Supporting information: Additional measures and supporting analyses

Cognitive Bias and Theory of Mind additional measures

The overall aim of the main paper was to examine trust processing in schizophrenia, and specifically to examine how people with schizophrenia use facial appearance and actual fairness to guide trusting decisions. Capacity to learn partner actual fairness is likely underpinned by ability to utilise experience when reasoning, and theory of mind (ToM), skills that are impaired in schizophrenia (Langdon, 2005; Woodward, Moritz, Cuttler, & Whitman, 2006). As a supplementary analysis, we therefore measured cognitive reasoning (Moritz & Woodward, 2005; Woodward et al., 2006) and ToM (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Langdon, 2005) to ascertain any links between these and trust behaviour.

Cognitive biases: We measured Jumping-To-Conclusions (JTC) with the draws-to-decision measure from the Beads Task using a 85/15% beads ratio (Langdon, Ward, & Coltheart, 2010; Moritz & Woodward, 2005) and Belief-Against-Disconfirmatory-Evidence (BADE) bias using the pictured stories task (Woodward et al., 2006).

The clinical group showed a marginally significant trend towards higher JTC bias in the Beads Task (drawing fewer beads before making a decision) than the control group: $t(46) = 1.80$, $p = .079$, $d = 0.52$ (Table S1). On the BADE, the clinical group showed significantly less integration of disconfirmatory evidence relative to controls on revealed-on-second trials, $t(46) = 3.01$, $p = .004$, $d = 0.87$, and on revealed-on-third trials, $t(46) = 2.65$, $p = .011$, $d = 0.77$ (Table S1). Clinical and control groups did not differ in integration of confirmatory evidence on revealed-on-second trials, $t(46) = 0.90$, $p = .371$, $d = 0.26$, but the clinical group showed significantly less integration of confirmatory evidence than controls on revealed-on-third trials,

$t(46) = 2.14$, $p = .038$, $d = 0.62$. The clinical group were also more willing to endorse absurd interpretations than controls, $t(46) = 2.13$, $p = .039$, $d = 0.61$.

Table S1

Means and standard deviations for the clinical group (N = 24) and control group (N = 24), on cognitive bias measures (BADE and JTC). Clinical < control group, ** $p < .01$, * $p < .05$, $^{+}$ $p < .08$.

	BADE Disconfirmatory		BADE Confirmatory		BADE Absurd	JTC
	Revealed-on-second	Revealed-on-third	Revealed-on-second	Revealed-on-third		Decide
Clinical	17.8**	26.6*	40.5	55.4*	12.5*	2.9^{+}
	(21.1)	(18.9)	(24.8)	(20.1)	(12.6)	(1.8)
Control	33.4**	39.3*	46.2	68.6*	6.5*	4.3^{+}
	(14.2)	(13.8)	(18.7)	(22.5)	(5.6)	(3.4)

Theory-of-mind (ToM): We measured ToM using the False-belief/Deception Comprehension task, a revised version of the Sally-Anne task designed for use with adults (Langdon, Connors, & Connaughton, 2017). We also included the Reading the Mind in the Eyes task (RMET), which may be particularly relevant as participants are measured on their ability to infer intentions from faces (Baron-Cohen et al., 2001).

As predicted, the clinical group showed lower second-order ToM relative to controls in the False-belief/Deception Comprehension task: $t(46) = 4.23$, $p < .001$, $d = 1.22$ (Table S2). The clinical group showed no difference in first-order ToM: $t(46) = 1.33$, $p = .190$, $d = 0.38$, and no general comprehension problems on control questions, thus poorer ToM in the clinical group was not due to a comprehension failure: both $t(46) < 0.94$, both $p > .353$, both $d < 0.3$ (Table S2). There was no evidence of a group difference in the RMET, $t(46) = 1.45$, $p = .155$, $d = 0.42$ (Table S2).

Table S2

Means and standard deviations for the clinical group (N = 24) and control group (N = 24), on Theory of Mind measures (False-belief/Deception Comprehension task and RMET). Clinical < control group, ** $p < .01$, * $p < .05$, $^{+}$ $p < .08$.

	False-belief/Deception Comprehension task				RMET
	1st order ToM	**2nd order ToM**	**1st order Control**	**2nd order Control**	**Total score**
Clinical	4.0	3.1**	1.9	1.8	26.0
	(1.8)	(1.5)	(0.4)	(0.4)	(6.0)
Control	4.6	4.8**	1.8	1.9	28.1
	(1.4)	(1.3)	(0.5)	(0.3)	(3.9)

Which aspects of the disorder relate to Trust Game performance?

