Exaggerated Volatility Beliefs Drive Social Hallucinations in Paranoia

Krisztina Jedlovszky (krisztina.jedlovszky.21@ucl.ac.uk)[1,2], Daniel Yon[2] & Philip R. Corlett[1]
[1]Department of Psychiatry, Yale University
[2] School of Psychological Sciences, Birkbeck, University of London

Abstract

The social content of paranoid delusions has prompted theories that paranoia stems from deficits in social cognition. For instance, paranoid individuals are more likely to falsely attribute harmful intentions to randomly moving visual patterns—'social hallucinations.' However, paranoia also involves domain-general deficits in belief updating and learning. To investigate if domain-general deficits are sufficient to explain social hallucinations, participants with varying paranoia levels completed reversal learning tasks (social/nonsocial × easy/hard), then judged intentional motion in ambiguous displays. High-paranoia participants showed increased win-switching and elevated social hallucinations, with the former predicting the latter — suggesting linked mechanisms. Hierarchical Gaussian Filter modeling revealed context-dependent mechanisms: when volatility was salient, prior volatility beliefs (μ_3) linked win-switching, social hallucinations, and paranoia; when social context was salient, meta-volatility learning (ω_3) mediated these relationships. These findings demonstrate that social hallucinations emerge from domain-general belief updating problems operating through context-sensitive pathways.

Keywords: Paranoia; Social hallucinations; Belief updating; Volatility; Hierarchical Gaussian Filter

Introduction

Paranoia — the fixed belief that others intend to do one harm — involves profound disturbances in social functioning. Individuals with paranoia hold strong convictions about the malicious intentions of others, accompanied by demonstrable impairments in how they process information about their environment, and update their beliefs in response to new evidence (Diaconescu et al., 2020; Fett et al., 2019; Hauke et al., 2024; Suthaharan et al., 2021; Wellstein et al., 2020). The fundamentally social nature of these deficits has led to a proliferation of theories proposing that paranoia is primarily a deficit in social cognition, manifesting through mentalizing deficits, theory of mind impairments, and disturbances in processing social threats (Barnby et al., 2023; Bell et al., 2021; Raihani & Bell, 2018).

However, recent computational work suggests a more domain-general basis for paranoid ideation, specifically through aberrant processing of environmental volatility. Research has consistently documented domain-general deficits in paranoid thinking, characterized by beliefs of heightened environmental uncertainty, reduced flexibility when integrating disconfirmatory information, and an increased sensitivity to change (Freeman et al., 2011; Jedlovszky et al., 2024; Prike et al., 2018; Reed et al., 2020; Suthaharan et al., 2021). Computational modeling approaches have revealed that these phenomena are accompanied by aberrant learning parameters, including higher volatility priors, increased learning rates, and altered weighting of prediction errors (Diaconescu et al., 2020; Gibbs-Dean et al., 2023; Reed et al., 2020; Stuke et al., 2017) . It seems that these domain-general impairments reflect fundamental differences in how beliefs are formed and updated across diverse contexts, providing a computational foundation for understanding the cognitive architecture of paranoid thoughts (Corlett et al., 2025; Feeney et al., 2017; Sheffield et al., 2022).

It has remained unclear whether the domain-general differences in volatility processing seen in paranoia are an ‘additional’ deficit alongside disruptions in social cognition, or whether these domain-general changes could in fact explain the social features of paranoia without impairments in social cognition per se. While general differences in learning and inference seen in paranoia should operate across many contexts (by definition), these general alterations in cognition could alter perception and behaviour in seemingly social-specific ways. For instance, an elevated sense that our surroundings are volatile could allow us to update our beliefs more quickly in the face of little evidence. This general predisposition to readily accept unreliable 'hypotheses' and ‘jump to conclusions’ (Garety et al., 2011) when deployed in social settings - could structure our perception of social interactions, such that we perceive structure and intentions where none genuinely exist.

Social hallucinations — aberrant agency attribution when processing visual information — provide an ideal test case for evaluating this possibility. Recent work by Castiello et al. (2024) identified ‘social hallucinations’ as a distinct feature of paranoia, where individuals inappropriately assign intentionality and agency to ambiguous visual stimuli, such as moving disks (Castiello et al., 2024). Unlike high-level social cognition deficits that involve explicit reasoning about others' mental states, social hallucinations emerge during fundamental visual processing of motion patterns, operating at a perceptual rather than cognitive level. This low-level nature makes them particularly informative, as they allow us to trace how information cascades downstream through the cognitive hierarchy. If domain-general volatility processing shapes these low-level perceptual interpretations, this would support the idea that alterations in domain-general mechanisms ’upstream’ can generate seemingly social-specific phenomena ‘downstream‘. However, if volatility processing and social hallucinations are not functionally connected, this would suggest that independent social cognition deficits operating at the perceptual level, separate to general differences in belief updating.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

The present study addresses two fundamental questions about the relationship between domain-general belief updating and low-level social hallucinations in paranoia. First, we investigate whether social hallucinations arise from the same computational mechanisms that produce domain-general learning impairments, or whether they reflect a distinct process specific to social cognition. Second, we examine how environmental volatility modulates both domain-general learning and social hallucinations. Across two experiments, we manipulated both volatility (easy vs. hard) and domain (social vs. non-social) in a probabilistic reversal learning task, followed by assessment of social hallucinations through perceived animacy in visual chasing displays. By combining behavioural measures with Hierarchical Gaussian Filter (HGF) computational modeling, we aimed to identify the specific uncertainty-processing mechanisms that could link domain-general belief updating deficits to social hallucinations in paranoid thinking.

