Inflexible social inference in individuals with subclinical persecutory delusional tendencies

Katharina V. Wellstein [a,*,1], Andreea Oliviana Diaconescu [a,b,c,1], Martin Bischof [d], Annia Rüesch [d,e], Gina Paolini [a,f], Eduardo A. Aponte [a], Johannes Ullrich [e], Klaas Enno Stephan [a,g,h]

[a] *Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland*
[b] *Department of Psychiatry (UPK), University of Basel, Switzerland*
[c] *Krembil Centre for Neuroinformatics (CAMH), University of Toronto, Canada*
[d] *Department of Psychiatry, University Hospital of Psychiatry (PUK), University of Zurich, Switzerland*
[e] *Department of Psychology, University of Zurich, Switzerland*
[f] *Klinik für Psychiatrie und Psychotherapie, Clienia Schlössli AG, Switzerland*
[g] *Wellcome Centre for Human Neuroimaging, University College London, UK*
[h] *Max Planck Institute for Metabolism Research, Cologne, Germany*

ARTICLE INFO

ABSTRACT

It has been suspected that abnormalities in social inference (e.g., learning others' intentions) play a key role in the formation of persecutory delusions (PD). In this study, we examined the association between subclinical PD and social inference, testing the prediction that proneness to PD is related to altered social inference and beliefs about others' intentions. We included 151 participants scoring on opposite ends of Freeman's Paranoia Checklist (PCL). The participants performed a probabilistic advice-taking task with a dynamically changing social context (volatility) under one of two experimental frames. These frames differentially emphasised possible reasons behind unhelpful advice: (i) the adviser's possible intentions (dispositional frame) or (ii) the rules of the game (situational frame). Our design was thus 2 × 2 factorial (high vs. low delusional tendencies, dispositional vs. situational frame). We found significant group-by-frame interactions, indicating that in the situational frame high PCL scorers took advice less into account than low scorers. Additionally, high PCL scorers believed more frequently that incorrect advice was delivered intentionally and that such misleading behaviour was directed towards them personally. Overall, our results suggest that social inference in individuals with subclinical PD tendencies is shaped by negative prior beliefs about the intentions of others and is thus less sensitive to the attributional framing of adviser-related information. These findings may help future attempts of identifying individuals at risk for developing psychosis and understanding persecutory delusions in psychosis.

1. Introduction

Delusions represent a hallmark of psychosis and are conceptualized as false beliefs based on incorrect inference about the external world that persist in the face of disconfirmatory evidence (DSM-IV 2000, p. 765 and DSM-5 2013, p. 819). The most prominent delusional beliefs pertain to the social world, specifically that other individuals' intentions are of persecutory nature (Bell and Halligan, 2013). Persecutory delusions (PD) shape the experience of 70% of first episode psychosis patients and 50% of patients from the schizophrenia spectrum(Freeman, 2007; Freeman and Garety, 2014; Sartorius et al., 1986), and are related

to reduced psychological well-being (Freeman et al., 2014) and higher risks of violence (Keers et al., 2014).

The understanding of delusions as abnormal beliefs and their immunity to disconfirmatory evidence led to influential concepts that suggested abnormalities of Bayesian inference (i.e. inference based on integrating observations with prior beliefs) as the cause of delusion formation (Coltheart et al., 2010; Hemsley and Garety, 1986). The notion that delusional ideation may be associated with abnormal inference has previously been related to the Jumping to Conclusions (JTC) bias in delusions (e.g., (Garety et al., 1991; Peters and Garety, 2006; So et al., 2012; Speechley et al., 2010); but see (Ermakova et al., 2017; Moutoussis et al., 2011) for alternative interpretations).

A more recent Bayesian account of delusions refers to the interplay between prior beliefs and prediction error (PE) signals (Corlett et al., 2016; Corlett et al., 2010; Fletcher and Frith, 2009; Sterzer et al., 2018). This derives from one prominent Bayesian perspective on

perception – predictive coding (Friston, 2005; Rao and Ballard, 1999) – which proposes that the brain infers the causes of its sensations using a hierarchical model of the external world. This model is assumed to represent beliefs that provide top-down predictions about sensory inputs, which are then adjusted by PEs at each level of the hierarchy. According to hierarchical Bayesian schemes, delusion formation might reflect a compensatory response to deficiencies of hierarchical inference (Adams et al., 2013; Corlett et al., 2016; Fletcher and Frith, 2009). Specifically, delusions might result from efforts to "explain-away" abnormally precise low-level PEs leading to adaptation of beliefs at higher levels in the processing hierarchy (Adams et al., 2013; Schmack et al., 2013). If these PEs are "chaotic" and result from usually unremarkable events (cf. "aberrant salience"; (Heinz, 2002; Kapur, 2003; Shaner, 1999)), adopting general and overly precise higher-level beliefs, may be a way of making sense of these events.

Adopting precise higher-level beliefs is crucial in social contexts, since human intentions are typically concealed (Biedermann et al., 2012). Previous studies linking Theory of Mind (ToM) and PD found that patients with PD have difficulty taking contextual factors into account when thinking about others' intentions and exhibit an enhanced externalising bias (i.e., a tendency to blame others rather than the situation for negative events, see (Craig et al., 2004; Langdon et al., 2006; Lincoln et al., 2010a); but also see (Martin and Penn, 2002; McKay et al., 2005)). This tendency to attributing harmful intent to others has also been associated with subclinical persecutory ideation (Raihani and Bell, 2017, 2018).

In this study, we investigated social inference in individuals with subclinical PD tendencies. Based on the notion of rigid high-level beliefs playing a crucial role in PD and assuming that PD is a dimensional construct, we generally expected that (1) individuals with subclinical tendencies towards PD should behave less adaptively in dynamic social contexts and (2) should be less sensitive to experimental manipulations of attributional biases (experimentally-induced priors).

