Supplementary Methods

Experimental Timing

An important feature of this study was the ability to investigate activity time-locked specifically to privileged information cues. Variable delays were introduced between the instruction cue (Doors) and the privileged information cue, as well as between the privileged information cue and the belief judgment cue. As in previous studies (1-3) this allowed us to isolate blood oxygen level-dependent (BOLD) activity time-locked to the privileged information and the prediction error signal without the contaminating effects of subsequent trial events (unrelated visual and motor responses). Events in each trial took place across 4 repetition times (TRs) (0–8 s; TR = 2 s, Fig. 1). In order to optimally sample evoked hemodynamic responses (EHRs), we randomly varied the interval between scan onset and each of the three elements of the trial (Doors, privileged information, belief response). We specifically, jittered the onset of the Instruction cue (Doors) 0-500ms from the onset of the first TR, the onset of the privileged information 0-3000ms from the onset of the second TR (thus it was jittered between the second and third TR), and the belief response was jittered 0-500ms from the onset of the fourth TR. This achieved an effective temporal sampling resolution much finer than one TR. These intervals were uniformly distributed, ensuring that EHRs time-locked to the privileged information cue were sampled evenly across the time period. Jittering the onset of privileged information cue relative to the start of each Echo Planar Imaging (EPI) volume additionally guarantees that stimuli are presented during the acquisition of every slice, and as such there is no spatial bias in our imaging protocol.

Since the privileged information cues were temporally uncorrelated with the preceding and subsequent elements of the task, they could be modeled as independent event types. This allowed us to investigate activity time-locked to privileged information cues (and prediction errors) without the contaminating effects previous or subsequent visual and attention confounds, or motor responses.

Supplementary Results

Behavioural results

To investigate accuracy we ran a four way ANOVA, using Group (ASD, TD), Outcome (PO, PE), Reward (Positive, Neutral), and Agent (1st person, 3rd person, Computer) as factors. Accuracy was calculated as the number of correct responses/(number of correct and incorrect responses). Missed responses were excluded from the calculation of accuracy.

All the main effects were significant. Specifically, there was a main effect of Outcome ($F(1,34) = 12.9$, $p < 0.001$, $\eta p^2 = 0.275$) with greater accuracy for PO trials (90.76% ± 0.922) compared to PE trials (88.08% ± 1.271). There was a main effect of Reward ($F(1,34) = 4.76$, $p < 0.05$, $\eta p^2 = 0.123$) with greater accuracy for positive outcome trials (90.262% ± 0.953) compared to neutral outcomes (88.58% ± 1.255). There was a Main effect of Agent ($F(2,68) = 14.42$, $p < 0.001$, $\eta p^2 = 0.298$) driven by significantly better performance on 1st person trials (91.95% ± 0.868) compared to both 3rd person trials (87.81% ± 1.294) and Computer trials (88.50% ± 1.239). However, there was also a Group x Agent interaction ($F(2,68) = 6.43$, $p < 0.005$, $\eta p^2 = 0.159$) given that this effect was more exaggerated in individuals with ASD compared to TD (see Figs. 1d,e). Individuals with ASD showed a significantly better performance for 1st person trials (90.28% ± 1.294) compared to both 3rd person trials (84.38% ± 2.282) and Computer trials (84.06% ± 2.427), whereas TD individuals showed a significant difference between 1st person (94.21% ± 0.913) and 3rd person trials (92.25% ± 1.278) but not between 1st person and Computer trials (93.71% ± 0.912). There was also a main effect of Group (F(1,34) = 13.05, $p < 0.001$, $\eta p^2 = 0.277$) given that the individuals with ASD were significantly less accurate (85.65% ± 1.558) compared to the TD group (93.2 ± 1.394). Finally, there was a Outcome x Reward interaction (F(1,34) = 23.34, $p < 0.001$, $\eta p^2 = 0.407$) given that on PO trials there was greater accuracy for positive (93.67% ± 0.818) compared to neutral outcome trials (87.86% ± 1.293). However, on PE trials there was greater accuracy for neutral outcome trials (89.3% ± 1.375) compared to positive trials (86.86% ± 1.421). There were no significant correlations between accuracy for each condition, IQ, ADOS, ADI, SRS, or SCQ values.

