weighting is the precision assigned to sensory inputs relative to higher-level (prior) beliefs. For example, in a generic and widely used hierarchical Bayesian formulation—the hierarchical Gaussian filter (HGF; Mathys, Daunizeau, Friston, & Stephan, 2011; Mathys et al., 2014)—the trial-wise updating of beliefs at any level of the hierarchy by PEs depends on their weighting by a ratio of sensory to prior precision. The intuition behind this weighting is that beliefs should be updated more readily the more precise the sensory information and the less precise (or certain) the prior belief. In other words, agents who are highly confident in their predictions are less likely to update their model of the world in the face of contradictory evidence.

From a Bayesian perspective, delusions can be understood as deficits of hierarchical inference (Corlett et al., 2010; Fletcher & Frith, 2009), with precision weighting playing a crucial role (Adams et al., 2013; Corlett, Honey, & Fletcher, 2016; Sterzer et al., 2018). More specifically, delusions have been proposed to arise from increased sensory precision of low-level PEs, rendering these PEs abnormally salient and leading to a chronic surprise about sensory inputs. In this condition, the formation of highly precise high-level beliefs may represent a compensatory response that is required to "explain away" the low-level PE signals (Fletcher & Frith, 2009).

Several recent studies have provided empirical evidence for this enhanced influence of higher-level prior beliefs across the psychosis spectrum. For instance, Teufel and colleagues (Teufel et al., 2015) have shown that the level of reliance on prior expectations during low-level visual perception was positively correlated with subclinical levels of psychotic symptoms. Similarly, precise prior beliefs were shown to govern belief updating processes in delusion-prone individuals (Schmack et al., 2013) and in individuals who reported hearing voices compared to those who did not, irrespective of psychosis diagnosis (Powers, Mathys, & Corlett, 2017).

In persecutory delusions, beliefs about the intentions of others play a central role (Biedermann, Frajo-Apor, & Hofer, 2012). Social information is by its nature ambiguous as human intentions are concealed and have to be inferred. Additionally, it implies a hierarchical form built from low-level features with increasing degrees of abstraction. In this highly uncertain context, the role of precision in belief updating is particularly critical (Diaconescu et al., 2014).

In the current study, we examined social inference (i.e., inference about the intentions of others) in subclinical persecutory delusion. We analyzed 151 participants who scored on opposite ends of Freeman's Paranoia Checklist (PCL; Freeman et al., 2005) and performed a probabilistic advice-taking task in a social context. We examined the role of precision in the belief updating process by manipulating (a) changes in the association strength between advice and the outcome (i.e., volatility) and (b) the social context by instructing participants about the task under one of two experimental frames.

The frames differentially emphasized possible causes of misleading advice, either drawing participants' attention to the adviser's intentions (dispositional frame) or the rules of the game (situational frame). Our design was 2 × 2 factorial with between-subjects factors "persecutory delusional tendencies" (high vs. low) and "frame" (dispositional vs. situational). In conventional analyses of the behavioral data using analysis of variance (Wellstein et al., 2019), we found significant Group × Frame interactions in advice-taking, which suggested that individuals with high persecutory delusional tendencies (high PD group) failed to incorporate the experimental frames into their learning about advice. Specifically, high PCL scorers did not exhibit differences in their advice-taking behavior as a function of the framing, whereas low PCL scorers (low PD group) did, taking advice into account less under the dispositional frame (which emphasized misleading advice as potentially related to the adviser's hidden intentions) compared to the situational frame. Furthermore, in a task-specific debriefing questionnaire, we found that the high PD group expressed more distrust regarding the adviser and attributed incorrect advice more to the adviser compared to the low PD group.

In this article, we extend our previous analyses using computational modeling and Bayesian model selection. In line with the Bayesian theories of delusion described above, we hypothesized that high PCL scorers might exhibit overly precise higher-level beliefs about the adviser's fidelity. This would explain the lack of difference between the two frames in the high-delusion group that we observed in our previous analysis (Wellstein et al., 2019), as rigid higher-level beliefs about others' intentions would prevent adaptive belief-updating by social information. An alternative explanation might be that high PCL scorers are less sensitive to social information because they hold overly negative prior beliefs about the adviser and therefore predict the adviser's intentions to be more misleading in general, as compared to low PCL scorers. To disentangle these two mechanisms of abnormal inference, we fit several computational models to participants' trial-by-trial decisions and compared their ability to explain the data, using Bayesian model selection. Our model space included simple reinforcement learning models as well as a set of hierarchical and nonhierarchical Bayesian models, which emphasized the role of precision-weighting in the belief-updating process.

Method

With the notable exception of the computational modeling approach, most of the methods used in this study have been described in a previous publication (Wellstein et al., 2019). In order to keep the article self-contained and easily readable, we include a brief description of the participant sample and the task in this article.

Ethics Statement

All experimental participants gave written informed consent before the study, which was approved by the Ethics Committee of the Canton of Zurich (KEK-ZH-Req-2016-00236).

Experimental Design

The study used a factorial between-subjects design with two participant groups and two experimental conditions (experimental frames). Group assignment was based on average PCL scores across three different time points (for details, see Wellstein et al., 2019).

First, 1,145 individuals from the general population were prescreened with an online questionnaire, in order to assign them into either the high PD group or the low PD group. We used the items of the German version of the PCL (Freeman et al., 2005) intermixed with distractor items Neuroticism-Extraversion-Openness Five-Factor Inventory (NEO-FFI; McCrae & Costa, 2004). The PCL is a self-report questionnaire assessing paranoid thoughts using a multidimensional approach to paranoid ideation. It consists of Frequency, Conviction, and Distress subscales. The German version of the PCL questionnaires has been validated by Lincoln, Peter, Schäfer, and Moritz (2009). The inclusion criteria for filling out the prescreening questionnaires were as follows: (a) age 18 or older, (b) fluent German, and (c) absence of ongoing psychological or psychiatric treatment. The group assignment was based on the PCL Frequency subscale, which indicates how frequently participants were thinking paranoid thoughts.

Participants were assigned to the high PD group if they scored 0.5 standard deviations above the mean Frequency score reported by Freeman et al. (2005), which corresponds to a cumulative score above 16. Participants scoring 0.5 standard deviations below mean (i.e., a sum score of below 3) were assigned to the low PD group.

In order to ensure that participants' group assignment was based on a trait-like construct rather than a temporary expression of paranoia, participants initially assigned to one of the two groups were invited to fill out another questionnaire 4 weeks later. This was the case for 340 participants of the prescreening sample. The second questionnaire contained the same items as the first one but intermixed in another sequence. Of these 340 participants, only 162 participants whose scores were consistent with the previous group assignment were invited to the study.

On experiment day, they performed the advice-taking task, underwent a brief cognitive assessment, and filled out a task-specific debriefing questionnaire and additional questionnaires on the computer, including the PCL.

Sample. Sample sizes were based on an a priori power analysis, for a power of 0.8 under a moderate effect size of Cohen's f = 0.25 (based on results by Diaconescu et al., 2014, obtained under the same task). Under a Type I error of 5%, the required sample size for the analyses of Group × Condition interactions across the four participant groups was $N = 146$. Assuming a dropout/exclusion rate of approximately 10%, we invited 162 participants: 78 volunteers were part of the high PD group, of whom 8 were excluded from the analyses based on exclusion criteria that had been predefined in a timestamped analysis plan (see below and Wellstein et al., 2019, for details). Eighty-one volunteers were included in the low PD group, resulting in an overall sample size of $N = 151$ eligible for analyses. All cells (groups and conditions) were balanced regarding age, education, and proportion of male versus female. Note that cells remained balanced after excluding participants from analyses.

Experimental procedures. Following informed consent, participants received standardized instructions on paper and were instructed to write a short summary of the task in their own words. This served to ensure that they had correctly understood the task. Afterward, they performed a practice round (eight trials), in which they were truthfully informed that advice validity was fixed to chance.

After performing the experimental task (duration 40 min; for details, see below), participants filled out a task-specific debriefing questionnaire. In the last phase of the experiment, a cognitive screening was administered consisting of two subtests of the Brief Cognitive Assessment (Fervaha et al., 2015) in order to control for the potential influence of cognitive deficits (including working memory) on task performance (Ventura, Wood, & Hellemann, 2013).

