Comparison of social prediction error signals obtained from fMRI data during a trust game in patients with schizophrenia and healthy controls

Schizophrenia is a complex and multifaceted mental disorder characterized by a range of cognitive deficits. Cognitive impairments, including deficits in memory, attention, and executive function precede the clinical diagnosis of schizophrenia, highlighting their role as core features of the disorder (Guo et al., 2018; Kochunov et al., 2017). Social cognitive dysfunction, a subset of cognitive impairment, refers to difficulties in understanding and processing social information, which can manifest as challenges in recognizing emotions, understanding social cues, managing complex tasks, and engaging in effective communication (Caqueo-Urízar et al., 2015; , Lesh et al., 2010). Social cognition is of great importance in adapting to an environment where there are many stimuli and changes. Difficulties in social functioning lead to social isolation and interpersonal problems exacerbate feelings of loneliness and depression. This further compounds the overall impact of the disorder on quality of life, causing relapses, significantly impairing quality of life, and negatively affecting rehabilitation processes (46, 47, Raykeer et al., 2019).

Components of social cognition include emotional face recognition, emotion recognition in social context, emotional awareness, theory of mind, and empathy skills (67). In the last meeting of the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) conducted by NIMH, it was suggested that social cognition be examined in 5 main areas. These areas are; 1. emotion processing, 2. social perception, 3. social knowledge, 4. attribution bias and 5. theory of mind (30). Due to the complexity of social cognition, it is difficult to assess it with clinical interviews or scales. To solve this problem, researchers use neuroeconomics games that are based on game theory. Interpersonal interactions can be simulated in a laboratory environment with these games. Then, social decision making can be experimentally investigated and the underlying neuroscientific mechanisms can be elucidated (68).

The trust game is one of the most studied neuroeconomics games for assessing social cognition (Berg, J., Dickhaut, J., & McCabe, K. (1995). This game typically involves two players: an investor and a trustee.The investor decides to invent, give an endowment to the trustee, or keep it himself. If the investor invests the money, the amount sent is usually tripled. The trustee then decides to share the money with an investor or keep it all to himself.

In a study of patients with schizophrenia, patients with schizophrenia invested less than controls during play with a cooperative partner, but did not differ when playing with a deceptive partner. These findings suggest that they are impaired in adapting to changing environmental conditions, have reduced behavioral flexibility after negative feedback, and develop poor strategies. Neuroimaging studies utilizing functional magnetic resonance imaging (fMRI) have provided valuable insights into the neural correlates of social cognition.

During the anticipation phase of the trust game in a safe environment, healthy individuals showed increased coupling between the right parietal cortex, the fusiform gyrus (FG), and the inferior/middle temporal gyrus. During cooperative responses in trust games, lower BOLD signals were observed in the right TPJ in patients with schizophrenia. The temporoparietal junction (TPJ) plays a crucial role in processes such as empathy, theory of mind, social decision-making tasks. (Fett et al., 2012; , 33, 34, 35). This diminished activation may reflect the underlying cognitive deficits that characterize the disorder, such as impaired theory of mind and difficulties in processing social cues (Jáni & Kašpárek, 2018).

Social interaction and social functioning involve a large number of cognitive processes, including social perception, understanding the actions of others, observational social learning, and social decision making. In order to adapt to a rapidly changing world, the brain makes numerous predictions. Motor control, perceptual inference, and reward-based learning occur through constant updating of predictions based on error (55, 56, 57). This mechanism is called prediction error (PE). (paraphrase)Prediction error refers to mismatch between prior prediction and actual results. If the outcome is better than expected it's called positive PE and if the outcome is worse than expected it is called negative PE. Effective learning occurs by monitoring prediction errors (59). Being able to predict the consequences of our actions is essential for effective decision making. Prediction mechanisms play a role in social communication and interpersonal interactions, affecting daily functioning. The TPJ has been shown to play a role in social prediction errors (60). Studies have shown that patients with schizophrenia demonstrate deficits in learning from social feedback, which may be linked to abnormal activation in the TPJ and other related brain regions (Lemmers-Jansen et al., 2018). These neural disruptions can lead to maladaptive social behaviors, such as distrust and social withdrawal, which further exacerbate the challenges faced by individuals with schizophrenia in their daily lives.