In order to ascertain whether these accuracy effects were due to an increase in incorrect responses or perhaps an increase in missed responses we analysed the number of misses. There was a main effect of Agent (F(2,68) = 6.41, $p = 0.003$, $\eta p^2 = 0.16$), driven by the decreased misses to 1st person compared to 3rd person and Computer. However, there were no significant Group differences (F(1,34) = 2.99, $p = 0.093$; $\eta p^2 = 0.08$), and no Group x Agent interaction (F(2,68) = 3.04, $p = 0.054$, $\eta p^2 = 0.08$). This suggests that the group differences in accuracy are due to incorrect responses rather than a failure to respond.

fMRI results

Effect of 3rd person PE>PO

The ACCg was differentially activated by PE trials versus PO trials, and this effect was significantly larger for the 3rd person perspective than for the 1st person or the Computer trials. Supplementary Fig. 1a shows the portion of ACCg showing a Belief x Agent interaction in the TD group (green), and the portion of ACCg showing a Group difference in the Belief x Agent interaction (red). The overlap between these regions is highlighted in yellow. Supplementary Fig. 1b shows the percent signal change responses for all conditions from the ACCg (Group x Agent x Belief interaction shown in Fig. 2 and Supplementary Fig. 1a in red). It is clear from this that the Group x Agent x Outcome interaction in the ACCg was driven by group differences in the 3rd person PE and PO trials compared to all other conditions. The largest effects are in the 3rd person PE trials in both groups, suggesting that this condition is driving activity in the ACCg.

These responses were for the parametric modulation regressors (i.e. positive hrf for positive reward and negative hrf for neutral rewards). For further clarification we provide percent signal change values for all conditions, separating positive and neutral outcome trials. Supplementary Fig. 2a shows positive outcome trials whilst the Supplementary Fig. 2b shows all neutral outcome trials. The one condition that stands out is the 3rd PE+ which is larger than all other conditions in the TD group and shows a significant interaction Belief x Agent interaction for positive outcome ($T(19) = 2.42$, $p < 0.05$) but not for negative outcome trials ($T(19) = -0.78$, $p > 0.05$). Only the 3rd person PE+ condition showed a significant difference between ASD and TD ($T(34) = 2.61$, $p < 0.05$).

Main effect of Outcome (PE > PO trials)

Fig S3 shows all the regions showing a significant increase in BOLD activity on PE trials compared to PO trials (red-yellow). This included the dorsomedial prefrontal cortex, left posterior superior temporal sulcus (bordering on the temporo-parietal junction), right temporal pole, bilateral parietal lobules and the precuneus. All these regions are described in the Supplementary Table 1. Regions in blue in Supplementary Fig. 3 show a significantly greater decrease in BOLD activity (significantly greater negative BOLD response) for PE trials compared to PO trials. This included the precuneus, right superior frontal sulcus, and right middle temporal gyrus.

Two regions showed significant group differences in Belief (PE<>PO) signals across Agents; the right inferior frontal gyrus, pars triangularis (MNI coordinates [x = 51, y = 29, z = 23], t = 4.4, k = 54, $p < 0.05$ FWE corrected using TFCE (4), assigned Area 9/46v (31%; 5), Supplementary Fig. 4a), and the sulcus of the anterior cingulate cortex (ACCs; MNI coordinates [x = 9, y = 29, z = 32], t = 4.32, k = 54, $p < 0.05$ FWE corrected using TFCE (4), assigned to anterior rostral cingulate zone (RCZa (36%); 6), Fig S4b). Neither region was associated with ASD social symptom severity ($p > 0.22$, see Table S2).

Effect of 1st person PE>PO

The left caudate nucleus (Supplementary Fig. 5a) and occipital lobe (Supplementary Fig. 5b; V3d 53%), showed a strong difference between 1st person PE and PO trials, although it did not show any differences between PO and PE trials for other agents (see bar plots in Supplementary Fig. 5). Both the caudate nucleus and occipital lobe activations were derived from a 1 sample t-test collapsing across ASD and TD individuals and were not significantly different between groups ($p > 0.05$).

Effect of Computer PE>PO

The occipital lobe (hOC1 77%; Supplementary Fig. 6), showed a strong difference between 1st person PE and PO trials, however, it also showed a strong activation for Computer PE trials. This activation was derived from a 1 sample t-test collapsing across ASD and TD individuals and were not significantly different between groups ($p > 0.05$).