Advice-taking task. The task used in this study is a modified version of Behrens's established advice-taking task (Behrens, Hunt, Woolrich, & Rushworth, 2008) and has been used in a similar form in several studies by Diaconescu et al. (2014, 2017). In brief, participants played a probabilistic binary lottery, where they had to predict the outcome of a color draw (blue or green) trial by trial. They had access to two sources of information: (a) a nonsocial cue, which was represented by a pie chart indicating what color was more likely to win on any given trial, and (b) advice from a more informed agent, represented by a videotaped adviser who gave a recommendation on which color to choose by holding up a card (blue or green). The participants had to make decisions by integrating the two sources of information. Motivation to perform well was provided by silver and gold targets, which were associated with an additional bonus (Swiss Francs 10 or Swiss Francs 20, respectively).

Videos of the adviser were recorded from face-to-face interactions (Diaconescu et al., 2014) and were presented alongside the nonsocial cue for 2 s. In order for participants to stay alert and make their decisions intuitively, they had 5 s to press the button for blue or green after the presentation of the two cues. After every decision, participants received feedback on their choice and on the correct outcome (see Figure 1).

In order to perform well, participants had to infer not only the accuracy of current advice but also the adviser's intention and how it might change over time (volatility). To examine the impact of belief precision on learning from advice, we manipulated volatility and thereby varied the association strength between the advice and the outcome. We predicted that the higher-level belief precision about the adviser's fidelity is low when volatility is high and vice versa.

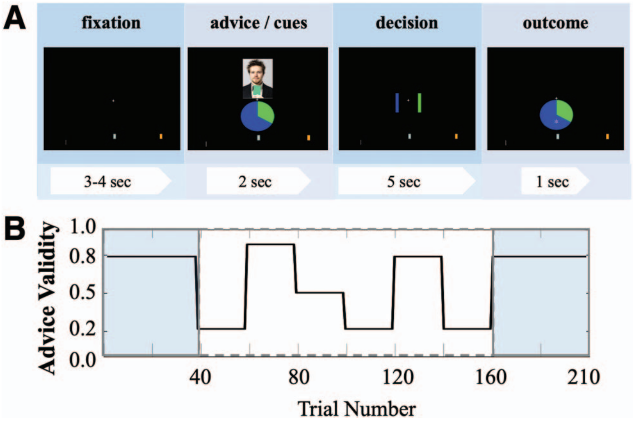

Figure 1. Advice-taking experimental paradigm: Participants predicted the outcome of a binary lottery (blue vs. green) based on a social and a nonsocial cue (advice and pie chart) presented simultaneously. Players accumulated points with every correct prediction. If the cumulative score exceeded the silver or gold target, players earned an additional bonus on top of the reimbursement paid out for participating in the study. After predicting what color would "win," they were informed about the real outcome. Pie chart probabilities varied between 50:50, 55:45, and 65:35. Advice validity was varied across 210 trials as indicated by the boxcar chart. The shaded areas above the input structure highlight the stable (blue) and volatile (white) phases of the task. Adapted with permission from "Inflexible social inference in individuals with subclinical persecutory delusional tendencies," by K. V. Wellstein, A. O. Diaconescu, M. Bischof, A. Rüesch, G. Paolini, E. A. Aponte, J. Ullriche, and K. E. Stephanagh, advanced online publication September 5, 2019, *Schizophrenia Research* (http://dx.doi.org/10.1016/j.schres.2019.08.031). CC BY-NC-ND.

Experimental framings. We used two experimental framing conditions, which differed in how potentially misleading advice was framed (dispositional vs. situational). This allowed us to probe whether the participants' inferences on the causes of play outcomes would be modulated by the frame and whether this modulation depended on the magnitude of persecutory ideation tendencies (i.e., Group × Frame interaction). Critically, neither of the frames provided false information but simply described the adviser's role from different perspectives. In the dispositional frame, participants' attention was directed to the adviser as a potential source of variability in advice validity and emphasized his or her ability to act intentionally in order to achieve his or her own (unknown) goals. In the situational frame, attention was directed to the role of the adviser as part of the task, highlighting that he or she was instructed to use the imperfect information available to him or her for guiding the player's behavior. We induced the two frames over three different channels: (a) one sentence in the instructions that differed between the two frames, (b) a reminder on the start screen of the task, and (c) the wording used regarding advice validity ("correct" and "incorrect" in situational frame vs. "helpful" and "misleading" in the dispositional frame). For more details on how the framing was induced, please see Wellstein et al. (2019).

Computational Modeling

Our computational modeling approach was guided by the general idea that participants use a generative model of trial outcomes (i.e., correct or incorrect advice) to infer on both the advice validity and the adviser's change in intentions (volatility). In accordance with previous studies of advice-taking under volatility (Diaconescu et al., 2014, 2017), we considered six learning models (referred to as "perceptual models"; Figure 2): (a) the HGF (Mathys et al., 2011, 2014) in a classical three-level formulation; (b) the HGF with a constant drift parameter; (c) a mean-reverting HGF; (d) a nonvolatility, two-level HGF (Diaconescu et al., 2014); (e) a reinforcement learning model with an adaptive learning rate (Sutton, 1988); and (f) a Rescorla–Wagner (RW) model of associative learning (Rescorla & Wagner, 1972).

For parameter estimation, we used the meta-Bayesian framework by Daunizeau et al. (Daunizeau, den Ouden, Pessiglione, Kiebel, Friston, et al., 2010, Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan, et al., 2010). This requires a response model that describes the probabilistic link from hidden beliefs to observable responses or outcomes. Here, we considered three possible forward mappings from beliefs to decisions based on previous studies (Diaconescu et al., 2014, 2017). These response models represent different mechanisms of how participants incorporate social and/or nonsocial sources of information to make decisions.

We accounted for individual differences in the inference process by employing random-effects Bayesian model selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009), a procedure that treats the model as a random variable in the population and allows for estimating which proportion of the population is best described by each of the models considered.

Generally, the models used in this study were chosen such that their parameter estimates could be used to test predictions from theories that conceptualize the emergence and persistence of delusional beliefs in terms of aberrant inference, described above (see Corlett et al., 2010; Fletcher & Frith, 2009; Sterzer et al., 2018). That is, we expected that individual delusional tendencies (as assessed by questionnaires) would be associated with individual parameter estimates that describe how flexibly or rigidly be-

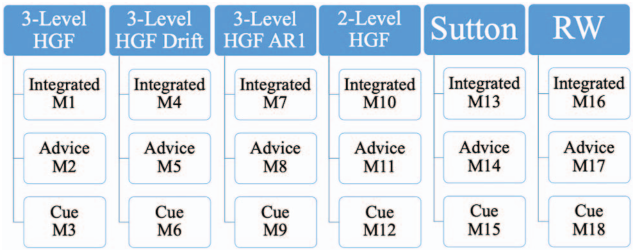

Figure 2. Model space: perceptual models, response models. The models considered in this study have a 6 × 3 factorial structure (six perceptual models paired with three response models). Each box represents an individual model of learning in which both or one of the two social and nonsocial sources of information are considered. The nodes at the top represent the perceptual model families (hierarchical Gaussian filter [HGF], HGF with drift, mean-reverting HGF [AR1], nonvolatility, two-level HGF, Sutton, and Rescorla–Wagner [RW] models). Three response models were formalized according to the weighing of social and nonsocial information. These models propose that participants' beliefs are based on (a) both cue and advice information (Integrated: Cue and Advice) and (b) advice only (Advice) or (c) cue only, that is, that only the given cue probabilities (i.e., the pie chart) enter the belief-to-response mapping (Cue).

liefs about the intentions of others behave when experiencing helpful and misleading advice in an ambiguous context.

Perceptual models. The HGF is a hierarchical model of learning under perceptual uncertainty and environmental volatility, which allows for inference on participants' beliefs and belief precisions about states in the world (i.e., in the current study, the validity of advice) from their observed behavior (see Mathys et al., 2011, 2014). In brief, the HGF assumes that an agent infers on a hierarchy of hidden states $x_1^{(k)}$, $x_2^{(k)}$, . . . , $x_n^{(k)}$, which cause the sensory inputs he or she experiences on each trial k. In the HGF, these states evolve in time as Gaussian random walks where, at any given level, the step size is controlled by the state at the next higher level.