Materials and Methods

Study Design

We administered a computerised version of a 3-arm Probabilistic Reversal Learning (PRL) task followed by a Perceived Animacy (PA) task (adapted from Study 2 of Castiello et al. 2024 (Castiello et al., 2024)). This was followed by a set of questionnaires in Qualtrics, also completed online.

We conducted an initial study with N = 400 participants (Experiment 1), followed by a pre-registered replication (https://aspredicted.org/ne85z9.pdf), that had N = 399 participants and a slightly adjusted task battery (Experiment 2). The setup and sequence of both tasks are depicted in Figure 1.

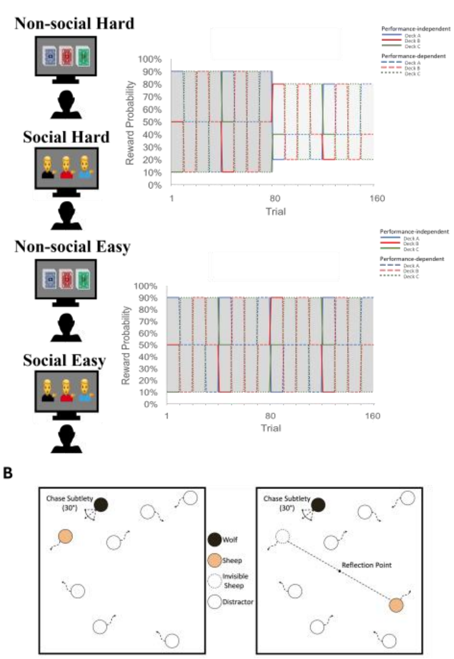

Figure 1: PRL task conditions (A), and the PA task setup with a chase-present and a chase absent trial, adapted from Castiello et al. 2024 (B).

In the PRL task, participants were instructed to choose between 3 arms to receive probabilistic +100pt rewards or –50pt punishments. The reward probabilities associated with each arm changed both randomly and depending on performance (for more details see Reed et al. 2020). Reward probabilities in the first half (2 blocks, 40 trials each) were uniformly 10%/50%/90% across all participants. In the easy version of the task, the second half (2 blocks, 40 trials each) had the same reward probabilities, while in the hard version, the reward contingencies changed to 20%/40%/60%. We also included a social manipulation, in which the arms were either represented by decks of cards (non-social) or human avatars (social). Participants in Experiment 1 were therefore randomly assigned to one of the four PRL conditions: Social Easy, Social Hard, Non-social Easy, Non-social Hard. In Experiment 2, we only used the hard version (Social Hard, Non-social Hard), to make the social manipulation more salient and to isolate the effects that are present in higher levels of environmental uncertainty. We measured the rate with which participants were able to select the arm with the highest reward probability (p(correct)), as well as their tendencies to win-switch (selecting a different arm despite receiving a reward on the preceding trial).

The PA task was administered uniformly to all groups. Participants viewed displays of eight disks for 200 trials (20 practice with feedback, 180 consecutive test trials) and had to indicate as fast as possible if they perceived any disk 'chasing' another one. If participants did not respond during the display, the question 'Chase or No chase' was presented on a blank screen. In the 180 test trials, they also had to indicate their confidence in their answer on a 1-5 scale. The practice block consisted of 10 chase-present and 10 chase-absent trials in random order, while the test block contained 90 chase-present and 90 chase-absent trials. Further details of this design are described in Study 2 of Castiello et al. (2024)

We excluded participants who either had reaction times of > 500 ms and < 8000 ms on less than 25% of their trials or scored below 55% accuracy when identifying chase-present and chase-absent trials. High confidence false alarms, that is, indicating chasing in chase-absent trials with a confidence level above three, were then used in the statistical analysis as a measure of social hallucinations.

Participants

All participants were recruited from Prolific, completed the tasks online, and were compensated with a flat rate of $8 for partaking. For both experiments we pre-screened for high (N =200) and low levels of paranoia (N = 200). Paranoia group was determined based on whether participants scored above 10 on the persecution subscale of the R-GPTS questionnaire (Freeman et al., 2019) - which is the cutoff value for elevated levels of persecutory thoughts. All participants were UK or US residents with no additional exclusion criteria. The study was approved by the Institutional Review Board (IRB) of Yale University, and all participants provided informed consent when commencing the study.

Questionnaires

Revised Green et al. Paranoid Thought Scale (R-GPTS) The R-GPTS measures paranoid ideation across two subscales: referential (10 questions) and persecutory (8 questions). Participants indicate their responses to statements based on a 5-point Likert scale that ranges from 0-Not at all to 4-Totally. Paranoia group membership was determined based on persecutory subscale scores, with 10 being the cutoff value for high paranoia and indicating moderately severe levels of persecutory thoughts (Freeman et al., 2019).