An interesting question is whether individual differences in other task and clinical measures associate with clinical participants' failure to distinguish between fair and unfair partners at the end of the Trust Game. Interestingly, there was a significant association between correctly selecting confirmatory statements on the BADE revealed-on-second trials, and higher experience-based trust discrimination (i.e. distinguishing between fair and unfair partners on the last trial) for the clinical participants, $r = .420$, $p = .041$, all $n = 24$. Higher tendency to endorse absurd explanations on the BADE was also marginally significantly associated with lower

experience-based trust discrimination: r = -.399, p = .054. No other associations with task measures approached significance, apart from matrix reasoning: r = .365, p = .080.

No associations between experience-based trust discrimination and clinical measures reached significance, although there was a near-significant negative correlation with the PDI paranoia score: r = -.392, p = .058.

Summary

In summary, the clinical group showed cognitive reasoning biases relative to the control group, including a poorer ability to update beliefs based on experience (replicating (Woodward et al., 2006). The clinical group also showed evidence of higher-order TOM impairment and higher paranoia, as expected by theory (Frith & Johnstone, 2003) and in line with other studies (Chan & Chen, 2011; Langdon, 2005). Performance in cognitive reasoning tasks and paranoia was linked to trust game performance in the clinical group, although not always significantly. Future research should confirm these links with a large-scale individual differences study.

Four-way Trust Game ANOVA

Our main aim was to examine how people with schizophrenia use facial appearance and/or experience with actual fairness to guide trusting decisions. Here we conducted a four-way mixed ANOVA with Group (clinical versus control) as a between-subjects factor, and Partner Appearance (trustworthy versus untrustworthy), Partner Behaviour (fair or unfair) and Time (first versus last block) as within-subjects factors (**figure S1**). We could compare the first and last block directly because these trials were identical. Note that four participants had to be excluded from the overall analysis due to missing trials (three controls and one patient). In the main paper, we examined appearance-based trust and experience-based trust separately, for simplicity, because these analyses allowed us to use the whole dataset and because there is no equivalent four-way ANOVA for the slot machine game (see **figure S1**). Note that there is no difference in conclusions between the main paper and the current supplementary analysis.

Importantly, the four-way interaction was not significant: $F(1,42) = 0.81$, $p = .37$, $\eta_p^2 = .02$. The three-way interaction between group, partner appearance and time was also not significant, as reported in the main paper: $F(1,42) = 0.98$, $p = .33$, $\eta_p^2 = .02$. Critically, the other significant effects described in the paper were also still significant in this overall four-way analysis, including the three-way interaction between group, partner behaviour and time; $F(1,42) = 7.61$, $p = .009$, $\eta_p^2 = .15$; the two-way interaction between time and partner behaviour: $F(1,42) = 13.39$, $p < .001$, $\eta_p^2 = .24$; the two-way interaction between group and partner behaviour: $F(1,42) = 8.42$, $p = .006$, $\eta_p^2 = .17$; and the main effects of partner appearance: $F(1,42) = 4.60$, $p = .04$, $\eta_p^2 = .10$, and partner behaviour: $F(1,42) = 17.69$, $p < .001$, $\eta_p^2 = .30$.

Interestingly, there was a significant three-way interaction between partner appearance, partner behaviour and time: $F(1,42) = 11.24$, $p = .002$, $\eta_p^2 = .21$, reflecting the fact that on

average, participants learned to discriminate between trustworthy looking partners in terms of fair behaviour: $F(1,44) = 26.53$, $p < .001$, $\eta_p^2 = .38$, but did not learn to discriminate between untrustworthy looking partners: $F(1,46) = 0.43$, $p = .51$, $\eta_p^2 = .009$. That is, partners who looked untrustworthy faced a penalty even if they behaved fairly. No other effects in the overall four-way analysis were significant: all $F(1,42) < 1.34$, $p > .25$, $\eta_p^2 < .03$.