Here, at the lowest level, x_1 represents the advice accuracy (i.e., a single presentation of advice is either accurate ($x_1^{(k)} = 1$) or inaccurate ($x_1^{(k)} = 0$)). x_1 is a probabilistic function of x_2 via the logistic sigmoid transformation s(·) (Equations 1–2).

$$p(x_1 \mid x_2) = s(x_2)^{x_1}(1 - s(x_2))^{1-x_1} = \text{Bernoulli}\ (x_1; s(x_2)) \qquad (1)$$

where

$$s(x) \overset{\text{def}}{=} \frac{1}{1 + \exp(-x)}. \qquad (2)$$

All states higher than x_1 are continuous. State x_2 represents the adviser's tendency to offer helpful advice (i.e., the adviser's fidelity) in logit space, whereas the highest state x_3 reflects how quickly the intentions of the adviser (x_2) are changing, that is, log volatility (Equations 3–4).

$$p\left(x_2^{(k)} \mid x_2^{(k-1)},\ x_3^{(k)},\ \kappa, \omega_2\right) = \mathcal{N}\left(x_2^{(k)}; x_2^{(k-1)}, \exp(\kappa x_3^{(k)} + \omega_2)\right), \qquad (3)$$

$$p\left(x_3^{(k)} \mid x_3^{(k-1)},\ \omega_3\right) = \mathcal{N}\left(x_3^{(k)}; x_3^{(k-1)}, \exp(\omega_3)\right). \qquad (4)$$

The evolution of these states is determined by three subject-specific parameters: First, κ determines the extent to which the second level x_2 is coupled to the third level x_3. In the context of this study, it represents the degree to which a participant utilizes his or her estimate of volatility (the adviser's changing intentions) to infer on the current advice validity. Second, ω_2 is the evolution rate, and it represents the tonic component of the log volatility at the second level, which is independent of the phasic influence by the volatility component x_3. In other words, it captures the subject-specific magnitude of belief updates about the adviser's fidelity that is independent of the volatility of the adviser's intentions. Finally, ω_3 (metavolatility) determines the evolution rate of x_3, determining the variance of the adviser's changing intentions (see Figure 3 for a graphical representation of the HGF and Table 1 and Table 2 for the priors over parameters).

Model inversion: The update equations. The HGF assumes that agents estimate the adviser's fidelity based on hierarchically coupled states in a trial-by-trial fashion by employing an efficient variational approximation to ideal Bayesian inference (see Mathys et al., 2011, for details). The update equations that emerge from this approximation have a simple and interpretable form comparable to reinforcement learning models, except for featuring a dynamic learning rate that is determined by the next higher level in the hierarchy.

At each hierarchical level i, belief updates (posterior means) $\mu_i^{(k)}$ on each trial k are proportional to precision-weighted PEs (Equation 5). In essence, the belief update is proportional to the product of the PE from the level below $\delta_{i-1}^{(k)}$, weighted by a precision ratio:

$$\Delta\mu_i^{(k)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}}\,\delta_{i-1}^{(k)} \qquad (5)$$

where $\hat{\pi}_{i-1}^{(k)}$ and $\pi_i^{(k)}$ represent estimates of the precision of the prediction about input from the level below (i.e., precision of the data) and of the belief at the current level, respectively.

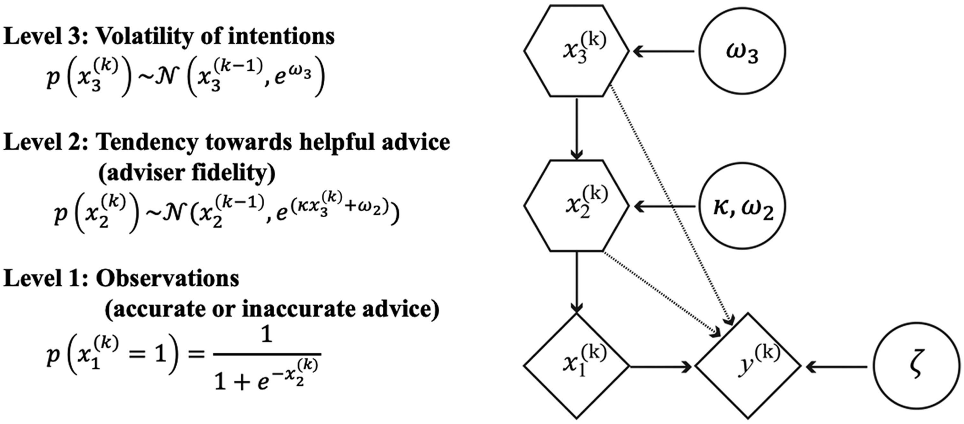
Level 3: Volatility of intentions
$p\left(x_3^{(k)}\right) \sim \mathcal{N}\left(x_3^{(k-1)}, e^{\omega_3}\right)$
Level 2: Tendency towards helpful advice (adviser fidelity)
$p\left(x_2^{(k)}\right) \sim \mathcal{N}(x_2^{(k-1)}, e^{(\kappa x_3^{(k)} + \omega_2)})$
Level 1: Observations (accurate or inaccurate advice)
$p\left(x_1^{(k)} = 1\right) = \frac{1}{1 + e^{-x_2^{(k)}}}$
$x_3^{(k)}$
ω_3
$x_2^{(k)}$
κ, ω_2
$x_1^{(k)}$
$y^{(k)}$
ζ

Figure 3. Graphical representation of the hierarchical Gaussian filter. In this graphical notation, circles represent constants and diamonds represent quantities that change in time (i.e., that carry a time/trial index). Hexagons, like diamonds, represent quantities, which change in time, but additionally depend on the previous state in time in a Markovian fashion. x_1 represents the accuracy of the current piece of advice, x_2 the adviser's fidelity or tendency to give helpful advice, and x_2 the current volatility of the adviser's intentions. Parameter κ determines how strongly x_2 and x_3 are coupled, ω_2 determines the tonic volatility component, and ω_3 represents the volatility of x_3. The response model has two layers: (a) the probability of the outcome given both the nonsocial cue and the advice and (b) the chosen action, drawn from the integrated belief using a sigmoid decision rule. Parameter ζ determines the weight of the advice compared to the nonsocial cue. y represents the subject's binary response ($y = 1$: deciding to accept the advice, $y = 0$: deciding against the advice).

Table 1
Prior Mean and Variance of the Perceptual Model Parameters

Parameter	Prior mean	Prior variance
(i) HGF $M_1, \ldots, M_3$		
κ	1	1
ω_2	−3	16
ω_3	−6	16
$\mu_2^{(k=0)}$	0	1
$\mu_3^{(k=0)}$	1	1
(ii) HGF with Drift $M_4, \ldots, M_6$		
κ	1	1
ω_2	−3	16
ω_3	−6	16
ρ_2	0	1
$\mu_2^{(k=0)}$	0	1
$\mu_3^{(k=0)}$	1	1
(iii) Mean-reverting HGF $M_7, \ldots, M_9$		
κ	1	1
ω_2	−3	16
ω_3	−6	16
m_2	0	1
$\mu_2^{(k=0)}$	0	1
$\mu_3^{(k=0)}$	1	1
(iv) HGF $M_{10}, \ldots, M_{12}$		
κ	1	0
ω_2	−3	16
ω_3	−20	0
$\mu_2^{(k=0)}$	0	1
$\mu_3^{(k=0)}$	1	0
(v) Sutton $M_{13}, \ldots, M_{15}$		
μ	1	100
$v^{(k=0)}$	0	16
(vi) Rescorla–Wagner $M_{16}, \ldots, M_{18}$		
α	0.25	1
$v^{(k=0)}$	0.5	1

Note. The prior variances are given in the space in which parameters are estimated. α, $\mu_2^{(k=0)}$, $\mu_3^{(k=0)}$, $v^{(k=0)}$ are estimated in logit-space, while κ and μ (Sutton model) are estimated in log-space. HGF = hierarchical Gaussian filter.