Computational modelling

We fit a Hierarchical Gaussian Filter model to the second block of PRL data using the tapas package in the openly available HGF Toolbox v7.1.2 (Mathys et al., 2011, 2014) and ran in Matlab 2024b. We analysed parameters fitted on the second block only, as this is where the key manipulations between stable and volatile conditions show up.

A three-level mean-reverting perceptual model was used in combination with a softmax decision model (that is set based on third-level volatility), due to it being the best fitting model of PRL data across multiple previous studies (Reed et al., 2020; Suthaharan et al., 2021).

Following previous findings (Reed et al., 2020; Suthaharan et al., 2021), our variables of interest were the prior beliefs about third-level volatility (μ_3), expected uncertainty that is used to update beliefs about the value of options (ω_2), meta-volatility that is used to update beliefs about tonic volatility (ω_3) and the influence of unexpected uncertainty on updating stimulus-outcome associations (κ_2). The modelling was done following the procedure and settings from Suthaharan et al. (2021).

Statistical analysis

All statistical analysis was conducted in R, using the wizaRdry package for pre-processing, car, emmeans and effectsize packages for the ANOVA, lme4 and lmerTest for the regression models and ggplot2 for plotting (Bates et al., 2025; Ben-Shachar et al., 2020; Fox et al., 2024; Kenney et al., 2025; Kuznetsova et al., 2026; Lenth & Piaskowski, 2017; Wickham et al., 2025).

The effects of PRL task type, task difficulty, and paranoia group membership on task performance were assessed with ANOVAs. In Experiment 1, this was a 2x2x2 ANOVA with two levels of task type (social, non-social), task difficulty (easy, hard) and paranoia group membership (high, low). In Experiment 2, we used a 2x2 ANOVA with only task type and paranoia group membership. The dependent variables were p(correct), win-switching and fitted HGF parameters in the PRL task, as well as high confidence false alarms in the PA task. Post-hoc tests on interactions were conducted using both visual assessment and comparison of estimated marginal means using a t-test.

To see how paranoia, win-switching, high confidence false alarms and HGF parameters relate to each other, we conducted logistic (paranoia group) and linear (win-switching, high confidence false alarms) regressions on the z-scored predictors. Post-hoc tests on interactions were conducted using both visual assessment and a pairwise comparison of estimated marginal means using a t-test with Tukey's HSD to correct for multiple comparisons.

To confirm whether behavioural biases and HGF parameters were unique characteristics of paranoia, we ran a logistic regression predicting paranoia group membership (high/low) from win-switch rates, high confidence false alarms, μ_3, ω_2, ω_3 and κ_2.We checked whether the significant predictors have unique contributions to the variance in paranoia group membership by running models with only behavioural predictors (win-switching, high confidence false alarms) and with only HGF parameters and comparing model fit.

Win-switch behaviour and high confidence false alarms were predicted from z-scored HGF parameters, paranoia group membership, and win-switch rates (only for the FA analysis) using multiple linear regression. Interaction effects with task type and task difficulty were also included in the full model, and model comparisons were performed between the interaction and no interaction models to trade model fit and complexity. Comparisons were performed using the AIC criterion, with lower AIC score indicating a better model. The best fitting models that were used in the analysis are the following: HGF predictors and interactions with paranoia and task type for Experiment 1 win-switching (1), HGF predictors and paranoia for Experiment 1 win-switching (2), full model for Experiment 1 high confidence false alarms (3), HGF parameters, win-switching and paranoia for Experiment 2 high confidence false alarms (4).

Results

Paranoia leads to characteristic behavioural biases in both the PRL and PA tasks

Task performance on both the PRL and PA tasks is depicted in Figure 2. Participants with high levels of paranoia displayed significantly higher rates of win-switching (Exp. 1: $F_{1,298} = 6.89$, $p = 0.009$, $\eta p^2 = 0.02$; Exp. 2: $F_{1,285} = 7.33$, $p = 0.007$, $\eta p^2 = 0.03$). Moreover, in Experiment 1, there was a paranoia x task type interaction effect ($F_{1,298} = 4.84$, $p = 0.03$, $\eta p^2 = 0.02$), with low paranoia participants win-switching significantly more in the social task, compared to the non-social task ($t(298) = 3.43$, $p < 0.001$).

The rate of p(correct) was not significantly different between paranoia groups (all $p > 0.07$) but was affected by overall task difficulty (Exp. 1: $F_{1,298} = 4.70$, $p = 0.031$, $\eta p^2 = 0.02$) and task type. In Experiment 1 participants in the low paranoia group had an impaired performance on the social task compared to the non-social (($F_{1,298} = 5.14$, $p = 0.024$, $\eta p^2 = 0.02$, $t(298) = -2.42$, $p = 0.016$), while in Experiment 2 — with a more salient social manipulation — the whole sample showed a decrease in performance compared to the non-social task ($F_{1,285} = 5.07$, $p = 0.025$, $\eta p^2 = 0.02$). While there seems to be a difference between how effectively participants make use of social and non-social cues, the win-switching effect is a strong indicator of high paranoia uniformly across the social and non-social variants of the task.