There are many studies indicating that PE is impaired in schizophrenia. Deactivation was detected in the right superior frontal lobe and inferior frontal lobes during PE compared to healthy controls. Increased activation was seen in the PCC, which is part of the DMN, indicating impairment in the estimation of self-worth in the schizophrenia group (62). In negative prediction error, differences in mPFC and right middle frontal cortex activations have been observed (62, 63). In positive prediction error, impairments are manifested by deactivation in PFC, which may explain the avolition among negative findings (64). It has also been suggested that impaired PE calculations in schizophrenia may cause delusions (65, 66). According to the incentive salience hypothesis, increased firing of chaotic or stress-related dopaminergic afferents in the striatum of schizophrenia patients has been proposed to attribute salience to irrelevant stimuli. Over-attribution of meaning to irrelevant cues may affect thought content and mood, leading to perceptual distortions and the formation of delusions (62). This condition has also been thought to affect cognitive processes in schizophrenia patients by causing paradoxical learning (over-learning of irrelevant and neutral information and under-learning of rewarding events) (64).

Moreover, the role of dopamine in schizophrenia cannot be overlooked. The dopamine hypothesis suggests that dysregulation of dopaminergic pathways contributes to the cognitive and emotional symptoms observed in schizophrenia (Millard et al., 2021). Specifically, the TPJ's interaction with dopaminergic systems may influence social reward processing, which is critical for establishing trust and cooperation in social contexts (Fett et al., 2011)

The interplay between neural mechanisms, social cognition, and dopamine dysregulation highlights schizophrenia's multifaceted nature. The TPJ emerges as a critical region for understanding the social deficits associated with the disorder, particularly in the context of trust and cooperation. The neural underpinnings of social cognition in schizophrenia with PE contribute to a deeper understanding of the disorder. Our investigation aimed to compare the BOLD signal in the TPJ region of schizophrenia patients with healthy controls by creating a prediction error with the trust game. To achieve this, we designed a trust game with a reinforcement model that produced Negative PE at first and then Positive PE. We extract prediction errors from the response phases of the trust game and then link those with BOLD signals. By showing the statistical significance of this relationship, we desire to investigate the BOLD signals of ROIs that differed between groups. Specifically, we predicted that prediction error signals generated during the trust game will cause decreased BOLD signals in the temporoparietal junction in schizophrenia patients compared to the healthy control group.

# Methods

## Task Paradigm

In the proposed fMRI task, the experimental paradigm is based on the "Trust Game," a widely used game-theoretical model in the literature that captures interpersonal trust (REF). Subject have met with the trustee, as if the subject was playing with the trustee simultaneously under fMRI. The introduction with trustee was standardized to not to effect basal trust level.

## In each trial, the participant starts with an initial endowment of 40 Turkish Lira (TL) and must decide whether to keep the money or invest it by transferring it to a trustee. If the participant keeps the money, it is evenly split between them and the trustee. If the participant chooses to invest, the amount is tripled to 120 TL, and the trustee then decides whether to share this tripled amount with the participant or keep all of it. The central task involves the participant evaluating whether to trust the trustee in each round. The trust game consists of three blocks, each with 20 trials, making a total of 60 trials. Each trial follows a standardized sequence presented across six screens (See Figure 1A).

Each trial of the Trust Game followed a structured sequence of six consecutive screens, designed to capture the participant's decision-making and trust behavior under controlled experimental conditions (Figure 1B).

In the first decision or anticipation screen, participants saw a photo of the trustee alongside their initial endowment of 40 Turkish Lira (TL). They were instructed to mentally decide whether to keep the money or invest it by transferring it to the trustee. No action or response was needed during this phase, and the screen remained visible for 6 seconds.

On the second choice screen, participants were asked to confirm their earlier decision (“invest” or “keep”) by pressing the appropriate button within a 3-second window. Before starting the task, they were informed that failing to respond would result in zero earnings for that trial (0 TL) and that missing responses on five trials—whether in a row or not—would lead to the termination of the entire task.

In the third waiting screen, a 3-second waiting screen followed, during which participants anticipated the trustee's response.