To examine the role of prior beliefs about the adviser's fidelity, in addition to the learning parameters κ, ω_2, and ω_3, we also estimated the initial values of the estimated adviser fidelity or $\mu_2^{(k)}$, that is, $\mu_2^{(k=0)}$.

Competing Models

The first model (Model 1) is similar to the winning model from previous studies by Diaconescu et al. (Diaconescu et al., 2014, 2017) where the same advice-taking task was used (see Tables 1 and 2 for the priors used). The first competing model (Model 2) differed from the standard HGF by means of introducing a constant drift parameter—ρ_2 at the second level (see Equation 6). This constant drift parameter can be interpreted as a hostility bias, serving to shift the belief trajectory toward more negative predictions about the adviser's fidelity. While its prior is zero, reflecting unbiased beliefs, the sign and direction of this parameter could be estimated from participants' choices.

$$p\left(x_2^{(k)} \mid x_2^{(k-1)},\ x_3^{(k)},\ \kappa, \omega_2,\ \rho_2\right) \\ = \mathcal{N}\left(x_2^{(k)};\ x_2^{(k-1)} + t^{(k)}\rho_2, \exp\left(\kappa x_3^{(k)} + \omega_2\right)\right), \tag{6}$$

where $t^{(k)}$ refers to the trial number.

The second competing model (Model 3) considers the possibility that rigid beliefs about the adviser's fidelity are adequately represented by "anchoring" them to an attractor state. In this case, the generative model represents state x_2 (adviser fidelity) as experiencing a drift toward an equilibrium point, m_2. This formulation is captured by a "mean-reverting HGF" in which the evolution of x_2 (the adviser's fidelity) is not only determined by the learning parameters κ and ω_2 but also by an additional parameter ϕ_2, which defines how quickly x_2 drifts toward an equilibrium value m_2.

$$p\left(x_2^{(k)} \mid x_2^{(k-1)},\ x_3^{(k)},\ \kappa, \omega_2,\ m_2,\ \phi_2\right) \\ = \mathcal{N}\left(x_2^{(k)};\ x_2^{(k-1)} + \phi_2\left(m_2 - x_2^{(k-1)}\right), \exp\left(\kappa x_3^{(k)} + \omega_2\right)\right). \tag{7}$$

Intuitively speaking, "rigidity" in this model corresponds to a tendency of holding a belief about adviser fidelity that is relatively immune to the experience of helpful or misleading advice.

The third competing model is a nonhierarchical Bayesian model or "two-level HGF" (Diaconescu et al., 2014), where only the evolution rate parameter ω_2 determines the magnitude of the belief update about the adviser's fidelity.

Finally, the last two competing models are both reinforcement learning models. The "Sutton" model is equipped with an adaptive learning rate (Sutton, 1988) while the RW model (Rescorla & Wagner, 1972) has a fixed learning rate parameter α.

Response models. The response models (see Figure 2) describe how the participant's beliefs are transformed into decisions.

Table 2
Prior Mean and Variance of the Response Model Parameters

Variable	Parameter	Prior mean	Prior variance
Integrated model M_1, M_4, M_7, M_{10}, M_{13}, M_{16}	ζ	0.5	100
Reduced: advice M_2, M_5, M_8, M_{11}, M_{14}, M_{17}	ζ	∞	0
Reduced: cue M_3, M_6, M_9, M_{12}, M_{15}, M_{18}	ζ	$-\infty$	0

Common parameters

Parameter	Prior mean	Prior variance
β	48	16

Note. The prior variances are given in the space in which parameters are estimated. ζ is estimated in logit-space, while β is estimated in log-space.

During the task, participants could either integrate social and nonsocial sources of information or use either source of information exclusively. On each trial, the visual pie chart indicated the true prior probability $\tilde{c}$ about the outcome, whereas the social information corresponded to the participant's current belief that the adviser would give correct advice, that is, $\hat{\mu}_1^{(k)}$. Agents who considered only one of the two sources of information could base their decisions on these quantities directly. In the more complex case of integrating cue and advice, participants were assumed to base their decisions on a weighted average of the two sources of information: Introducing a parameter ζ that represents the weight of the advice (a value in the unit interval: $\zeta \in [0, 1]$), the expected outcome probability is represented by

$$b^{(k)} = \zeta\, \hat{\mu}_1^{(k)} + (1 - \zeta)\tilde{c}^{(k)}. \tag{8}$$

The probability that the participant followed the advice (i.e., response $y = 1$, as opposed to $y = 0$ when going against the advice) was described by a sigmoid function, which maps the unit interval [0, 1] onto itself (note that this function differs from the logistic sigmoid above, which maps the whole real line onto the unit interval).

$$p(y^{(k)} = 1 \mid b^{(k)}) = \frac{b^{(k)\beta}}{b^{(k)^{\beta}} + (1 - b^{(k)})^{\beta}}. \tag{9}$$

Parameter β in Equation 9 represents the inverse decision temperature. A low decision temperature (high β) means always choosing the highest probability option, whereas a high decision temperature (low β) means random sampling from a uniform distribution.

Based on previous work (see Diaconescu et al., 2014, 2017), the belief-to-response mapping was assumed to vary trial-by-trial with a decision noise parameter (i.e., β that was estimated for each participant) and the predicted adviser volatility, $\exp(-\mu_3^{(k-1)})$. In other words, as the estimated volatility of the adviser's intentions decreases, the sigmoid function becomes steeper and participants act more according to their estimates of advice validity. By contrast, when the volatility increases, participants become more uncertain about the adviser's intentions and behave in a more exploratory fashion.

Maximum a posteriori (MAP) estimates of model parameters were obtained using the HGF toolbox Version 5.1 (http://www.translationalneuromodeling.org/tapas). Furthermore, we used family-level inference (Penny et al., 2010) to determine (a) the most likely class of perceptual models, combining across all response models, and (b) the most likely class of response models, combining across all perceptual models.

Hypotheses. The hypotheses and analysis procedures were specified in an analysis before the start of data analysis. The analysis plan was time-stamped and stored on a GitLab at the Translational Neuromodeling Unit. Both the analysis plan and the code for analyzing the data can be accessed at https://gitlab.ethz.ch/sibak.

Most of the hypotheses specified in the previous analysis plan concern statistical analyses of questionnaire and debriefing data (the associated results are reported in Wellstein et al., 2019). Two hypotheses (I and II) are of relevance for the model-based analyses reported in this article. Hypothesis II, however, has become irrelevant since it was conditional on a specific model emerging as superior from model comparison; our model selection procedure, however, found a different model to provide the best explanation of the behavioral data. This article is therefore restricted to testing those predictions that derive from Hypothesis I in the analysis plan: This general hypothesis states that participants in the high PD group take situational information less into account than participants in the low PD group. As specified in the analysis plan, this hypothesis gives rise to three concrete predictions that can be tested by using the subject-specific MAP estimates of model parameters estimates as inputs for a two-factorial analysis of variance (ANOVA), with factors group (high vs. low PCL scorers) and condition (situational vs. dispositional experimental frame).

First, assuming that one of the HGF-based models would be selected as the winning model across both groups, we expected to find that the initial prior beliefs about adviser fidelity (i.e., the estimates of $\mu_2^{(k=0)}$) would differ between groups but in a way that depended on the experimental frame. In other words, we expected to find an interaction of group and condition with regard to $\mu_2^{(k=0)}$.

Second, again assuming that one of the HGF-based models would be selected as the winning model across both groups, the rigidity of beliefs would be captured by the evolution rate parameter ω_2, which represents the tonic component of the log-volatility at the second level (which represents adviser fidelity). We predicted a main effect of group—that is, participants in the high PD group compared to the low PD group were expected to exhibit lower (more negative) values of ω_2 and thus lower learning rates (slower updating of beliefs about adviser fidelity) as a reflection of their inflexible beliefs.

Third, assuming that a response model with a weighted mixture of social and nonsocial information would be found superior in the model selection procedure, we expected that individuals in the high PD group would make less use of information provided by social advice, as compared to participants in the low PD group. This is because social information is ambiguous and may not provide useful information for an individual who expects advice to be misleading in general. Therefore, we expected high PD participants to show a reduced weight of the social advice (response model parameter ζ) in both conditions but particularly in the situational compared to the dispositional frame. Put differently, we expected to find both a significant main effect of group and a Group × Condition interaction.