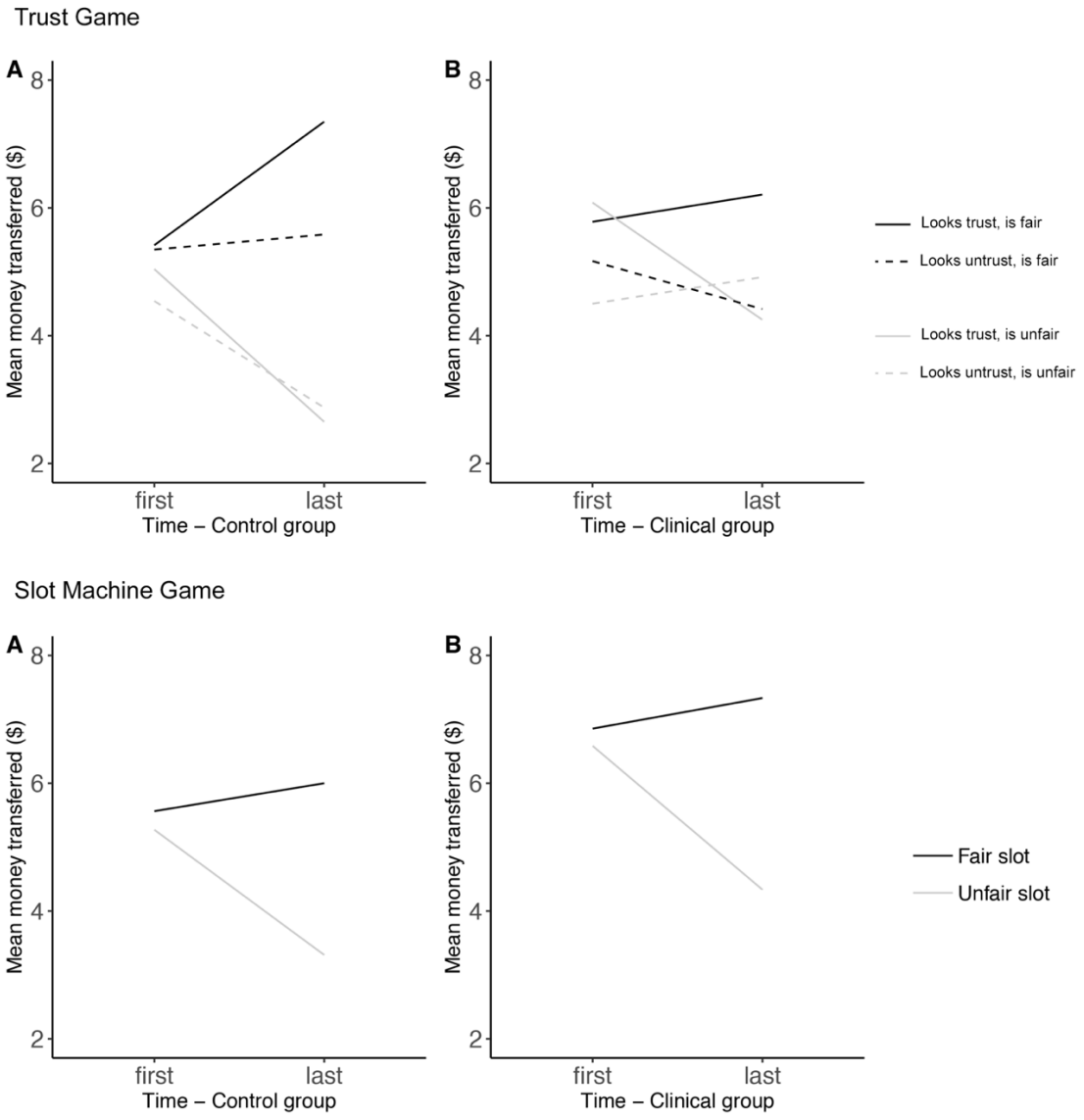

Figure S1. Upper panel: Money transferred in the Trust Game on average for A) control (left) and B) clinical participants (right panel), based on partner trustworthy (light grey) or untrustworthy (dark grey) appearance, as well as partner actual fair (dark grey) or unfair (light grey) behaviour, throughout the Trust Game. **Lower panel:** Money transferred in the Slot Machine Game on average for A) control (left) and B) clinical participants (right panel), based on slot machine rewarding (dark grey) or unrewarding (light grey) behaviour, in the first and last trials of the Slot Machine Game. *NB: Confidence intervals are omitted as these are potentially misleading, given the mixed design, and highly variable, given that each point here reflects one transaction per participant only.*