Adopting a dimensional perspective on PD (Van Os et al., 1999), we invited participants from the general population scoring either high or low on the Paranoia Checklist (PCL; (Freeman et al., 2005; Lincoln et al., 2010b)). To investigate social inference, we employed an iterative probabilistic advice-taking paradigm. We probed the influence of attributional priors on social inference by introducing two experimental frames with minor differences in how the cause of incorrect advice was framed: First, a dispositional frame served to direct participants' attention to the adviser's character – namely, that the adviser acted intentionally in order to achieve his/her own (unknown) goals. Second, a situational frame directed participants' attention to the contextual aspects of the task – namely, that the adviser's behaviour was not only influenced by his/her intentions but also by the incompleteness of the information available to him/her. Individuals with relatively agnostic beliefs about the adviser were expected to learn and adhere to advice differently depending on how the task was framed. By contrast, individuals with proneness to PD were expected to be governed more by their own high-level beliefs than task-induced attributional priors. Thus, we predicted group-by-frame interactions in participants' decisions to adhere to advice during our social inference paradigm (Hypothesis I). Furthermore, we expected that high PCL scorers vs. low PCL scorers would attribute incorrect advice *less* to the adviser having incomplete information (Hypothesis II), and rather attribute negative events (e.g. bad performance on the task) to external-personal causes (Hypothesis III). Specifically, we hypothesized that high PCL scorers would attribute incorrect advice *more* to the adviser as a person than to themselves or the social context (Hypothesis IV). Additionally, we predicted that high PCL scorers would expect misleading advice (Hypothesis V) and believe that the adviser's giving incorrect advice was directed towards them personally (Hypothesis VI). For an overview of the hypotheses, their operationalization, and the results see Supplementary Fig. 1 in the Supplementary material.

These hypotheses were defined in an analysis plan prior to data analysis (https://gitlab.ethz.ch/sibak/sibak-analysis-plan). For a summary of all hypotheses and results, please see Supplementary Fig. 1. Notably, the hypotheses addressed in this paper refer exclusively to behavioural readouts from the task or to self-report measures. An independent analysis of the behavioural data that adresses additional hypotheses using a computational model of learning and inference is presented in a separate paper (Diaconescu et al., 2019).

2. Materials and methods

2.1. Pre-screening

Participants were recruited from the general population via online platforms for students and locals and via print adverts at stores. $N = 1145$ individuals were pre-screened online with the Paranoia Checklist (PCL; (Freeman et al., 2005; Lincoln et al., 2010b), intermixed with distractor items from the NEO-FFI (McCrae and Costa, 2004)), allowing us to assign individuals to groups characterised by high ("high PD") or low tendencies towards PD ("low PD"). The PCL is a self-report questionnaire consisting of 18 items representing statements linked to paranoid ideation. Subscales include frequency of the thoughts occurring, conviction, and distress. Group assignment was based on the frequency subscale. See Fig. 1 for the distribution of PCL frequency scores in our pre-screened sample.

The inclusion criteria for participating in the online pre-screening were as follows: (i) age 18 or older, (ii) fluent German comprehension, and (iii) absence of current treatment. The groups were defined in reference to the mean and standard deviation of the PCL scores obtained in healthy volunteers by Freeman et al. (2005). We used these reference values in order to enable continuous inclusion of participants during ongoing prescreening. Participants scoring $0.5sd$ above this mean were assigned to the high PD group and those scoring $0.5sd$ below were assigned to the low PD group.

In order to reduce the possibility of group assignment being based on a transient expression of persecutory ideation, participants assigned to one of the two groups were invited to fill in the online questionnaire again (screening), four weeks after participating in the pre-screening, which was the case for $N = 344$. Only individuals whose score remained outside the $0.5\ sd$ intervals described above when completing the PCL
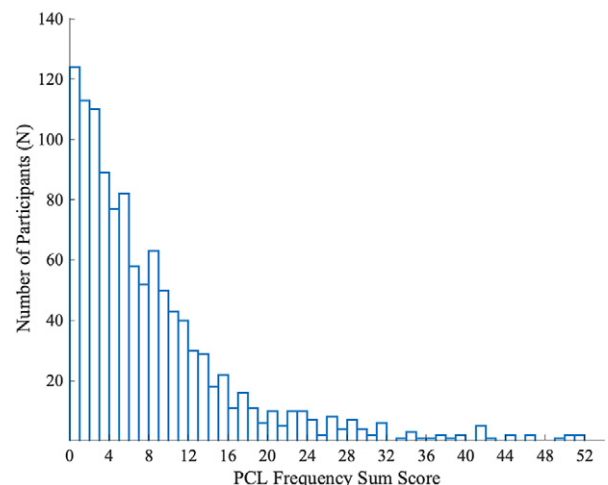

Fig. 1. Histogram of PCL frequency scores in pre-screening sample. The PCL frequency subscale assesses frequency of paranoid thoughts (18 items), with 0=*"less than once every month"* and 4=*"at least once a day"*; the highest possible score is 72. $N = 1145$ individuals filled in the questionnaire during pre-screening. This sample: *mean* = 7.73, *median* = 5.0, *sd* = 8.39, *range* = 0–52. Measures reported by Freeman et al. (2005): *mean* = 11.9, *median* = 9.0, *sd* = 10.5, *range* = 0–64.

for the second time were invited to the experiment, where they completed the PCL for a third time. The latter served to exclude any systematic differences in PCL scores obtained online versus on site.

2.2. Sample

We used a 2 × 2 factorial between-subject design with two participant groups assigned pseudo-randomly to two experimental conditions. Based on a power analysis (using g-Power (Faul et al., 2007)), a sample size of $N = 146$ was computed for a minimum of 80% power at $\alpha = 0.05$ under a moderate effect size (Cohen's $f^2 = 0.25$), required to run a two- and three-way ANOVA. Concerning the effect size, while there are no previous results that exactly relate to our questions, we sought guidance by a previous study (Diaconescu et al., 2014) which used the same type of task and reported a moderate effect size for a related question.

Assuming a drop-out rate of 10% (based on studies using the same task by Diaconescu et al. (2014, 2017)), we invited 162 participants to the experiment, and matched low PD and high PD participants regarding age, education, and proportion of male vs. female. Eleven participants were excluded from analyses based on previously defined exclusion criteria (specified in an analysis plan, time-stamped before completion of data acquisition and analyses (https://gitlab.ethz.ch/sibak/sibak-analysis-plan)).

Only participants with an average score over all three questionnaires (pre-screening, screening, and experiment day) that was outside the ± 0.5 sd intervals were included in the analyses, which was the case for 151 of the 162 individuals.

2.3. Experiment

2.3.1. Experimental procedure

After providing informed consent, participants received written standardised task instructions. To ensure that they understood the task, they were asked to explain it in their own words and performed a practice round (8 trials) in which they were truthfully informed that advice validity was fixed to chance.

After completing the task, participants filled out a task-specific debriefing questionnaire and were administered a cognitive screening – the symbols-test (coding) and the letters-numbers test (working memory) of the Brief Neurocognitive Assessment (BNA; (Fervaha et al., 2014)) to control for the potential influence of cognitive deficits on social inference. (Ventura et al., 2013) It was administered after the task to avoid possible influences of cognitive load on social inference (Gilbert and Osborne, 1989). After filling out the PCL again (Freeman et al., 2005; Lincoln et al., 2010b), participants were reimbursed and debriefed about the study before they left.