Results

Bayesian Model Selection

Bayesian model selection indicated that both groups were characterized best by the same model. The winning model for both groups together was the classical HGF with volatility-influenced decision noise, including the social weighing parameter ζ (Model 1 of Figure 2; posterior probability across the two groups, $p(r \mid y) = 0.5646$ and protected exceedance probability: PXP = 1; Table 3). The same winning model was obtained when testing the groups separately ($p(r \mid y) = 0.5791$ and PXP = 1 for the low PD group and $p(r \mid y) = 0.5785$ and PXP = 1 for the high PD group).

Family-level inference across all perceptual model classes demonstrated that the HGF perceptual model family had highest model evidence (PXP = 1, Table 4). Family selection across all response model classes showed that the family comprising models with the

Table 3
Results of Bayesian Model Selection: Posterior Model Probabilities or p(r | y)

Variable	HGF	HGF with drift	Mean-reverting HGF (AR1)	Two-level HGF	Sutton	RW
Integrated	0.5646	0.1486	0.0695	0.0574	0.0207	0.0175
Advice	0.0312	0.0060	0.0063	0.0071	0.0059	0.0059
Cue	0.0193	0.0067	0.0083	0.0083	0.0083	0.0083

Note. HGF = hierarchical Gaussian filter; RW = Rescorla–Wagner.

social weighing factor ζ as a free parameter outperformed the other response models (PXP = 1, Table 5). This suggests that participants integrated social and nonsocial information instead of relying exclusively on one of the two sources.

Model Parameter Comparison

We extracted the parameters of the winning model (the classical HGF with volatility-influenced decision noise; see Table 6 for descriptive statistics) and examined the aforementioned hypotheses by performing a two-way ANOVA with an interaction term (group, framing condition, and Group × Framing condition) on the MAP estimates for $\mu_2^{(k=0)}$, ω_2, and ζ.

Concerning the first prediction described above, no significant main effects or interactions were observed with regard to the initial prior belief about the adviser's fidelity, $\mu_2^{(k=0)}$ (group: $df = (1, 150)$, $F = 0.28$, $p = .59$; frame: $df = (1, 150)$, $F = 1.56$, $p = .21$; interaction: $df = (1, 150)$, $F = 1.16$, $p = .28$).

With regard to the second prediction, we did not find the hypothesized main effect of group ($df = (1, 150)$, $F = 0.11$, $p = .74$). We did, however, observe a significant Group × Frame interaction for the evolution rate parameter ω_2 (Figure 4a; interaction: $df = (1, 150)$, $F = 4.75$, $p = .03$), with lower and less differential values across both experimental frames for the high PD group compared to the low PD group. Furthermore, we observed a main effect of frame, with overall larger belief updates for the dispositional frame compared to the situational frame ($df = (1, 150)$, $F = 6.12$, $p = .01$).

Third, concerning the hypotheses about the social bias parameter ζ, we found a significant Group × Frame interaction, as predicted ($df = (1, 150)$, $F = 6.58$, $p = .01$). This interaction suggests that individuals in the high PD group showed less differences in how they took advice into account across the framing conditions (Figure 4b). The form of the interaction differed from our expectations, however, in that ζ showed relatively similar values across framing conditions in the high PD group. Concerning main effects, we failed to find the predicted main effect of group ($df = (1, 150)$, $F = 0.76$, $p = .38$). Instead, we found a significant main effect of frame, reflecting reduced weighting of social advice in the dispositional frame compared to the situational frame ($df = (1, 150)$, $F = 10.49$, $p = .001$).

After testing our prespecified hypotheses and finding Group × Frame interactions for learning and decision-making parameters, we performed additional (exploratory) analyses to see how these interactions were reflected by the evolution of beliefs on the different hierarchical levels of the model. For parts of these analyses, we distinguished three contexts of learning from advice: (a) the first stable phase, when advice was correct with $p = .8$; (b) the volatile phase; and (iii) the second stable phase, when advice was correct with $p = .8$ (see Figure 1).

First, we extracted individual belief precision trajectories and computed a subject-specific average across each phase. Second, we performed a mixed-factor ANOVA with between-subjects and within-subject factors in order to investigate Frame × Group × Phase interactions.

We found that the Group × Frame interaction described for the learning parameter ω_2 described above was reflected by the estimated belief precision about the adviser's fidelity. In addition to a significant Group × Frame interaction for the belief precision π_2 ($df = (1, 147)$, $F = 7.05$, $p = .008$), we also found a significant main effect for frame (frame: $df = (1, 147)$, $F = 4.93$, $p = .02$). Whereas low PCL scorers showed significantly reduced belief precision in the dispositional compared to the situational frame, high PCL scorers showed no differences across frames. In general, the precision of the prediction about the adviser fidelity was reduced for the dispositional frame compared to the situational frame, presumably due to the attribution bias that was established by the frame (i.e., seeing the adviser's intentions as a primary cause of incorrect advice).

Furthermore, there was a significant impact of volatility on belief precision, with individuals showing more uncertainty and thus a decrease in the precision of predictions after volatility, even when the advice becomes stable and helpful, as in the initial phase of the task (main effect of phase: $df = (2, 294)$, $F = 221.82$, $p = 4.2745e\text{-}32$), but no two-way significant interactions (Phase ×

Table 4
Family-Level Inference (Perceptual Model Set): Posterior Model Probability or p(r | y) and Protected Exceedance Probabilities (pxp)

Variable	HGF	HGF with drift	Mean-reverting HGF (AR1)	Two-level HGF	Sutton	RW
p(*r* | *y*)	0.7127	0.0095	0.1244	0.0696	0.0741	0.0095
pxp	1	0	0	0	0	0

Note. HGF = hierarchical Gaussian filter; RW = Rescorla–Wagner.

Table 5
Family-Level Inference (Response Model Set): Posterior Model Probability or $p(r \mid y)$ and Protected Exceedance Probabilities (pxp)

Variable	Integrated	Reduced: advice	Reduced: cue
$p(r \mid y)$	0.9763	0.0124	0.0112
pxp	1	0	0

Condition: $df = (2, 294)$, $F = 3.73$, $p = .053$; Phase × Group: $df = (2, 294)$, $F = 1.49$, $p = .22$). Importantly, however, the three-way interaction of Group × Frame × Phase was significant ($df = (2, 294)$, $F = 3.98$, $p = .04$). This result suggests that the impact of volatility was stronger in the situational compared to the dispositional frame in the low PD group, whereas in the high PD group, volatility had a similar impact across the two experimental frames (see Figure 5).

Although we expected that belief precision would increase again in the second period of stability, no differences between the volatile phase and the second stable phase were observed. This was because the second stable phase was not long enough for the belief precision to return to the original (prevolatility) values. The second stable phase would have needed to be a lot longer than 42 trials, which would have led to a substantially longer experiment.

Discussion

The current study aimed at providing a computational characterization of subclinical persecutory ideation with respect to hierarchical inference about others' intentions. To this end, individuals who scored reliably on either extreme of the PCL were invited to take part in the behavioral experiment, which consisted of playing a social advice-taking task under volatility.

With this task, we probed how participants inferred on the intentions of an expert adviser in order to decide whether to follow his or her recommendations when predicting the outcome of a binary lottery (represented by a pie chart) for monetary rewards. Additionally, we induced volatility by manipulating the advisers' strategy and association strength between the advice and the outcome. This allowed us to examine how participants' learning rate adapted in the face of increasing uncertainty, due to the changing intentions of the adviser. Finally, we varied the context by introducing two experimental frames that emphasized different causes for incorrect advice: (a) dispositional, focusing on the adviser as having his or her own (hidden) intentions, and (b) situational, focusing on the task structure.

Applying several computational models of behavior to participants' responses, we found that all participants, irrespective of group assignment, used the same model—a classical three-level version of the HGF—to learn about intentions. Furthermore, participants integrated both sources of information—the advice and the nonsocial cue (pie chart)—to predict the outcome of the lottery. This replicates findings from previous studies that used the same paradigm (Diaconescu et al., 2014, 2017).