In the fourth outcome screen showed the trustee’s decision—whether to share the tripled amount or keep it entirely—along with the outcome of the trial, including the participant’s earnings. This screen was displayed for 3 seconds. Participants were informed in advance that they were interacting with a computer algorithm simulating trustee behavior, and that the responses were pre-programmed to seem random

In the fifth jitter screen inter-trial interval followed, lasting randomly between 3 and 6 seconds. During this time, a blank screen was displayed to introduce temporal variability between trials.

In the last sixth fixation screen, to maintain a total inter-trial interval of 9 seconds, a fixation cross (“+”) was shown for the remaining time—ranging from 3 to 6 seconds—depending on the length of the preceding jittered interval.

Additionally, after every 10 trials, participants were prompted to rate their level of trust in the trustee using a scale from 1 (not at all) to 7 (completely). This rating screen remained visible for 9 seconds.

Each trial lasted 24 seconds, leading to a total task duration of approximately 24 minutes and 54 seconds.

Computational Learning Model Based on Rescorla-Wagner

We implemented a reinforcement learning (RL) model based on the Rescorla-Wagner rule to estimate participants’ trial-by-trial reward predictions and learning dynamics during a sequential investment task. The model was applied separately to three experimental blocks (phases, each with 20 trials) featuring varying gain probabilities (80%, 50%, and 80%).

On each trial , the predicted reward value was updated according to the standard delta rule:

Here, is the actual reward and is the trial-specific learning rate. Unlike classical fixed-𝛼 Rescorla-Wagner models, we introduced an adaptive learning rate mechanism that updated 𝛼 based on the social partner’s behavior:

1. If the participant invested and the trustee shared:
2. If the participant invested and trustee keeps:
3. If the participant did not invest:

## fMRI Analysis

### fMRI Preprocessing

### First Level

### Second Level

Second-level statistical analyses were conducted using SPM12 to assess within-group and between-group effects of prediction error (PE)–related activation. First-level contrast images representing trial-by-trial PE modulation were entered into second-level models for group-level inference.

For the within-group one-sample t-tests were conducted separately for the HC and SZ groups to identify brain regions showing significant PE-related activation within each population. For the HC group, sex was included as a covariate. For the SZ group, both sex and duration of illness (DoI) were included as covariates to account for individual differences that may influence PE-related brain responses. Covariates were mean-centered prior to inclusion. All second-level models were estimated using classical (maximum likelihood) estimation, and results were corrected for multiple comparisons at the cluster level (FWE-corrected p < 0.05) using a cluster-forming threshold of p < 0.001 uncorrected.

Additionally, between-group analysis, a two-sample t-test was used to compare individuals with schizophrenia (SZ) to healthy controls (HC). To control for confounding variables, sex (coded as 0 = female, 1 = male) was included as a covariate of no interest. Contrast images for each subject were grouped accordingly, and individuals with missing or incomplete data were excluded from the analysis. Group comparisons (e.g., SZ > HC and HC > SZ) were tested while accounting for sex-related variance.

## Prediction Error Acquisition

## Hypothesis Testing

### Binomial Generalized Linear Mixed-Effects Model for Investment Decisions

We analyzed binary investment responses using a generalized linear mixed-effects model (GLMM) with a binomial distribution and a logit link function, implemented in the lme4 package in R. The aim was to investigate the effects of experimental group and task phase on the likelihood of making an investment decision, while accounting for repeated measures within participants. The model included fixed effects for group (two levels: 0 = reference, 1) and task phase (three levels: phase 1 = reference, phase 2, and phase 3). Participant-level variability was modeled with a random intercept for subject.

Here, the term is the probability that the participant on trial makes a positive response (e.g invests). ​ follows the assumption that the response variable is Bernoulli distributed, with success probability defined by the fixed and random effects through a logistic link. is the fixed intercept, where is fixed effect for Group (patient vs control) and indicates effect of Task phase relative to first phase as reference. Sex and Duration of Illness (DOI) included in the model, with sex is a categorical factor DOI is a continuous variable. denoting random intercept for subject with denoting residual error.

Model diagnostics were performed using the DHARMa package (ref). Simulated residuals showed no significant deviation from uniformity (KS test: p = 0.77), no overdispersion (p = 0.96), and no evidence of outlier inflation (p = 0.19). Residuals were homogeneously distributed across the range of predicted values, indicating that model assumptions were adequately met (see Supplementary Material). Multicollinearity among fixed effects was assessed using generalized variance inflation factors (GVIF). All adjusted GVIF values were well below 2 (range: 1.00–1.46), indicating no evidence of multicollinearity among predictors.