References

1. Balsters J, Ramnani N (2011) Cerebellar plasticity and the automation of first-order rules. *Journal of Neuroscience* 31(6):2305–2312.

2. Balsters J, Ramnani N (2008) Symbolic representations of action in the human cerebellum. *NeuroImage* 43(2):388–398.

3. Balsters J, Whelan CD, Robertson IH, Ramnani N (2013) Cerebellum and cognition: evidence for the encoding of higher order rules. *Cereb Cortex* 23(6):1433–1443.

4. Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44(1):83–98.

5. Sallet J, et al. (2013) The Organization of Dorsal Frontal Cortex in Humans and Macaques. *Journal of Neuroscience* 33(30):12255–12274.

6. Neubert F-X, Mars RB, Sallet J, Rushworth MFS (2015) Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proc Natl Acad Sci USA* 112(20):201410767–704.

Supplemental Figure Legends

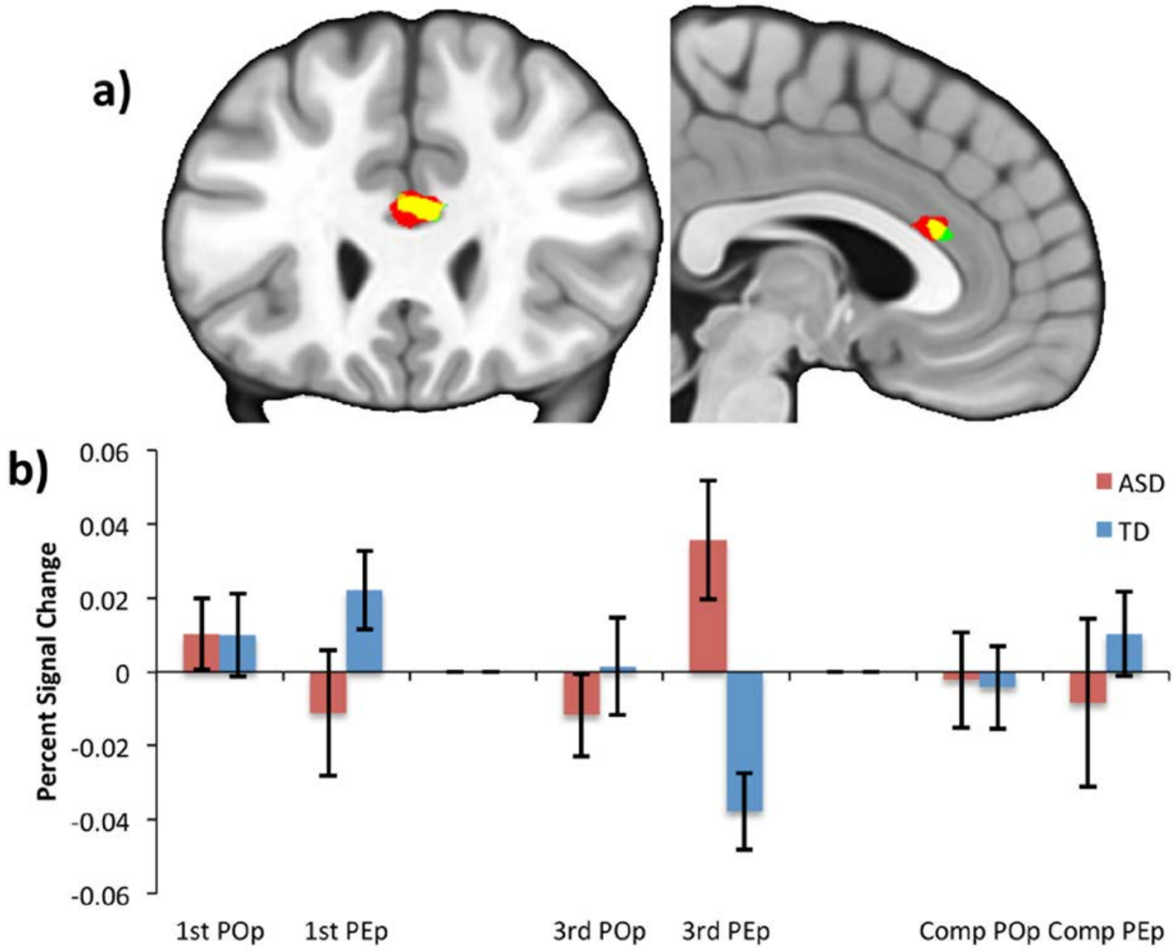