2.3.2. Task

The task used in this study is a modified version of the advice-taking task used by Behrens and colleagues (Behrens et al., 2008) (see Fig. 2), which has been used in a similar form in previous studies (Diaconescu et al., 2014, 2017).

Participants played a probabilistic lottery for monetary rewards trying to predict a binary outcome (blue or green) trial by trial (210 trials). Two sources of information, a social and a non-social cue, were presented on each trial. The latter was a pie-chart displaying a veridical probability distribution indicating what colour was more likely to win on any given trial. The pie-chart displayed different green-blue ratios (50:50, 55:45, 60:40, and 75:25) on each trial thus varying in its predictive uncertainty. The social cue (the advice) was represented by a videotaped adviser (recorded in a previous study (Diaconescu et al., 2014)) who gave a recommendation on which colour to choose, by holding up a card (blue or green). All participants were truthfully informed that the adviser did not have full information – and thus could make errors unintentionally – and that each piece of advice had been videotaped in a prior study with the same task (Diaconescu et al., 2014). In the Diaconescu et al. (2014) study, an additional control task which was matched in terms of volatility was conducted, in order to ensure that the main task truly captured social learning (learning from
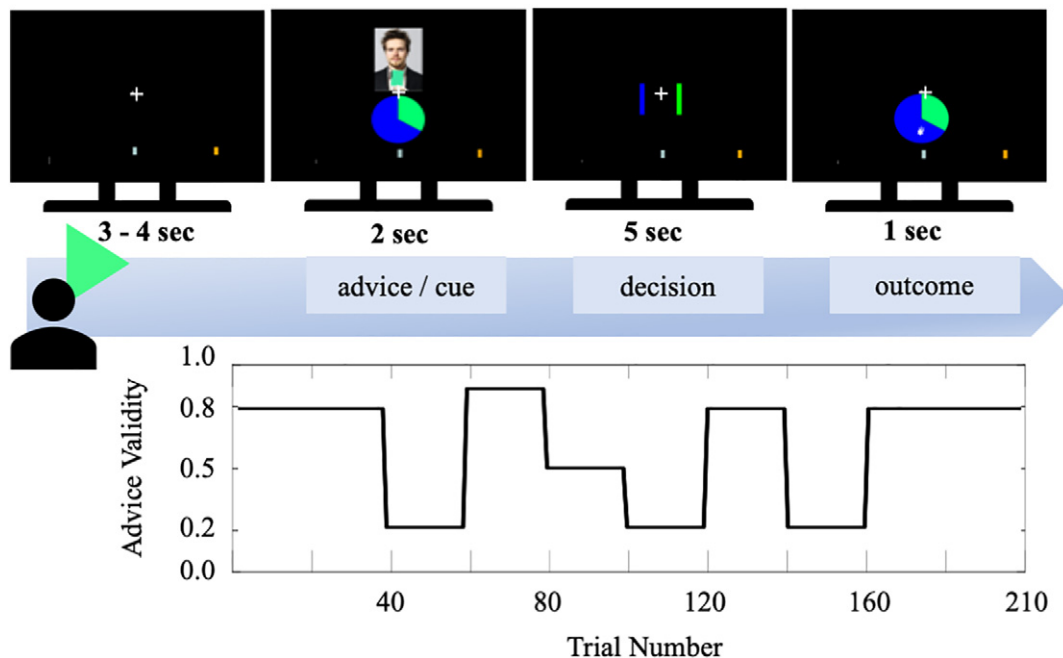

Fig. 2. Plot of task structure and trajectory of advice validity. Participants made binary decisions (blue vs. green) based on a social and a non-social cue (advice and pie-chart) presented simultaneously. Correct predictions result in the accumulation of points growing a progress-bar towards the targets displayed. Surpassing the targets earned participants additional CHF 10 (silver) or 20 (gold). After predicting the 'winning' colour they were informed regarding the real outcome. Pie-chart probabilities varied between 50:50, 55:45, 60:40, and 75:25. Advice validity varied across 210 trials (boxcar-chart): the first 42 trials consisted of predominantly ($p = 0.8$) correct advice (1st stable helpful phase), followed by 126 trials of highly variable advice validity (volatile phase). In the last 42 trials, advice was highly valid with $p = 0.8$ again (2nd stable helpful phase). The sequence of trials was fixed and identical across all participants.

changing intentions). In the control task participants acting as advisers chose the green or blue cards (their "advice") from a card deck which was accurate in either 80% or 20% of the trials. "Advisers" were blind-folded. This allowed the authors to control for the social aspect of the task, while matching the statistical structure of both pie chart information and cue-outcome relations identical. Performance in the control task differed significantly from performance in the main task, even when the card deck was 80% accurate, suggesting that an adviser hold-ing up the cards in a seemingly intentional way was processed differ-ently, which was also captured by a learning model incorporating a social-bias parameter (for more information see (Diaconescu et al., 2014)). Learning in this task should thus be driven by social inference about the adviser's intentions and not represent pure learning of statis-tics structure. We used one female and one male adviser, and adviser sex was balanced across groups and conditions.

Participants had to make predictions by integrating the two cues. After each decision, they received feedback on their choice and the cor-rect outcome. Players accrued points with every correct prediction. By achieving a cumulative score exceeding the silver or gold targets, partic-ipants earned an additional bonus, amounting to approx. 1/6 (silver) or 1/3 (gold) of the experiment's base reimbursement.

The task contained phases of differing advice validity; it began and ended with periods of high advice validity ($p = 0.80$, 42 trials each) and an intermediate period (126 trials) with changes in the advice-outcome contingency (volatility). The trial structure and sequence were fixed and identical across all participants.

2.3.3. Experimental conditions

The two experimental conditions (attributional frames) differed in how potentially unhelpful advice was framed: dispositional (caused by the adviser) vs. situational (caused by the rules of the game). Criti-cally, neither of the frames provided false information but summarised the adviser's role from different angles.

In the dispositional frame, participants' attention was directed to the adviser as a potential source of variability in advice validity, em-phasizing his/her ability of acting intentionally in order to achieve his/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/she was instructed to use the information available to him/her for guiding the player's behaviour. We induced the two frames by (i) one sentence in the instructions that differed between the two frames, (ii) a reminder on the task start-screen, and (iii) the wording used regarding advice validity. For more details, see the Supplementary material.

3. Results

3.1. Sample characteristics

151 of the 162 participants who took part in the experiment scored outside the ± 0.5 sd intervals of the PCL Frequency scale over all three questionnaire assessments (pre-screening, screening, experiment day), which was the cut-off for group assignment, and were thus eligible for analysis. They were on average 28 years old and had 15.8 years of ed-ucation (Table 1).