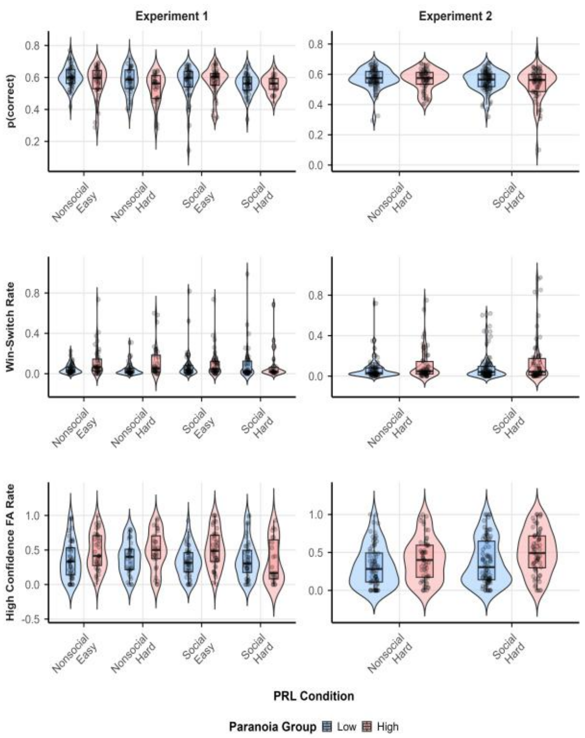

Figure 2: Box and violin plots of PRL performance and PA high confidence false alarms across both experiments. Data is plotted by PRL pre-training condition and paranoia group; boxes show sample median and the interquartile range.

In the PA task, high paranoia participants exhibited significantly elevated rates of high-confidence false alarms (Exp. 1: $F_{1,298} = 12.70$, $p < 0.001$, $\eta p^2 = 0.04$; Exp. 2: $F_{1,285} = 9.80$, $p = 0.002$, $\eta p^2 = 0.03$), with rates also elevated following the social PRL condition in Experiment 2 ($F_{1,285} = 4.49$, $p = 0.035$, $\eta p^2 = 0.02$). Both high confidence false alarms (Exp. 1: $\beta = 0.39$, $p = 0.001$, Exp. 2: $\beta = 0.31$, $p = 0.013$) and win-switching (Exp. 1: $\beta = 0.28$, $p = 0.031$, Exp. 2: $\beta = 0.26$, $p = 0.050$) were significant predictors of paranoia group membership in the logistic regression, although, as expected, win-switching did not have a contribution unique from the fitted HGF parameters.

Win-switching predicts social hallucinations

In Experiment 1, the best fitting regression model to predict high confidence false alarms included interaction effects from paranoia group, task type, and task difficulty. There was a significant win-switch x task type x task difficulty interaction ($\beta = 0.41$, $p = 0.044$) as a predictor of high confidence false alarms. This interaction is dissected through looking at how the slope values (β) of win-switching predicting high confidence false alarms change when broken down by task conditions as this was our key effect of interest. None of the post-hoc comparisons survived the corrections for multiple comparisons (all $p > 0.200$). Nevertheless, descriptive assessment of the interactions suggests that in the hard task, the predictive relationship of win-switching is mostly positive in the social condition ($\beta = 0.02$, 95% CI [-0.12, 0.16]), while completely negative in the nonsocial condition ($\beta = -0.15$, 95% CI [-0.28, -0.01]).

In Experiment 2, the best fitting regression model did not include task type interactions, and there was a significant positive main effect of both paranoia group membership ($\beta = 0.09$, $p = 0.010$) and win-switch rates ($\beta = 0.12$, $p = 0.003$), when predicting high confidence false alarms.

Win-switching and high confidence false alarms are linked through volatility priors or meta-volatility, depending on the salient manipulation

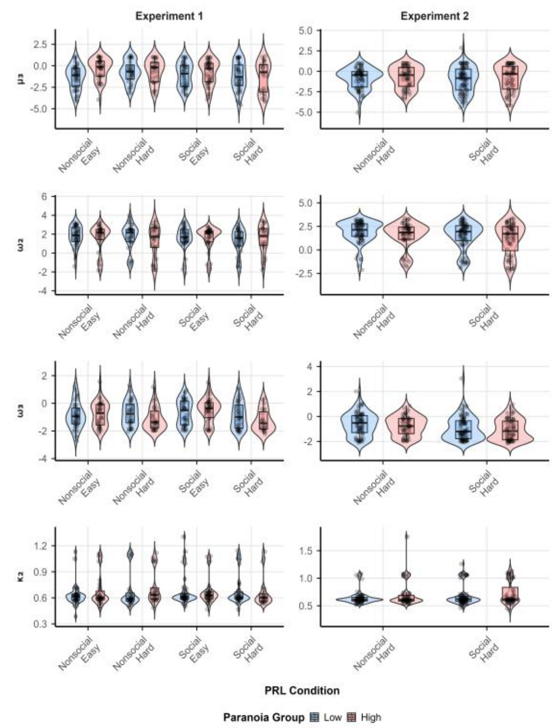

Figure 3: Box and violin plots of HGF parameters of interest across both experiments. Parameters are fitted to Block 2 of the PRL task.