Subclinical Persecutory Ideation Is Associated With Less Contextual Influence on Belief Updating

Using the model parameter estimates, we tested several hypotheses (that had been prespecified in an analysis plan) about differences in social inference across the two groups. As described in the Results section, several of these hypotheses were not supported by our analysis results. However, some of the predictions were confirmed, and additional interesting findings emerged from systematic ANOVA applied to the model parameter estimates. For example, as predicted, we found a significant Group × Frame interaction for the social bias parameter ζ, which represents how much weight an individual assigned to the advice when taking a decision. Specifically, this interaction demonstrated that individuals in the high PD group showed less differences in advice-taking across the framing conditions than participants from the low PD group (Figure 4b).

A similar Group × Frame interaction was detected for evolution rate parameter ω_2, the subject-specific tonic log volatility estimate that determines the learning rate for updating beliefs about the adviser's fidelity. Low PCL scorers exhibited higher evolution rates in the dispositional compared to the situational frame, suggesting that they updated their beliefs about the adviser's fidelity more rapidly in the condition when the adviser was emphasized as the potential source of incorrect advice. By contrast, high PCL scorers exhibited lower and comparable evolution rates across the two experimental frames. This suggests that their belief-updating process was similar across framing conditions and—consistent with the Group × Frame interaction for the response model parameter ζ—that they made less use of social context when learning from advice (Figure 4a).

These interaction effects on the evolution parameter ω_2 should be expressed as a lack of differences in second-level belief precision in the high PCL group, in contrast to the low PCL scorers where a differential effect of the framing conditions on belief precision should be visible. In additional analyses, we verified this by comparing the average of trial-wise belief precision estimates across the different phases of the task (see Figure 5). This figure provides a complementary illustration that high PCL scorers were more resistant to changing their beliefs and learning about the adviser's intentions than low scorers.

We also examined the impact of volatility on belief precision across groups and frames and found a main effect of phase and a Group × Frame × Phase interaction (see Figure 5), suggesting that belief precision estimates decreased following periods of increas-

Table 6
Average Maximum a Posteriori Estimates of the Learning and Decision-Making Parameters of the Winning Model

Variable	*M*	*SD*
Perceptual model parameters		
κ	0.96	0.46
ω_2	−6.76	3.03
ω_3	−5.71	1.35
$\mu_2^{(k=0)}$	0.48	0.41
$\mu_3^{(k=0)}$	0.86	0.47
Response model parameters		
ζ	0.49	0.23
β	8.15	8.00

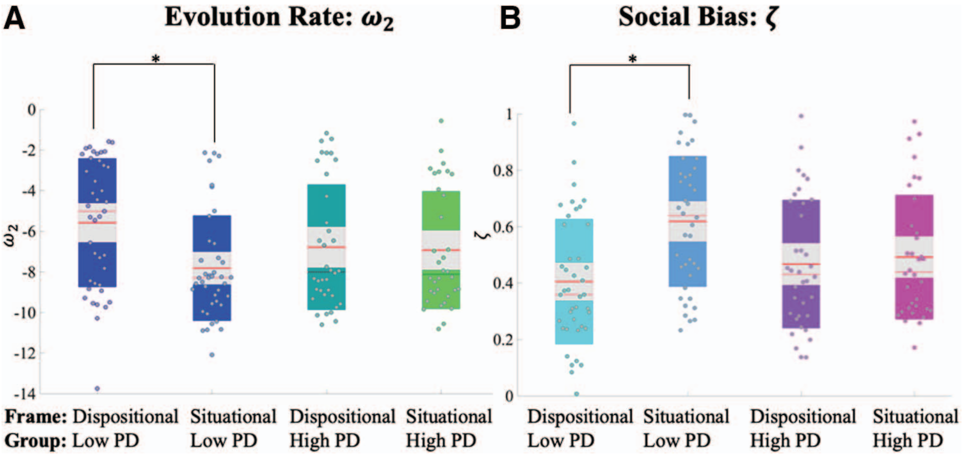

Figure 4. Maximum a posteriori estimates for both groups and conditions: (A) ω_2, the parameter representing tonic log-volatility, showed a significant Group × Frame interaction and a significant main effect of frame. (B) ζ, the response model parameter reflecting the weight of social information, also showed a significant Group × Frame interaction and a significant main effect of frame. The interaction suggests significant differences between the two conditions in the low persecutory delusion (PD) group, which were absent in the high PD group. See main text for details. Jittered raw data are plotted for each perceptual model parameter. The red line refers to the mean, the colored background reflects the 95% confidence intervals for the mean, and the gray background refers to 1 standard deviation of the mean. * $p < .05$ is indicated to emphasize the Group × Frame interactions.

ing volatility but that volatility had a stronger impact on belief precision in the situational frame compared to the dispositional frame in low PCL scorers. In other words, volatility-induced increases in uncertainty were larger for the unbiased task, where the social information was deemed to be "safe." A reduced sensitivity to social context was observed again in high PCL scorers, in whom volatility showed a similar impact on belief precision across the two experimental frames.

Subclinical Persecutory Ideation and Advice Discounting

Less variable belief precision estimates across experimental frames may be interpreted as high PD participants having a model of the adviser's intentions that is less susceptible to contextual information (i.e., the frame). The same interpretation is suggested by an additional finding, that is, a reduced influence of experi-

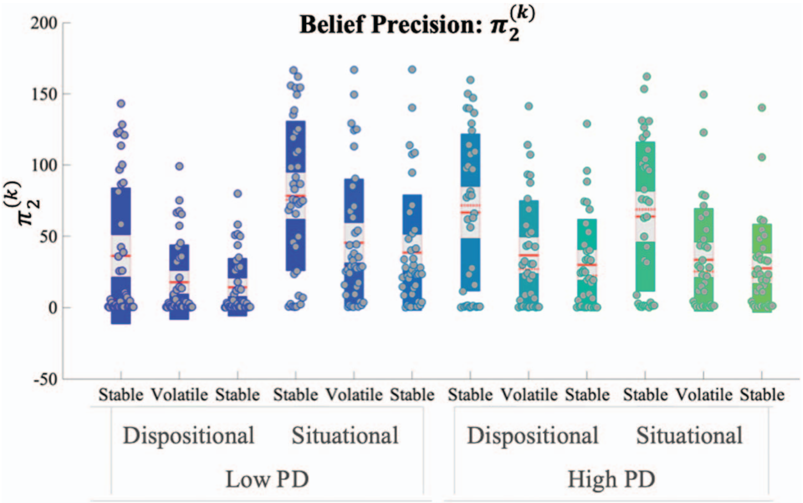

Figure 5. Average estimates of belief precision (π_2) for task phases, groups, and framing conditions. A mixed-factor analysis of variance (group, frame, and phase) found significant main effects of phase and significant Group × Frame × Phase interactions. See main text for details. Jittered raw data are plotted for each perceptual model parameter. The red line refers to the mean, the colored background reflects the 95% confidence intervals for the mean, and the gray background refers to 1 standard deviation of the mean. PD = persecutory delusion.

mental framing in the high versus low PCL group with regard to how much participants took advice into account, the significant Group × Frame interaction for the social bias parameter ζ. As is visible in Figure 4, low PCL scorers adapted more flexibly to context more when deciding to take the advice into account and therefore relied less on advice in the dispositional frame compared to the situational frame (Figure 4b).

In our study, these group differences are unlikely to be explained by learning deficits per se, since high and low PCL scorers did not differ in cognitive performance as assessed by cognitive screening. Also, as reported in more detail in a separate study, the groups did not differ in terms of performance accuracy or with regard to the amount of monetary rewards they gained during the task (Wellstein et al., 2019).

Furthermore, the results of the current study do not suggest that high PCL scorers are more likely to jump to conclusions about the adviser's fidelity. Jumping to conclusions represents a cognitive bias (arising from altered learning or decision-making) that has previously been suggested to be associated with delusions (Ermakova et al., 2019; Fine et al., 2007; Moutoussis, Bentall, El-Deredy, & Dayan, 2011; So et al., 2012; Speechley, Whitman, & Woodward, 2010). However, these results were obtained using very different (nonsocial) cognitive tasks lacking volatility. Instead, in the task used in this study, high PCL scorers exhibited increased belief precision and a propensity to go against the adviser's suggestions from the beginning of the interaction, across both periods of stability and volatility.