Sixty-four of 151 participants did not have an academic background. As expected, participant groups differed across all subscales of the PCL (see Table 2).

Furthermore, no significant difference between groups was detected regarding the cognitive screening administered (BNA working memory: $t = 0.43$, $p = 0.67$; BNA coding: $t = -1.20$, $p = 0.23$). It is thus unlikely that differences in task behaviour could stem from differences in cogni-tive capacity.

Participants also did not differ significantly in terms of performance accuracy on the task (two-tailed t-test, $df = 149$, $t = -1.82$, $p = 0.07$) with high PD participants' accuracy averaging at 0.60 and low PD

Table 1
Descriptive statistics of participants eligible for analyses.

Experimental frame	Low PD group		High PD group	
	M (SD)	Range	M (SD)	Range
Dispositional				
Age	29.42 (9.65)	18–67	26.47 (8.29)	18–49
Education (years)	17.12 (3.70)*	11–26	15.31 (3.11)*	9–21
N	41 (15 male)		36 (15 male)	
Situational				
Age	28.8 (9.74)	18–54	28.15 (11.03)	18–56
Education (years)	15.71 (4.22)	8–26	14.97 (3.86)	7–23
N	40 (17 male)		34 (15 male)	

$N = 151$, no differences in demographic variables detected between groups. Groups are not of equal size due to drop-outs (not responding to the invitation or no-shows) and par-ticipants being excluded from analyses after data acquisition based on previously defined criteria (see Supplementary material).
* $p < 0.05$, two-tailed t-tests, does not survive Bonferroni correction. All other variables: $p > 0.05$, two-tailed t-tests.

participants' at 0.61. Furthermore, both groups achieved a similar amount of points (high PD participants earned 41 and low PD partici-pants 47 on average; two-tailed t-test, $df = 149$, $t = -1.74$, $p = 0.08$) and reached the silver target on average.

3.2. Hypotheses

All following results represent main outcomes of hypothesis tests that were defined in a time-stamped analysis plan prior to data analysis (Analysis plan: https://gitlab.ethz.ch/sibak/sibak-analysis-plan, data: https://www.research-collection.ethz.ch/handle/20.500.11850/333102, and code: https://gitlab.ethz.ch/sibak/sibak-behavior-paper).

Additional effects (both significant and nonsignificant ones) of sec-ondary importance are reported in the Supplementary material (see Supplementary Fig. 1 for a summary of all hypotheses and results). This paper only reports analyses relating to task-behaviour (Hypotheses I–VI) and the "non-model based approach" in the analysis plan; for cor-responding computational analyses, see (Diaconescu et al., 2019).

In addition to inferential statistics (parametric or non-parametric tests, depending on the distribution of the data), we report effect size estimates ($\eta^2_{partial}$ for multiple regression analyses, Cohen's d for t-tests, and Cohen's r for non-parametric tests; (Fritz et al., 2012)).

3.2.1. Hypothesis I: high PD participants take information provided by the frame less into account than low PD participants (H_0 rejected)

We hypothesized that individuals with tendencies to PD behave less adaptively towards differences in social information as they rely on rigid high-level beliefs. We thus expected high PD participants to be less sen-sitive to the framing effect than low PD participants. We applied a two-way ANOVA with an interaction term (group, frame, and group x frame) to the participants' choices, i.e., how often they chose the colour recom-mended by the adviser. This group-by-frame interaction was significant

Table 2
Questionnaire scores of participants eligible for analyses.

	Low PD group		High PD group	
	M (SD)	Range	M (SD)	Range
PCL				
Frequency***	0.33 (0.34)	0–2	19.06 (5.43)	11–33
Conviction***	0.92 (1.57)	0–11	20.44 (5.77)	6–35
Distress***	17.88 (14.80)	0–46	27.96 (7.86)	11–44
BNA numbers-letters test	15.09 (3.49)	6–21	15.33 (3.43)	8–21
BNA symbols test	86.64 (14.10)	59–133	83.76 (15.41)	60–121

$N = 151$. Average PCL scores across the three questionnaire assessments. PCL subscales differed between groups.
*** $p < 0.001$, two-tailed t-test, unequal variances, Bonferroni corrected. Regarding BNA scores (cognitive performance) no differences were detected between groups, $p > 0.23$.

regarding participants choosing according to advice overall ($df = (1150), F = 5.77, p = 0.018, \eta^2_{partial} = 0.04$). Specifically, the interaction plot in Fig. 3a suggests that low PD participants more readily made decisions in accordance with the advice under the situational frame (highlighting incorrect advice as circumstantial) than under the dispositional frame (emphasizing incorrect advice as potentially caused by the adviser's intentions).

A post-hoc two-tailed t-test demonstrated that this difference between frames was significant in the low PD group ($df = 79, t = -5.55, p_{Bonferroni} = 7.3e-07$, Cohen's $d = 0.36$). In contrast, no significant difference in advice-taking behaviour between framing conditions could be detected in high PD participants ($df = 68, t = -1.52 \, p = 0.13$, Cohen's $d = 0.11$), suggesting that they did not integrate and utilize the information provided by the frame.

3.2.2. Hypothesis II: high PD participants attribute misleading advice validity to the adviser's character rather than to the possibility of the adviser making a mistake due to incomplete information. (H_0 not rejected)

We tested this hypothesis with regard to participants' advice-taking in the task phases where advice was stable and helpful vs. when advice validity was highly variable. We predicted that in the situational frame,

high PD participants would more frequently choose against the advice when it was volatile, and that this frame-specific difference between volatile and stable phases would be stronger than in low PD participants. In the dispositional frame, we expected smaller group differences. However, this three-way interaction was not significant (phase*group*frame: $F = 0.29, p = 0.59, \eta^2_{partial} = 0.002$, see Fig. 3c and the Supplementary material for more information and analyses).

3.2.3. Hypothesis III: high PD compared to low PD participants generally attribute negative events to more external-personal causes. (H_0 not rejected)

We tested this with debriefing question 14 ("*In your opinion, what factors determined your performance in the task?*", participants distributed a total of 100% to the following answer options: "a. *You as the player*", "b. *The adviser which was appointed to you*", and "c. *The rules/ the structure of the game*"). This hypothesis was not confirmed (see Supplementary material for more information).