Behaviour on the PRL task was modelled with the Hierarchical Gaussian Filter and we found that when volatility manipulation was salient (Experiment 1), prior beliefs about volatility (μ_3) were significantly higher for participants with high paranoia ($F_{1,298} = 5.44$, $p = 0.02$, $\eta p^2 = 0.02$), while meta-volatility (ω_3) was significantly lower in the hard task ($F_{1,298} = 5.57$, $p = 0.02$, $\eta p^2 = 0.02$) and within the hard task, it was lower for participants in the high paranoia group ($F_{1,298}$ = 5.78, p = 0.02, ηp^2 = 0.02, t(298) = -2.44, p = 0.015).

When social manipulation was salient (Experiment 2), ω_3 was significantly lower in the social task ($F_{1,285}$ = 9.15, p = 0.003, ηp^2 = 0.03), with no significant paranoia effect (p = 0.267). The effect of paranoia also did not show in μ_3 (p = 0.359). However, expected uncertainty (ω_2) was lower in the social task ($F_{1,285}$ = 5.09, p = 0.025, ηp^2 = 0.02) and in the high paranoia group ($F_{1,285}$ = 3.80, p = 0.052, ηp^2 = 0.01) and the influence of unexpected uncertainty on updating stimulus-outcome associations (κ_2) was higher in the high paranoia group ($F_{1,285}$ = 4.81, p = 0.029, ηp^2 = 0.02). HGF parameters for both experiments are depicted in Figure 3.

We linked the HGF parameters to elevated win-switch rates by findings that across both experiments, higher μ_3 predicted higher win-switching (Exp. 1: β = 0.05, p = 0.026; Exp. 2: β = 0.10, p < 0.001). This parameter also linked win-switching to paranoia by predicting paranoia group membership (β = 0.58, p = 0.009) in Experiment 1 and capturing the variance explained by win-switching. Task type also interacted with μ_3 when predicting win-switching (Exp.1: β = 0.07, p < 0.001), an effect that disappeared once task types were salient (in Experiment 2 task type was not included in the best fitting model). In Experiment 2, higher ω_2 and κ_2 also predicted less win switching (β = -0.12, p < 0.001; β = -0.08, p < 0.001), but these parameters were not predictive of paranoia in either experiment (all p > 0.06).

To see whether the μ_3 parameter linking win-switch behaviour to paranoia also provided links to high confidence false alarms, we ran linear regressions. In Experiment 1, μ_3 had 3-way interactions with paranoia group and difficulty (β = -0.36, p = 0.026), paranoia group and task type (β = -0.46, p = 0.005), as well as a four-way interaction with paranoia group, task type and task difficulty (β = 1.03, p < 0.001). In the hard version of the PRL task, paranoia groups showed opposite μ_3 slopes in nonsocial (β_{low} = 0.09, β_{high} = -0.18, t(249) = -0.28, p = 0.021) versus social (β_{low} = -0.11, β_{high} = 0.19, t(249) = 0.30, p = 0.106) conditions. Within high paranoia participants, μ_3 effects in social contexts reversed from strongly negative in the easy task (β = -0.33) to positive in the hard task (β = 0.19, t(249) = 0.52, p = 0.009).

On the contrary, in Experiment 2, μ_3 did not significantly predict high confidence false alarms (p = 0.913), rather, it was ω_3 in an interaction with paranoia group membership, that predicted high confidence false alarms (β = 0.09, p = 0.040). In the high paranoia group, ω_3 positively predicted high confidence false alarms (β = 0.06), while in the low paranoia group, this relationship was negative (β = -0.03, t(285) = -0.09, p = 0.04).

Discussion

Theoretical debates have long centered on whether paranoia represents primarily a social disorder or a domain-general impairment that happens to manifest in social contexts. Some accounts emphasize paranoia's fundamentally social nature, focusing on mentalizing deficits, theory of mind impairments, and disturbances in processing social threats (Barnby et al., 2023; Bell et al., 2021; Raihani & Bell, 2018). Others highlight domain-general probabilistic reasoning deficits that affect learning across contexts and argue that these are sufficient to induce rich social effects (Corlett et al., 2025; Feeney et al., 2017; Sheffield et al., 2022). Recent work has identified a phenomenon of social hallucinations (Castiello et al., 2024), where individuals with elevated paranoia inappropriately attribute agency to ambiguous visual motion patterns, raising a fundamental question: Are these low-level perceptual phenomena a manifestation of domain-general computational deficits applied to social contexts, or do they reflect distinct mechanisms specific to social cognition? To address this question, we conducted two experiments manipulating both volatility (stable vs. volatile) and domain (social vs. non-social) in a probabilistic reversal learning (PRL) task, followed by assessment of social hallucinations through high-confidence false alarms in a perceived animacy (PA) task. By combining behavioural measures with Hierarchical Gaussian Filter computational modeling, we aimed to identify the specific mechanisms linking domain-general belief updating deficits to social hallucinations in paranoid thinking.