Clinical Implications

Recent computational efforts to capture the jumping-to-conclusions bias in psychosis suggested explanations in terms of aberrant inference (and an asymmetric mapping of beliefs to probabilities) as well as increased response stochasticity (Adams, Napier, Roiser, Mathys, & Gilleen, 2018). We examined a similar mechanism by including a mean-reverting HGF model in our model space. This model suggests that beliefs about adviser fidelity drift dynamically toward a constant level, which can be seen to represent the rigidity of socially relevant beliefs of an agent with persecutory ideation. However, this model did not capture participants' decisions in the current task as well as the classical HGF. This may be because we tested individuals with subclinical persecutory ideation and not psychosis, who showed stable paranoid beliefs about others. Finally, in the current study, we did not observe any evidence of increased response stochasticity in the high PCL scorers, as indexed by decision noise parameter β.

The results of the current study do, however, support the general idea that delusions can be conceptualized as beliefs with overly high precision. Adams et al. (2013) suggested that abnormally high belief precision may represent a compensatory response and function to attenuate unpredicted sensory inputs. While the results of this study do not provide any direct evidence for this notion, they are at least compatible with this idea. Notably, however, our study examined individuals without a clinical diagnosis who did not suffer from clinically significant delusions but merely exhibited proneness to persecutory ideation.

The influence of prior beliefs in delusion-prone individuals and psychosis patients has been debated. Although the utilization of prior knowledge correlated with positive symptom severity in psychosis patients in a perceptual discrimination task, the study also reported decreased impact of experimentally induced priors (Schmack, Rothkirch, Priller, & Sterzer, 2017). Positive symptom severity was also associated with an overcounting of sensory evidence in a probabilistic decision-making paradigm (Jardri, Duverne, Litvinova, & Denève, 2017). Furthermore, a recent study found that delusion-prone individuals showed a reduced influence of experimental priors in perceptual but not cognitive discrimination tasks (Stuke, Weilnhammer, Sterzer, & Schmack, 2019). These somewhat inconsistent results might possibly be reconciled by considering the distinct impact of sensory compared to belief precision on bottom-up PE signals (Adams et al., 2013; Sterzer et al., 2018).

The results of the current study may contribute to the long-term goal of developing computational assays, which can identify different stages of psychosis, particularly in the prodromal phase, under a dimensional perspective on psychosis as a continuum (van Os et al., 1999). In this study, the participants scored highly on the PCL and reported having persecutory thoughts frequently, consistently over three time points. These subclinical tendencies toward delusional ideation were associated with enhanced expression of higher-level belief precision about the adviser's fidelity, which leads to reduced learning from PEs. Although our participants did not suffer from clinically relevant persecutory delusions, our computational analysis suggests that their learning style was clearly distinct from individuals in the low PCL group who did not preoccupy themselves with persecutory thoughts at all.

In the future, this computational approach could be extended to examining not only delusion persistence but also formation. The prodromal phase of psychosis, which has been associated with an enhanced aberrant salience (Kapur, 2003; Shaner, 1999) or increased bottom-up PE signaling, might be characterized by a reduced precision of higher-level beliefs or, alternatively, increased precision of low-level PEs. It has been speculated by previous authors that the rigidity of high-level beliefs in fully developed psychosis represents a compensatory response to these putative initial processes (Adams et al., 2013; Corlett et al., 2010).

Limitations

It is important to note that our results do not explicitly support models of paranoia that rely specifically on abnormalities in social inference, as opposed to inferential abnormalities that may affect complex cognitive processes in general, including social cognition. In fact, impairments in learning about environmental uncertainty (cf. Kaplan et al., 2016) may lead to deficits in higher-level inferential processes, which are not specifically social but nonetheless impair learning about intentions (Ermakova et al., 2019).

To examine this issue in more detail for our paradigm, a control task with environmental volatility but without intentionality is needed. We have previously used such a control task in the context of a related paradigm, but using two interacting humans instead of videos (Diaconescu et al., 2014). This control task had the same statistical structure, volatility, and near-identical visual stimuli as our social learning task but used blindfolded advisers who selected their advice from predefined decks of cards, thereby eliminating the impact of intentionality. The order of the card decks was defined to match the incentive structure of the adviser. In this condition, we found that the three-level HGF best explained participants' responses in the social task and in the control task, but trial-wise volatility estimates were determining the mapping from beliefs to responses only in social task. Future studies will be needed to clarify whether the maladaptive inferential processes identified here are specific to the social domain, that is, tracking the adviser's changing intentions.

Another limitation of the current study is that not all parameters of our winning model could be recovered well (online supplemental Figures S2 and S4). Although the simulations based on the empirical data led to superior parameter recovery, the estimated coupling and metavolatility parameters were closely distributed around their prior mean. This is a general challenge for paradigms based on binary outcomes, where many trials and more complex volatility structures may be needed to sufficiently inform the estimation of parameters at higher levels in the hierarchy. Having said this, in the current study, the input structure does allow for recovery of the parameters we had predicted to show differences between low and high PD groups, such as $\mu_2^{(k=0)}$, ω_2, and ζ, which showed good recovery in both sets of simulations we conducted (online supplemental Figures S2 and S4).

Conclusion and Future Directions

Since the computational approach employed here casts testable hypotheses regarding the mechanism of delusion formation and persistence in psychosis, it may provide a useful starting point for the development of tools that predict transition to psychosis in clinical high-risk state (CHR) individuals. The CHR refers to the presence of one or more of the following criteria: attenuated psychotic symptoms, brief limited intermittent psychotic episodes, trait vulnerability (including family history), and a clear decline in psychosocial functioning. Although clinical measures have good prognostic accuracy for determining who will not develop psychosis, there is a need to increase the prediction accuracy of future transition to psychosis (Schmidt et al., 2017). A general strategy from neuromodeling and computational psychiatry is to use generative models for providing mechanistically interpretable quantities in individual patients that may inform predictions about disease trajectories and treatment response (Stephan et al., 2017). Model-based parameter estimates, which capture the individual belief-updating process when learning about intentions, may prove useful for predicting the future transition to psychosis in CHR individuals.

Mechanistically interpretable computational models such as the ones compared here could potentially enable inference on disease mechanisms in individual patients across the stages of psychosis. Furthermore, the computational quantities derived from the model—such as the belief precision or the precision-weighted PEs—may be associated with distinct neuromodulatory systems, such as dopamine or acetylcholine (Diaconescu et al., 2017; Iglesias et al., 2013), which are ultimately the targets of pharmacological treatment in psychosis. Future studies with prospective designs will be needed to examine the usefulness of this approach for predicting treatment response in individual patients.

References

Adams, R. A., Napier, G., Roiser, J. P., Mathys, C., & Gilleen, J. (2018). Attractor-like dynamics in belief updating in schizophrenia. *Journal of Neuroscience, 38,* 9471–9485. http://dx.doi.org/10.1523/JNEUROSCI.3163-17.2018

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry, 4,* 47. http://dx.doi.org/10.3389/fpsyt.2013.00047

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456,* 245–249. http://dx.doi.org/10.1038/nature07538

Bentall, R. P., Rowse, G., Shryane, N., Kinderman, P., Howard, R., Blackwood, N., . . . Corcoran, R. (2009). The cognitive and affective structure of paranoid delusions: A transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. *Archives of General Psychiatry, 66,* 236–247. http://dx.doi.org/10.1001/archgenpsychiatry.2009.1

Biedermann, F., Frajo-Apor, B., & Hofer, A. (2012). Theory of mind and its relevance in schizophrenia. *Current Opinion in Psychiatry, 25,* 71–75. http://dx.doi.org/10.1097/YCO.0b013e3283503624

Blackwood, N. J., Howard, R. J., Bentall, R. P., & Murray, R. M. (2001). Cognitive neuropsychiatric models of persecutory delusions. *American Journal of Psychiatry, 158,* 527–539. http://dx.doi.org/10.1176/appi.ajp.158.4.527