3.2.4. Hypothesis IV: high PD participants attribute differences in advice validity to the adviser being malevolent. (H_0 rejected)

We hypothesized that high PD participants exhibited more negative beliefs about the adviser and more readily attributed differences
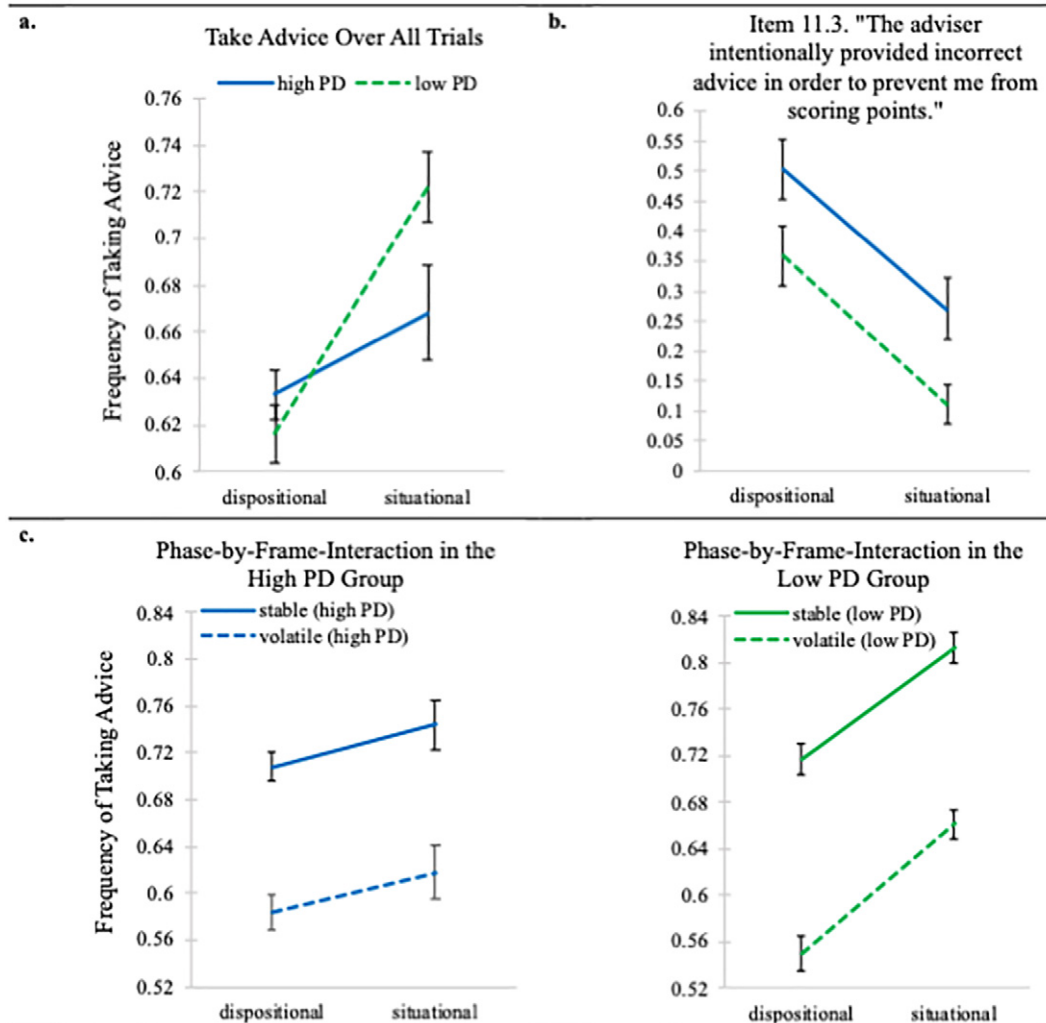

Fig. 3. ANOVAs for advice-taking behaviour ANOVAs. While no differences in advice-taking between frames were identified in high PD participants, low PD participants took advice into account less frequently in the dispositional vs. the situational frame. a. Two-way interaction for taking advice overall: group*frame: $F = 5.77, p = 0.018, \eta^2_{partial} = 0.04$ frame: $F = 22.71, p = 4.4754e-06, \eta^2_{partial} = 0.13$. b. Main effects regarding agreement with statement: "*The adviser intentionally provided incorrect advice in order to prevent me from scoring points*" (H. VI) group: $F = 11.01, p = 0.001, \eta^2_{partual} = 0.07$ frame: $F = 27.39, p = 5.63e-07, \eta^2_{partial} = 0.16$. c. Three-way interaction for taking advice in stable vs. volatile phases: phase*group*frame: $F = 0.29, p = 0.59, \eta^2_{partial} = 0.002$ phase*frame: $F = 0.15, p = 0.7, \eta^2_{partial} = 0.001$ phase*group: $F = 3.86, p = 0.051, \eta^2_{partial} = 0.03$ Y-axes show percentages, error bars indicate standard errors of the mean.

in advice validity to the adviser's character than low PD individuals and that this effect would be moderated by the framing. We tested this with debriefing item "*The adviser intentionally gave false information, without profiting from it.*", which states a possible cause for incorrect advice. Participants had to indicate how much they believed this cause played a role for the adviser providing incorrect advice (percentage assignment 0%–100%). While the interaction was not significant, the main effect of group was ($df = (1150), F = 16.87 \, p = 6.63e{\text -}05, \eta^2_{partial} = 0.1$). On average, high PD participants assigned $38.6 \pm 29.2\%$ to this statement whereas low PD participants assigned $20.3 \pm 25.9\%$.

In the debriefing questionnaire, participants were asked to rate what caused incorrect advice, and assign percentages to the adviser and to the rules of the game (debriefing question 13: "*When the adviser provided you with* misleading/incorrect *advice: What do you believe was the cause of this?*"). Participants distributed a total of 100% to the following answer options: "a. *The adviser which was appointed to you*" and "b. *The rules/the structure of the game*". Given the nature of the distributions (percentages assigned to the different answer options were tied because they had to sum up to 100%), we used a Wilcoxon rank sum test to compare groups. As expected, high PD participants were more likely to rate the adviser as the cause for incorrect advice ($40.0 \pm 23.8\%$), compared to low PD participants ($29.6 \pm 27.7\%; z = 2.42, p = 0.008, Cohen's r = 0.2$).

3.2.5. Hypothesis V: high PD participants expect receiving misleading advice. (H₀ not rejected)

We tested this hypothesis with the debriefing questionnaire (administered after the task), where we asked what advice participants expected to receive before playing the game ($1 = correct\ advice, 6 = incorrect\ advice$). We expected an interaction effect with high PD scoring higher than low PD participants in the situational frame but not the dispositional frame. While this interaction did not reach significance, both main effects did, showing that low PD participants scored lower on this scale on average (2.81 ± 1.18), indicating a tendency towards expecting correct advice compared to high PD participants (3.45 ± 1.19) than low PD participants (main effect of group; $df = 147, F = 11.72, p = 8.0e{\text -}04, \eta^2_{partial} = 0.07$, see Supplementary for full statistics).