Our findings revealed a clear link between domain-general belief updating deficits and social hallucinations. Participants with high paranoia displayed elevated win-switching rates regardless of whether the learning task involved social or non-social information, confirming well-established domain-general deficits (Barnby et al., 2022; Reed et al., 2020; Sheffield et al., 2022; Waltz, 2017). Similarly, high-confidence false alarms were elevated in high paranoia participants. These effects were further amplified following social PRL conditions, when the social manipulation was salient (Experiment 2). Critically, win-switching predicted high-confidence false alarms across both experiments, establishing that despite their different sensitivity to social context, these processes are computationally linked.

The computational modeling results suggested that the mechanisms linking belief updating to social hallucinations varied systematically depending on which dimension of uncertainty was most salient. In Experiment 1, where volatility manipulation was primary, prior beliefs about environmental volatility (μ_3) connected win-switching, paranoia, and social hallucinations, with high paranoia participants showing elevated μ_3 that predicted increased win-switching as well as high confidence false alarms. However, in Experiment 2, when social context was the salient manipulation, the computational architecture shifted entirely: μ_3 no longer predicted high-confidence false alarms, and instead, meta-volatility learning (ω_3) emerged as the critical mechanism linking paranoia to social hallucinations. This shift demonstrates that social hallucinations are overall driven by domain-general deficits in volatility beliefs but emerge through different computational pathways depending on the source of uncertain information.

The HGF parameters μ_3 and ω_3 that emerged in our study map directly onto well-established cognitive biases paranoia as well as previous computational modelling results. Prior beliefs about environmental volatility (μ_3) represent baseline expectations about how frequently the environment changes, with elevated μ_3 reflecting beliefs in a highly unstable world. This contributes to jumping to conclusions because expecting constant change creates urgency to act before circumstances shift, manifesting behaviourally as increased win-switching even following successful outcomes (Cole et al., 2020; Reed et al., 2020). Meta-volatility (ω_3) governs how quickly individuals update beliefs about environmental volatility itself, with lower ω_3 indicating slower learning about volatility dynamics and maintaining rigid beliefs about threat despite disconfirmatory evidence (Reed et al., 2020; Sheffield et al., 2022; Sloan et al., 2025). The context-dependent shift between μ_3 and ω_3 when predicting high confidence false alarms reveals that different layers of the hierarchical uncertainty processing system become critical depending on which aspect of the environment is the most salient. In Experiment 1, where both volatility and social context varied, volatility manipulation emerged as the dominant factor, with elevated priors about environmental instability (μ_3) driving aberrant agency attribution. However, in Experiment 2, where all participants experienced volatile environments, but social context varied, the salience of the social manipulation shifted the computational architecture. This meant that impaired capacity to learn about volatility dynamics (ω_3) became the primary mechanism linking paranoia to social hallucinations, while μ_3 no longer played a predictive role.

Overall, the consistent predictive role of HGF parameters across both experiments supports the non-social, domain-general account; however, the context-dependent shift in key parameters suggests a more integrated perspective. We showed that while belief updating impairments (win-switching) operate in a domain-general fashion, social hallucinations depend critically on the source and structure of environmental information. The dissociation between win-switching (invariant to social context) and high-confidence false alarms (amplified by social context) suggests that domain-general computational deficits cascade from higher-level cognitive belief updating processes downstream to shape lower-level perceptual interpretations. Similar dissociations between cognitive and perceptual-level belief updating have been documented in paranoia (Stuke et al., 2019), hierarchical model where domain-general impairments in volatility processing flow through the cognitive architecture to produce social hallucinations at the perceptual level, with the specific pathway modulated by environmental context and task demands.

This hierarchical cascade can also explain how domain-general deficits can produce seemingly social-specific symptoms. Exaggerated beliefs that the environment is volatile—operating across all contexts—create a general tendency to rapidly update beliefs with minimal evidence (jumping to conclusions). When this domain-general predisposition encounters social stimuli at lower perceptual levels, it manifests as aberrant agency attribution to ambiguous visual patterns, producing social hallucinations without requiring specialized social cognition impairments. These results also have practical implications, as effective interventions may need to target multiple levels of the cognitive hierarchy. Previous work has demonstrated successful treatment outcomes by reducing volatility priors (Sheffield et al., 2025). Our findings suggest such approaches may be particularly effective when volatility is the salient source of uncertainty, but interventions addressing social contexts may additionally need to target processes related to meta-volatility, to also affect lower levels of the cognitive hierarchy.

The effects shown in this study, however, have some limitations that warrant consideration. Many interaction effects with task type that were observed in Experiment 1 (such as task type × HGF parameter interactions when predicting high-confidence false alarms, task type × paranoia effects on win-switching and accuracy) disappeared in Experiment 2. This suggests that these were relatively weak effects that failed to replicate with the larger sample size and more uniformly difficult task design in Experiment 2. This raises questions about the independent influence of social manipulations and suggesting, that for some effects social information may exert meaningful effects primarily when interacting with other environmental features like volatility.