### The intraclass correlation coefficient (ICC) was calculated to assess the proportion of variance in investment decisions attributable to between-subject differences. The ICC was approximately 0.167, indicating that 16.7% of the variance was due to differences between participants, thereby justifying the inclusion of a random intercept in the model.

### Lognormal Linear Mixed-Effects Model for Prediction Errors

Because prediction error (PE) values included zero or negative values, we applied a constant shift to ensure all observations were strictly positive. This transformation preserved the relative structure of the data while enabling the use of log-transformation. The log-transformed PE values were then modeled using a linear mixed-effects model, assuming a lognormal distribution. This approach is appropriate for positively skewed data with heteroscedasticity and allows for multiplicative interpretation of fixed effects. Residual diagnostics indicated no issues with dispersion or heteroscedasticity; however, tests for residual uniformity and outliers revealed minor deviations, suggesting the model may be sensitive to extreme values (see Supplementary Figure).

To account for repeated measures and individual differences in task, a random intercept for Subject was included in the linear mixed-effects model. As same with Binomial model, all adjusted GVIF values were well below 2 (range: 1.00–1.46), indicating no evidence of multicollinearity among predictors. The intraclass correlation coefficient (ICC = 0.053) indicated that approximately 5.3% of the variance in log-transformed prediction error was attributable to between-subject variability, supporting the inclusion of a random effect for Subject. Based on the data characteristics and model diagnostics, the final model was defined as:

Here, is the PE for subject in trial . is the shifted PE value, where is the constant added to make values strictly positive. follows the assumption that PE is lognormal distributed. is the fixed intercept, where is fixed effect for Group (patient vs control) and indicates effect of Task phase relative to first phase as reference. As with binomial model, Sex and Duration of Illness (DOI) included in the model, with sex is a categorical factor DOI is a continuous variable. denoting random intercept for subject with denoting residual error. Model implemented in R 4.3.3 (REF) with lme4 package (REF).

# Results

## Behavioral Results

A binomial generalized linear mixed-effects model (GLMM) with a logit link was used to examine the effects of group, sex, task phase, and duration of illness (DoI) on investment decisions, while accounting for repeated measures with a random intercept for participants . Results are reported as odds ratios (OR) with 95% confidence intervals (see Methods – Hypothesis Testing).

The schizophrenia (SZ) group showed no significant difference in investment behavior compared to healthy controls (HC) (OR = 0.85, p = .590) in any of the phases, and sex was also not a significant predictor (OR = 1.05, p = .820). However, investment odds were significantly lower in phase 2 compared to phase 1 (OR ≈ 0.76, 95% CI ≈ [0.65, 0.89], p < .001) for both groups, suggesting a reduction in cooperative behavior during this phase. No significant difference was observed between phase 3 and phase 1 (OR ≈ 1.03, p = .750). Duration of illness (DoI) was not significantly associated with investment likelihood (OR ≈ 0.99, p = .503).

These findings indicate that while group membership and sex did not significantly influence investment decisions, phase 2 consistently decreased the odds of investing across participants. There is no evidence that group or sex moderated this phase effect in the current model specification.

A graph with black lines and white text

AI-generated content may be incorrect.

## Prediction Error

A linear mixed-effects model was fitted to examine the effects of group, sex, task phase, and duration of illness (DoI) on prediction error (PE), using the natural log of PE as the dependent variable. The model included fixed effects for group, sex, task phase (with phase 1 as the reference), their interaction (group × sex), and duration of illness (DoI), and a random intercept to account for repeated measures across subjects (see Methods – Hypothesis Testing).

The schizophrenia group exhibited a multiplicative effect of 0.96 (95% CI ≈ [0.904, 1.026], p = .22) compared to healthy controls, indicating a non-significant 4% reduction in PE. Male participants had significantly higher PE than females, with a multiplicative effect of 1.08 (95% CI ≈ [1.028, 1.124], p = .001), reflecting an 8% increase. A significant interaction between group and sex was observed (multiplicative effect = 1.12, 95% CI ≈ [1.044, 1.207], p = .001), indicating that males with schizophrenia had 12% higher PE than expected from the additive effects of group and sex alone.