3.2.6. Hypothesis VI: high PD participants view incorrect advice as directed towards them. (H₀ not rejected)

This could be directly assessed via debriefing questionnaire item "*The adviser intentionally provided incorrect advice in order to prevent me from scoring points.*" (percentage assignment 0%–100%). We originally expected an interaction here, with low PD participants assigning higher percentages to this statement in the dispositional compared to the situational frame and high PD participants assigning high percentages across framing conditions. However, only the main effect of group reached significance ($df = (1150), F = 11.01, p = 0.001, \eta^2_{partial} = 0.07$), with high PD participants endorsing the statement more ($39 \pm 31.9\%$) than low PD participants ($23.6 \pm 29.2\%$, see Fig. 3b).

4. Discussion

In this study, we investigated how individuals scoring high vs. low on the Paranoia Checklist (PCL; (Freeman et al., 2005)) incorporate attributional priors into learning from advice. Participants performed a probabilistic advice-taking paradigm with variable advice-outcome contingencies under one of two experimental frames which differentially emphasised causes of social information (dispositional vs. situational attributional frames).

We found that low PCL scorers took advice into account less under the dispositional (highlighting the adviser as the cause for incorrect advice) vs. the situational frame (highlighting incorrect

advice as circumstantial), whereas high PCL scorers did not differ between framing conditions (Fig. 3a). High PD participants' behaviour is similar to what was reported in a recent study using the "dictator game" (Raihani and Bell, 2017); the authors induced two differing experimental contexts (being at the receiving end of a dictator's decisions or a third-person observer) and found that persecutory ideation was related to attributing harmful intent, irrespective of context (Raihani and Bell, 2017). Low PD participants' advice-taking behaviour corresponds to what Fouragnan et al. found in a recent study, namely that prior knowledge about an agent's reputation generally influences learning about intentions (Fouragnan et al., 2013). Specifically, healthy participants relied on experimentally-induced reputation priors in a social learning task, even in light of disconfirming evidence.

Our dispositional frame, introduced a negative prior about the adviser's intentions, prompting low PD participants to disregard advice more often. The group-by-frame interaction (Fig. 3a) indicates that low scorers' advice-taking behaviour seemed to be driven by these experimentally-induced priors. High scorers however disregarded the advice irrespective of experimental frame, even in the "safer" social context when the rules of the game were emphasised as causes for incorrect advice.

One explanation for high PD participants' lack of differences in advice-taking behaviour across frames, might be that – rather than the experimentally-induced priors – their high-level prior beliefs about the adviser's intentions influenced their decisions. Specifically, from the perspective of hierarchical models of Bayesian inference, when low-level beliefs (e.g. about trial-by-trial behaviour) are ambiguous (have high variance) and higher-level beliefs (e.g. about the intentions of others) are held with more conviction (precision), perception and learning will be more strongly influenced by higher-level beliefs. Thus, the reduced impact of the framing in this study might reflect high PD participants' reliance on strong (precise) higher-level prior beliefs in the face of weaker (less precise) experimentally-induced, lower-level predictions. This explanation is in line with findings showing an increased influence of prior beliefs on the perception of ambiguous stimuli (Schmack et al., 2013) as well as reduced use of experimentally-induced priors in delusion-prone individuals (Stuke et al., 2018), suggesting that delusion(−proneness) is characterised by differences in the precision of prior beliefs at different levels of the processing hierarchy.

Indeed, high PD participants reported viewing the adviser as the main cause of incorrect advice, as opposed to considering the adviser having incomplete information (e.g., Hypothesis 4, debriefing question 13 "*When the adviser provided you with* misleading/incorrect *advice: What do you believe was the cause of this?*", high PD participants agreed more with "*The adviser which was appointed to you*"). Furthermore, compared to low PD participants, they reported expecting misleading advice more (Hypothesis 5) and viewing the adviser as acting intentionally malevolent towards them more (Hypothesis 6).

These findings suggest that high PCL scorers relied on overly precise negative higher-level prior beliefs when inferring on the adviser's intentions Following a reviewer's suggestion to scrutinise this interpretation in additional exploratory analyses, we investigated whether participants' conviction scores on the PCL (obtained prior to the experiment) were related to how they answered debriefing questionnaire item 11.3 ("*The adviser intentionally provided incorrect advice in order to prevent me from scoring points.*"). We found that the higher the conviction ratings regarding persecutory beliefs as assessed with the PCL, the more the participant judged the adviser as acting intentionally and malevolently ($F = 12.8, p = 0.0005, \eta^2_{partial} = 0.07$).

An alternative interpretation for the lack of between-condition differences in high PD participants might be that they did not believe the experimental framing that was induced via task instruction. We do not think this was the case, for the following reasons:

(i) We asked participants at the end of the debriefing questionnaire for feedback on different aspects of the study (debriefing questions 17 and 18), whether the instructions were understandable (Q 19), and if they felt influenced by them in how they viewed the adviser (Q 20) and how they played the game (Q 21). None of the participants mentioned suspecting anything or being influenced. Furthermore, after having been debriefed about the study and the framing, participants stated not having suspected anything.

(ii) Additionally, in the debriefing questionnaire, we asked participants to indicate their agreement with two statements probing the information highlighted in the two experimental frames: The adviser not having full information being the cause of incorrect advice in the situational frame (Q 11.5, "The adviser, in general, did not have full information and made mistakes.") and the possibility of the adviser being the cause of incorrect advice in the dispositional frame (Q 11.4, "The adviser intentionally provided false information because this was part of his/her instructions/task.").

We found a significant main effect of framing, together with non-significant main effect of group and interactions (see Supplementary Table 3 for detailed statistics). This suggests that (i) both high and low PD participants viewed the adviser's incomplete information as a more likely cause for incorrect advice in the situational compared to the dispositional frame (Q 11.5) and that the adviser providing false information due to their instructions was the more likely cause for incorrect advice in the dispositional frame compared to the situational frame (Q 11.4). These effects were expected based on the framing instructions. Thus, participants generally seemed to have a similar understanding of the framing and did not disregard the framing-specific instructions (see 8.2.1 and 8.2.1 in the Supplementary material for detailed statistics).

Our findings align with previous reports of theory of mind (ToM) deficits in psychosis patients with delusions, indicating an external-personal attribution bias (Bliksted et al., 2014; Craig et al., 2004; Frith and Corcoran, 1996; Langdon et al., 2001; Langdon et al., 2006), and extend these findings to subclinical PD. The association between social cognition and PD in subclinical populations has thus far been inconclusive ((McKay et al., 2005) and for a review see (Garety and Freeman, 2013)), potentially due to ToM paradigms being predominantly questionnaire-based measures. Although recent investigations of external-attributions in persecutory ideation captured dynamic aspects of social inference (Raihani and Bell, 2017), how predictions are updated as a result of contradicting evidence or PEs when information is continuously changing has not been examined yet.