Additionally, in Experiment 1, we did not find a significant main effect of win-switching on high-confidence false alarms (only interaction effects). However, the lack of win-switching main effect could reflect that its variance was captured by significantly predictive HGF parameters in the same regression model. Importantly, this finding also does not negate the domain-general basis of social hallucinations: HGF parameters consistently predicted social hallucinations across both experiments and provide a more mechanistically precise account of the underlying computational processes, underscoring the robustness of domain-general computational mechanisms in explaining both phenomena.

Our findings link social hallucinations in paranoia to domain-general belief updating deficits. Critically, these deficits cascade from higher cognitive levels down to perceptual-level social hallucinations, with the specific pathway depending on environmental context: volatility priors drive social hallucinations when environmental uncertainty is salient, while meta-volatility learning becomes critical when social context dominates. These results support a domain-general account of paranoia while revealing hierarchical vulnerability in how uncertainty processing shapes perception, with implications for targeted interventions addressing multiple levels of the cognitive hierarchy.

Acknowledgments

This work was undertaken with funding from the Experimental Psychology Society that supported KJ visiting Yale University. KJ is funded by the Medical Research Council under the reference MR/W006774/1

References

Barnby, J. M., Dayan, P., & Bell, V. (2023). Formalising social representation to explain psychiatric symptoms. *Trends in Cognitive Sciences*, *27*(3), 317–332. https://doi.org/10.1016/J.TICS.2022.12.004

Barnby, J. M., Mehta, M. A., & Moutoussis, M. (2022). *The computational relationship between reinforcement learning, social inference, and paranoia*. https://doi.org/10.1371/journal.pcbi.1010326

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2025). Linear Mixed-Effects Models using 'Eigen' and S4 [R package lme4 version 1.1-38]. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.PACKAGE.LME4

Bell, V., Raihani, N., & Wilkinson, S. (2021). Derationalizing Delusions. *Clinical Psychological Science*, *9*(1), 24–37. https://doi.org/10.1177/2167702620951553;PAGEGROUP:STRING:PUBLICATION

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/JOSS.02815

Castiello, S., Danielle, J., Ongchoco, K., Van Buren, B., Scholl, B. J., & Corlett, P. R. (2024). Paranoid and teleological thinking give rise to distinct social hallucinations in vision. *Communications Psychology 2024 2:1*, *2*(1), 117-. https://doi.org/10.1038/s44271-024-00163-9

Cole, D. M., Diaconescu, A. O., Pfeiffer, U. J., Brodersen, K. H., Mathys, C. D., Julkowski, D., Ruhrmann, S., Schilbach, L., Tittgemeyer, M., Vogeley, K., & Stephan, K. E. (2020). Atypical processing of uncertainty in individuals at risk for psychosis. *NeuroImage: Clinical*, *26*, 102239. https://doi.org/10.1016/J.NICL.2020.102239

Corlett, P., Rossi-Goldthorpe, R., Suthaharan, P., Sheffield, J. M., de Obeso, S. C., & Heyes, C. (2025). Pseudosocial cognition and paranoia. *Trends in Cognitive Sciences*, *29*(11), 997–1006. https://doi.org/10.1016/j.tics.2025.05.019

Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., & Stephan, K. E. (2020). Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, *129*(6), 556–569. https://doi.org/10.1037/ABN0000500

Feeney, E. J., Groman, S. M., Taylor, J. R., & Corlett, P. R. (2017). Explaining Delusions: Reducing Uncertainty Through Basic and Computational Neuroscience. *Schizophrenia Bulletin*, *43*(2), 263–272. https://doi.org/10.1093/SCHBUL/SBW194

Fett, A. K. J., Mouchlianitis, E., Gromann, P. M., Vanes, L., Shergill, S. S., & Krabbendam, L. (2019). The neural mechanisms of social reward in early psychosis. *Social Cognitive and Affective Neuroscience*, *14*(8), 861–870. https://doi.org/10.1093/SCAN/NSZ058

Fox, J., Weisberg, S., & Price, B. (2024). Companion to Applied Regression [R package car version 3.1-3]. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.PACKAGE.CAR

Freeman, D., Loe, B. S., Kingdon, D., Startup, H., Molodynski, A., Rosebrock, L., Brown, P., Sheaves, B., Waite, F., & Bird, J. C. (2019). The revised Green et al., Paranoid Thoughts Scale (R-GPTS): psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine*, *51*(2), 244. https://doi.org/10.1017/S0033291719003155

Freeman, D., McManus, S., Brugha, T., Meltzer, H., Jenkins, R., & Bebbington, P. (2011). Concomitants of paranoia in the general population. *Psychological Medicine*, *41*(5), 923–936. https://doi.org/10.1017/S0033291710001546