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research, 17,* 5–13. http://dx.doi.org/10.1016/0920-9964(95)00024-G

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology, 30,* 1145–1155. http://dx.doi.org/10.1177/0269881116650087

Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology, 92,* 345–369. http://dx.doi.org/10.1016/j.pneurobio.2010.06.007

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Friston, K. J., & Stephan, K. E. (2010). Observing the observer (II): Deciding when to decide. *PLoS ONE, 5,* e15555. http://dx.doi.org/10.1371/journal.pone.0015555

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): Meta-Bayesian models of learning and decision-making. *PLoS ONE, 5,* e15554. http://dx.doi.org/10.1371/journal.pone.0015554

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., . . . Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology, 10,* e1003810. http://dx.doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience, 12,* 618–634. http://dx.doi.org/10.1093/scan/nsw171

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (2011). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Ermakova, A. O., Gileadi, N., Knolle, F., Justicia, A., Anderson, R., Fletcher, P. C., . . . Murray, G. K. (2019). Cost evaluation during decision-making in patients at early stages of psychosis. *Comprehensive Psychiatry, 3,* 18–39. http://dx.doi.org/10.1162/cpsy_a_00020

Fervaha, G., Hill, C., Agid, O., Takeuchi, H., Foussias, G., Siddiqui, I., . . . Remington, G. (2015). Examination of the validity of the Brief Neurocognitive Assessment (BNA) for schizophrenia. *Schizophrenia Research, 166,* 304–309.

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry, 12,* 46–77. http://dx.doi.org/10.1080/13546800600750597

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10,* 48–58. http://dx.doi.org/10.1038/nrn2536

Freeman, D. (2007). Suspicious minds: The psychology of persecutory delusions. *Clinical Psychology Review, 27,* 425–457. http://dx.doi.org/10.1016/j.cpr.2006.10.004

Freeman, D., & Garety, P. (2014). Advances in understanding and treating persecutory delusions: A review. *Social Psychiatry and Psychiatric Epidemiology, 49,* 1179–1189. http://dx.doi.org/10.1007/s00127-014-0928-7

Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Fowler, D., . . . Dunn, G. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *British Journal of Psychiatry, 186,* 427–435. http://dx.doi.org/10.1192/bjp.186.5.427

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology, 41,* 331–347. http://dx.doi.org/10.1348/014466502760387461

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

Frith, C. D., & Corcoran, R. (1996). Exploring ‘theory of mind’ in people with schizophrenia. *Psychological Medicine, 26,* 521–530. http://dx.doi.org/10.1017/S0033291700035601

Frith, R. C. (1996). Conversational conduct and the symptoms of schizophrenia. *Cognitive Neuropsychiatry, 1,* 305–318. http://dx.doi.org/10.1080/135468096396460

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron, 80,* 519–530. http://dx.doi.org/10.1016/j.neuron.2013.09.009

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications, 8,* 14218. http://dx.doi.org/10.1038/ncomms14218

Kaplan, C. M., Saha, D., Molina, J. L., Hockeimer, W. D., Postell, E. M., Apud, J. A., . . . Tan, H. Y. (2016). Estimating changing contexts in schizophrenia. *Brain: A Journal of Neurology, 139,* 2082–2095. http://dx.doi.org/10.1093/brain/aww095

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry, 160,* 13–23. http://dx.doi.org/10.1176/appi.ajp.160.1.13

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, Optics, Image Science & Vision, 20,* 1434–1448. http://dx.doi.org/10.1364/JOSAA.20.001434

Lincoln, T. M., Peter, N., Schäfer, M., & Moritz, S. (2009). Impact of stress on paranoia: An experimental investigation of moderators and mediators. *Psychological Medicine, 39,* 1129–1139. http://dx.doi.org/10.1017/S0033291708004613

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5,* 39. http://dx.doi.org/10.3389/fnhum.2011.00039

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Frontiers in Human Neuroscience, 8,* 825. http://dx.doi.org/10.3389/fnhum.2014.00825

McCrae, R. R., & Costa, P. T., Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences, 36,* 587–596. http://dx.doi.org/10.1016/S0191-8869(03)00118-1

Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of jumping-to-conclusions bias in delusional patients. *Cognitive Neuropsychiatry, 16,* 422–447. http://dx.doi.org/10.1080/13546805.2010.548678

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology, 6,* e1000709.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science, 357,* 596–600. http://dx.doi.org/10.1126/science.aan3458

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

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement*. New York, NY: Appleton-Century-Crofts.

Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., . . . Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience, 33,* 13701–13712. http://dx.doi.org/10.1523/JNEUROSCI.1778-13.2013

Schmack, K., Rothkirch, M., Priller, J., & Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human Brain Mapping, 38,* 1767–1779. http://dx.doi.org/10.1002/hbm.23480

Schmidt, A., Cappucciati, M., Radua, J., Rutigliano, G., Rocchetti, M., Dell’Osso, L., . . . Fusar-Poli, P. (2017). Improving prognostic accuracy in subjects at clinical high risk for psychosis: Systematic review of predictive models and meta-analytical sequential testing simulation. *Schizophrenia Bulletin, 43,* 375–388.

Shaner, A. (1999). Delusions, superstitious conditioning and chaotic dopamine neurodynamics. *Medical Hypotheses, 52,* 119–123. http://dx.doi.org/10.1054/mehy.1997.0656

So, S. H., Freeman, D., Dunn, G., Kapur, S., Kuipers, E., Bebbington, P., . . . Garety, P. A. (2012). Jumping to conclusions, a lack of belief flexibility and delusional conviction in psychosis: A longitudinal investigation of the structure, frequency, and relatedness of reasoning biases. *Journal of Abnormal Psychology, 121,* 129–139. http://dx.doi.org/10.1037/a0025297

Speechley, W. J., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the “jumping to conclusions” bias associated with delusions in schizophrenia. *Journal of Psychiatry & Neuroscience, 35,* 7–17. http://dx.doi.org/10.1503/jpn.090025

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46,* 1004–1017. http://dx.doi.org/10.1016/j.neuroimage.2009.03.025

Stephan, K. E., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E. A., Brodersen, K. H., . . . Heinz, A. (2017). Computational neuroimaging strategies for single patient predictions. *NeuroImage, 145,* 180–199. http://dx.doi.org/10.1016/j.neuroimage.2016.06.038

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., . . . Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry, 84,* 634–643. http://dx.doi.org/10.1016/j.biopsych.2018.05.015

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

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning, 3,* 9–44.

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., . . . Fletcher, P. C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 13401–13406. http://dx.doi.org/10.1073/pnas.1503916112

van Os, J., Verdoux, H., Maurice-Tison, S., Gay, B., Liraud, F., Salamon, R., & Bourgeois, M. (1999). Self-reported psychosis-like symptoms and the continuum of psychosis. *Social Psychiatry and Psychiatric Epidemiology, 34,* 459–463. http://dx.doi.org/10.1007/s001270050220

Ventura, J., Wood, R. C., & Hellemann, G. S. (2013). Symptom domains and neurocognitive functioning can help differentiate social cognitive processes in schizophrenia: A meta-analysis. *Schizophrenia Bulletin, 39,* 102–111. http://dx.doi.org/10.1093/schbul/sbr067

Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch Ranganadan, A., Aponte, E. A., Ulrich, J., & Stephan, K. E. (2019). Social inference and beliefs differ in individuals with low and high subclinical persecutory delusional tendencies. *Schizophrenia Research*. Advance online publication. http://dx.doi.org/10.1016/j.schres.2019.08.031

White, T. P., Borgan, F., Ralley, O., & Shergill, S. S. (2016). You looking at me? Interpreting social cues in schizophrenia. *Psychological Medicine, 46,* 149–160. http://dx.doi.org/10.1017/S0033291715001622

Young, H. F., & Bentall, R. P. (1997). Probabilistic reasoning in deluded, depressed and normal subjects: Effects of task difficulty and meaningful versus non-meaningful material. *Psychological Medicine, 27,* 455–465. http://dx.doi.org/10.1017/S0033291796004540

Received September 30, 2018
Revision received September 12, 2019
Accepted December 13, 2019 ■