The results of this study suggest that individuals with PD tendencies may incorporate PEs less into their predictions about the adviser's intentions due to more rigid high-level prior beliefs about the adviser's intentions. In a separate analysis, computational modelling of our data suggested that maladaptive social inference in the task may result from overly precise higher-level beliefs about the adviser's fidelity, leading to reduced belief-updating in the face of incoming PEs (Diaconescu et al., 2019).

In this study, we aimed to capture the behaviour of individuals for whom persecutory ideation was a stable trait. Since impairments of social cognition might contribute to risk for developing psychosis and are found in first-episode psychosis (FEP) patients (Bora and Pantelis, 2013; Sun et al., 2011), future extension of our approach might serve to address clinically-relevant predictions, such as transition to psychosis in clinical high risk individuals and treatment response in FEP patients.

Contributors

Katharina V. Wellstein, Andreea O. Diaconescu, and Klaas Enno Stephan equally contributed to designing the study and establishing the study protocol. Martin Bischof and Johannes Ulrich provided expert advice, which improved the study design and the subsequent analyses. Andreea O. Diaconescu programmed the task used in this study. Katharina V. Wellstein and Annia Rüesch acquired the data. Katharina V. Wellstein and Andreea O. Diaconescu equally managed the literature searches and analyses. Katharina V. Wellstein and Andreea O. Diaconescu undertook the statistical analysis. Eduardo A. Aponte performed the code review; i. e., independently examining the robustness of the analysis

pipeline and tested the reproducibility of the results. Katharina V. Wellstein, Andreea O. Diaconescu, Martin Bischof, Annia Rüesch, Gina Paolini, Eduardo A. Aponte, Johannes Ullrich, and Klaas Enno Stephan wrote the paper. All authors approved the final manuscript.

Declaration of competing interest

The authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.schres.2019.08.031.

References

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The computational anatomy of psychosis. Front. Psychiatry 4 (May), 1–26. https://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 (7219), 245–249. https://doi.org/10.1038/nature07538. Associative.

Bell, V., Halligan, P.W., 2013. The neural basis of abnormal personal belief. The Neural Basis of Human Belief Systems, pp. 191–224.

Biedermann, F., Frajo-Apor, B., Hofer, A., 2012. Theory of mind and its relevance in schizophrenia. Curr. Opin. Psychiatry 25 (2), 71–75. https://doi.org/10.1097/YCO.0b013e3283503624.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., Videbech, P., 2014. Social cognition and neurocognitive deficits in first-episode schizophrenia. Schizophr. Res. 153 (1–3), 9–17. https://doi.org/10.1016/j.schres.2014.01.010.

Bora, E., Pantelis, C., 2013. Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. Schizophr. Res. 144 (1–3), 31–36. https://doi.org/10.1016/j.schres.2012.12.013.

Coltheart, M., Menzies, P., Sutton, J., 2010. Abductive inference and delusional belief. Cogn. Neuropsychiatry 15 (1–3), 261–287. https://doi.org/10.1080/13546800903439120.

Corlett, P.R., Taylor, J.R., Wang, X.J., Fletcher, P.C., Krystal, J.H., 2010. Toward a neurobiology of delusions. Prog. Neurobiol. 92 (3), 345–369. https://doi.org/10.1016/j.pneurobio.2010.06.007.

Corlett, P.R., Honey, G.D., Fletcher, P.C., 2016. Prediction error, ketamine and psychosis: an updated model. J. Psychopharmacol. 30 (11), 1145–1155. https://doi.org/10.1177/0269881116650087.

Craig, J.S., Hatton, C., Craig, F.B., Bentall, R.P., 2004. Persecutory beliefs, attributions and theory of mind: comparison of patients with paranoid delusions, Asperger's syndrome and healthy controls. Schizophr. Res. 69 (1), 29–33. https://doi.org/10.1016/S0920-9964(03)00154-3.

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 Comput. Biol. 10 (9), e1003810. https://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. Soc. Cogn. Affect. Neurosci. 12 (November 2016), 1–17. https://doi.org/10.1093/scan/nsw171.

Diaconescu, A.O., Wellstein, K.V., Mathys, C., Stephan, K.E., 2019. Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. J. Abnorm. Psychol. (under revi(Predictive Coding and Psychopathology)).

Ermakova, A.O., Gileadi, N., Knolle, F., Diaz, A.J., Anderson, R., 2017. Cost evaluation during decision making in patients at early stages of psychosis. Comput. Psychiatry, 1–39 https://doi.org/10.1101/225920 in press.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. GPower 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191. https://doi.org/10.3758/BF03193146.

Fervaha, G., Hill, C., Agid, O., Takeuchi, H., Foussias, G., Siddiqui, I., Remington, G., 2014. Examination of the validity of the Brief Neurocognitive Assessment (BNA) for schizophrenia. Schizophr. Res. 166 (1–3), 304–309. https://doi.org/10.1016/j.schres.2015.05.015.

Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat. Rev. Neurosci. 10 (1), 48–58. https://doi.org/10.1038/nrn2536.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., Coricelli, G., 2013. Reputational priors magnify striatal responses to violations of trust. J. Neurosci. 33 (8), 3602–3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013.

Freeman, D., 2007. Suspicious minds: the psychology of persecutory delusions. Clin. Psychol. Rev. 27 (4), 425–457. https://doi.org/10.1016/j.cpr.2006.10.004.

Freeman, D., Garety, P., 2014. Advances in understanding and treating persecutory delusions: a review. Soc. Psychiatry Psychiatr. Epidemiol. 49 (8), 1179–1189. https://doi.org/10.1007/s00127-014-0928-7.

Freeman, D., Garety, P.A., Bebbington, P.E., Smith, B., Rollinson, R., Kuipers, E., Katarzyna, R., 2005. Psychological investigation of the structure of paranoia in a non-clinical population. Br. J. Psychiatry 186, 427–435. https://doi.org/10.1192/bjp.186.5.427.

Freeman, D., Startup, H., Dunn, G., Wingham, G., Černis, E., Evans, N., Kingdon, D., 2014. Persecutory delusions and psychological well-being. Soc. Psychiatry Psychiatr. Epidemiol. 49 (7), 1045–1050. https://doi.org/10.1007/s00127-013-0803-y.

Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc., B 360 (1456), 815–836. https://doi.org/10.1098/rstb.2005.1622.