Garety, P., Freeman, D., Jolley, S., Ross, K., Waller, H., & Dunn, G. (2011). Jumping to conclusions: the psychology of delusional reasoning. *Advances in Psychiatric Treatment*, *17*(5), 332–339. https://doi.org/10.1192/APT.BP.109.007104

Gibbs-Dean, T., Katthagen, T., Tsenkova, I., Ali, R., Liang, X., Spencer, T., & Diederen, K. (2023). Belief updating in psychosis, depression and anxiety disorders: A systematic review across computational modelling approaches. *Neuroscience & Biobehavioral Reviews*, *147*, 105087. https://doi.org/10.1016/J.NEUBIOREV.2023.105087

Hauke, D. J., Wobmann, M., Andreou, C., Mackintosh, A. J., DE BOCK, R., Karvelis, P., Adams, R. A., Sterzer, P., Borgwardt, S., Roth, V., & Diaconescu, A. O. (2024). Altered Perception of Environmental Volatility During Social Learning in Emerging Psychosis. *Computational Psychiatry*, *8*(1), 1–22. https://doi.org/10.5334/CPSY.95

Jedlovszky, K., Corlett, P. R., & Yon, D. (2024). *Subjective volatility, learning and paranoia*. https://doi.org/10.31234/OSF.IO/SRE9Y

Kenney, J. G., Williams, T. F., Pappu, M. K., & Spilka, M. J. (2025). wizaRdry: A Magical Framework for Collaborative & Reproducible Data Analysis. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.PACKAGE.WIZARDRY

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2026). Tests in Linear Mixed Effects Models [R package lmerTest version 3.2-0]. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.PACKAGE.LMERTEST

Lenth, R. V., & Piaskowski, J. (2017). emmeans: Estimated Marginal Means, aka Least-Squares Means. *CRAN: Contributed*

Packages. https://doi.org/10.32614/CRAN.PACKAGE.EMMEANS

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*. https://doi.org/10.3389/FNHUM.2014.00825

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*(MAY), 9. https://doi.org/10.3389/FNHUM.2011.00039

Prike, T., Arnold, M. M., & Williamson, P. (2018). The relationship between anomalistic belief and biases of evidence integration and jumping to conclusions. *Acta Psychologica*, *190*, 217–227. https://doi.org/10.1016/J.ACTPSY.2018.08.006

Raihani, N. J., & Bell, V. (2018). An evolutionary perspective on paranoia. *Nature Human Behaviour*, *3*(2), 114. https://doi.org/10.1038/S41562-018-0495-0

Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *ELife*, *9*, e56345. https://doi.org/10.7554/ELIFE.56345

Sheffield, J. M., Sloan, A. F., Corlett, P. R., Rogers, B. P., Vandekar, S., Liu, J., Beals, K. M., Hall, L. M., Gautier, T., Moussa-Tooks, A. B., Torregrossa, L. J., Achee, M., Armstrong, K., Woodward, N. D., Belt, K., Freeman, D., Isham, L., Diamond, R., Brinen, A. P., & Heckers, S. (2025). Prior Expectations of Volatility Following Psychotherapy for Delusions: A Randomized Clinical Trial. *JAMA Network Open*, *8*(6), e2517132–e2517132. https://doi.org/10.1001/JAMANETWORKOPEN.2025.17132

Sheffield, J. M., Suthaharan, P., Leptourgos, P., & Corlett, P. R. (2022). Belief Updating and Paranoia in Individuals With Schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(11), 1149–1157. https://doi.org/10.1016/J.BPSC.2022.03.013

Sloan, A. F., Kittleson, A. R., Torregrossa, L. J., Feola, B., Rossi-Goldthorpe, R., Corlett, P. R., & Sheffield, J. M. (2025). Belief Updating, Childhood Maltreatment, and Paranoia in Schizophrenia-Spectrum Disorders. *Schizophrenia Bulletin*, *51*(3), 646–657. https://doi.org/10.1093/SCHBUL/SBAE057

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*(1). https://doi.org/10.1093/SCHBUL/SBX189

Stuke, Heiner, Stuke, Hannes, Weilnhammer, V. A., & Schmack, K. (2017). Psychotic Experiences and Overhasty Inferences Are Related to Maladaptive Learning. *PLoS Computational Biology*, *13*(1). https://doi.org/10.1371/JOURNAL.PCBI.1005328

Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., Litman, L., Robinson, J., Moss, A. J., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2021). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour 2021 5:9*, *5*(9), 1190–1202. https://doi.org/10.1038/s41562-021-01176-8

Waltz, J. A. (2017). The neural underpinnings of cognitive flexibility and their disruption in psychotic illness. *Neuroscience*, *345*, 203–217. https://doi.org/10.1016/j.neuroscience.2016.06.005

Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E. A., Ullrich, J., & Stephan, K. E. (2020). Inflexible social inference in individuals with subclinical persecutory delusional tendencies. *Schizophrenia Research*, *215*, 344–351. https://doi.org/10.1016/j.schres.2019.08.031

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & van den Brand, T. (2025). Create Elegant Data Visualisations Using the Grammar of Graphics [R package ggplot2 version 4.0.1]. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.PACKAGE.GGPLOT2