Compared to phase 1, other task phases had no significant effect on PE. Duration of illness was positively associated with PE, with a multiplicative effect of 1.003 (95% CI ≈ [1.000, 1.006], p<.05), reflecting a 0.3% increase in PE per unit increase in illness duration.

These findings suggest that prediction error is increased specifically in male patients with schizophrenia, rather than uniformly across all individuals with the disorder, and is not significantly modulated by task phase.

A graph with black lines and white text

AI-generated content may be incorrect.

## PE-related Activations

Anatomical locations were identified using the Neuromorphometrics atlas provided in SPM12. In the HC group, significant activation was observed in right middle and superior occipital gyrus (peak MNI: [30, -76, 11], T = 7.52, p < .05 FWE-corrected, cluster size = 28 voxels) and left superior parietal lobule and angular gyrus in response to PE controlling for sex (peak MNI: [-30, -64, 26], T = 7.46, p < .05 FWE-corrected, cluster size = 58 voxels).

However, in the SZ group significant activation was observed in right occipital fusiform and inferior occipital gyrus in response to PE controlling for sex and DoI (peak MNI: [27, -85, -7], T = 4.98, p < .001 uncorrected, cluster size = 48 voxels).

In the between-group analysis, the SZ group exhibited greater PE-related activation than the HC group in the left cerebellum exterior and occipital fusiform gyrus (peak MNI: [−18, −85, −25], T = 4.22, p < .001 uncorrected, cluster size = 26 voxels). No significant clusters were observed for the reverse contrast (HC > SZ).

In summary, the results indicate that the temporal-parietal junction (TPJ) shows increased PE-related activation in healthy controls, whereas this effect is absent in individuals with schizophrenia. In contrast, the schizophrenia group exhibited PE-related activation predominantly in lower-level sensory regions, including the occipital cortex and cerebellum. These findings may reflect a deficit in mentalizing processes in schizophrenia, whereby prediction errors are processed primarily in unimodal sensory regions rather than in higher-order transmodal areas such as the TPJ.

**Discussion**

This study examined behavioral and neural correlates of social prediction error (PE) processing during a trust game in individuals with schizophrenia (SZ) compared to healthy controls (HC). While behavioral investment patterns were largely comparable across groups, pronounced group differences emerged in PE processing and associated neural activations. These findings advance our knowledge of the computational and circuit-level mechanisms underlying social dysfunction in schizophrenia.

**Behavioral Adaptation and Trust Decisions**

*Contrary to several previous reports of diminished trust behavior in schizophrenia (Fett et al., 2012; Gromann et al., 2013), statistical results yielded no significant differences in investment across the task phases* While sex and duration of illness (DoI) did not predict trust, both SZ and HC participants reduced investment during Phase 2 suggesting a sensitivity to dynamic changes in social uncertainty. This result partially aligns with previous work in first-episode psychosis (FEP) and clinical high-risk populations, where trust adaptation remains preserved (Lemmers-Jansen et al., 2019; Wisman-van der Teen et al., 2022). However, the relatively high remission rates and low PANSS scores in our cohort, compared to the general schizophrenia population, may have attenuated behavioral manifestations.Although investment patterns appear equal, this equality may be achieved through different neural pathways. Our results indicate that under certain structured and repetitive task conditions, individuals with schizophrenia may exhibit behaviors similar to healthy controls in response to social feedback.

**Prediction Error Processing and Sex-Specific Modulation**

Despite behavioral parity, group differences emerged in the neural encoding of prediction errors. A key finding was a significant group × sex interaction, wherein male patients exhibited elevated PE signals relative to females and controls. This effect was not attributable to main effects of sex or group alone and may reflect increased volatility or aberrant salience attribution in male patients—a group disproportionately affected by early-onset SZ and severe symptom burden (Grossman et al., 2006; Waltz et al., 2018). That PE amplitude was positively associated with duration of illness further suggests that prolonged exposure to illness processes may alter PE computations, potentially reinforcing maladaptive belief updating.

# Supplementary Material

## Model Diagnostics

Binary Model

A comparison of a diagram

AI-generated content may be incorrect.

PE model

A diagram of a model

AI-generated content may be incorrect.