Frith, C.D., Corcoran, R., 1996. Exploring 'theory of mind' in people with schizophrenia. Psychol. Med. 26 (03), 521. https://doi.org/10.1017/S0033291700035601.

Fritz, C.O., Morris, P.E., Richler, J.J., 2012. Effect size estimates: current use, calculations, and interpretation. J. Exp. Psychol. Gen. 141 (1), 2–18. https://doi.org/10.1037/a0024338.

Garety, P.A., Freeman, D., 2013. The past and future of delusions research: from the inexplicable to the treatable. Br. J. Psychiatry 203 (5), 327–333. https://doi.org/10.1192/bjp.bp.113.126953.

Garety, P.A., Hemsley, D.R., Wessley, M.R.C., 1991. Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. J. Nerv. Ment. Dis. 179 (4), 149–201.

Gilbert, D.T., Osborne, R.E., 1989. Thinking backward: some curable and incurable consequences of cognitive busyness. J. Pers. Soc. Psychol. 57 (6), 940–949. https://doi.org/10.1037/0022-3514.57.6.940.

Heinz, a., 2002. Dopaminergic dysfunction in alcoholism and schizophrenia–psychopathological and behavioral correlates. Eur. Psychiatry 17 (1), 9–16 https://doi.org/S0924933802006284 [pii].

Hemsley, D.R., Garety, P.A., 1986. The formation of maintenance of delusions. Br. J. Psychiatry 149 (1), 51–56. https://doi.org/10.1192/bjp.149.1.51.

Kapur, S., 2003. Psychosis as a state of aberrant salience: and pharmacology in schizophrenia. Am. J. Psychiatr. 160, 13–23. https://doi.org/10.1176/appi.ajp.160.1.13.

Keers, R., Ullrich, S., DeStavola, B.L., Coid, J.W., 2014. Association of violence with emergence of persecutory delusions in untreated schizophrenia. Am. J. Psychiatr. 171 (3), 332–339. https://doi.org/10.1176/appi.ajp.2013.13010134.

Langdon, R., Coltheart, M., Ward, P.B., Catts, S.V., 2001. Mentalising, executive planning and disengagement in schizophrenia. Cogn. Neuropsychiatry 6 (2), 81–108. https://doi.org/10.1080/13546800042000061.

Langdon, R., Corner, T., McLaren, J., Ward, P.B., Coltheart, M., 2006. Externalizing and personalizing biases in persecutory delusions: the relationship with poor insight and theory-of-mind. Behav. Res. Ther. 44 (5), 699–713. https://doi.org/10.1016/j.brat.2005.03.012.

Lincoln, T.M., Mehl, S., Exner, C., Lindenmeyer, J., Rief, W., 2010a. Attributional style and persecutory delusions. Evidence for an event independent and state specific external-personal attribution bias for social situations. Cogn. Ther. Res. 34 (3), 297–302. https://doi.org/10.1007/s10608-009-9284-4.

Lincoln, T.M., Ziegler, M., Lüllmann, E., Müller, M.J., Rief, W., 2010b. Can delusions be self-assessed? Concordance between self- and observer-rated delusions in schizophrenia. Psychiatry Res. 178 (2), 249–254. https://doi.org/10.1016/j.psychres.2009.04.019.

Martin, J.A., Penn, D.L., 2002. Attributional style in schizophrenia: an investigation in outpatients with and without persecutory delusions. Schizophr. Bull. 28 (1), 131–141. https://doi.org/10.1093/oxfordjournals.schbul.a006916.

McCrae, R.R., Costa, P.T., 2004. A contemplated revision of the NEO Five-Factor Inventory. Personal. Individ. Differ. 36 (3), 587–596. https://doi.org/10.1016/S0191-8869(03)00118-1.

McKay, R., Langdon, R., Coltheart, M., 2005. Paranoia, persecutory delusions and attributional biases. Psychiatry Res. 136 (2–3), 233–245. https://doi.org/10.1016/j.psychres.2005.06.004.

Moutoussis, M., Bentall, R.P., El-Deredy, W., Dayan, P., 2011. Bayesian modelling of jumping-to-conclusions bias in delusional patients. Cogn. Neuropsychiatry 16 (5), 422–447. https://doi.org/10.1080/13546805.2010.548678.

Peters, E., Garety, P., 2006. Cognitive functioning in delusions: a longitudinal analysis. Behav. Res. Ther. 44 (4), 481–514. https://doi.org/10.1016/j.brat.2005.03.008.

Raihani, N.J., Bell, V., 2017. Paranoia and the social representation of others: a large-scale game theory approach. Sci. Rep. 7 (1), 1–9. https://doi.org/10.1038/s41598-017-04805-3.

Raihani, N.J., Bell, V., 2018. Conflict and cooperation in paranoia: a large-scale behavioural experiment. Psychol. Med. 48 (9), 1523–1531. https://doi.org/10.1017/S0033291717003075.

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2 (1), 79–87.

Sartorius, N., Sartorius, N., Jablensky, A., Jablensky, A., Korten, A., Korten, A., ... Day, R., 1986. Early manifestations and first-contact incidence of schizophrenia in different cultures. Psychol. Med. 16, 909–928.

Schmack, K., Gomez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rossler, H., Haynes, J.-D., Sterzer, P., 2013. Delusions and the role of beliefs in perceptual inference. J. Neurosci. 33 (34), 13701–13712. https://doi.org/10.1523/JNEUROSCI.1778-13.2013.

Shaner, A., 1999. Delusions, superstitious conditioning and chaotic dopamine neurodynamics. Med. Hypotheses 52 (2), 119–123. https://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. J. Abnorm. Psychol. 121 (1), 129–139. https://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. J. Psychiatry Neurosci. 35 (1), 7–17. https://doi.org/10.1503/jpn.090025.

Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Corlett, P.R., 2018. The predictive coding account of psychosis. Biol. Psychiatry https://doi.org/10.1016/j.biopsych.2018.05.015.

Stuke, H., Weilnhammer, V.A., Sterzer, P., Schmack, K., 2018. Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. Schizophr. Bull., 1–7 https://doi.org/10.1093/schbul/sbx189.

Sun, H., Young, N., Hwan, J., Kim, E., Shim, G., Yoon, H., ... Soo, J., 2011. Social cognition and neurocognition as predictors of conversion to psychosis in individuals at ultra-high risk. Schizophr. Res. 130 (1–3), 170–175. https://doi.org/10.1016/j.schres.2011.04.023.

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. Soc. Psychiatry Psychiatr. Epidemiol. 34 (9), 459–463. https://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. Schizophr. Bull. 39 (1), 102–111. https://doi.org/10.1093/schbul/sbr067.