Figure S1: Social PE signals in the ACCg. A) A Belief x Agent interaction in the TD group (green) and the group difference in the Belief x Agent interaction (red; also seen in Fig 2a). The overlap between these regions is shown in yellow. B) Percent signal change responses from the ACCg (Group x Agent x Belief interaction shown in Fig 2 and Fig S1a in red). The labels 1st, 3rd, and Comp refer to 1st person, 3rd person and Computer trials respectively. The labels POp and PEp refer to Predicted Outcome and Prediction Error parametric modulators respectively. This bar plot shows that group differences in the Belief x Agent interaction in the ACCg were driven by social PE trials (3rd PEp). Error bars indicate standard error.

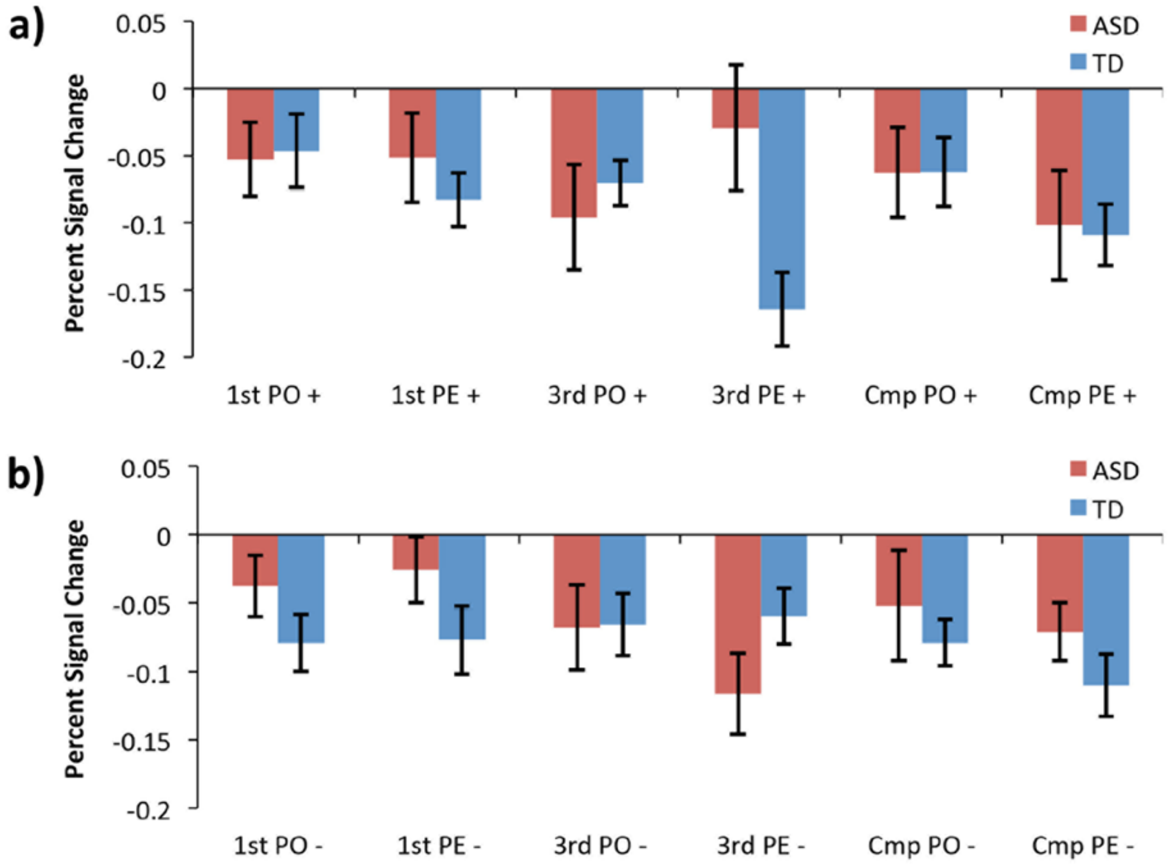

Figure S2: Percent signal change responses for all twelve conditions from the ACCg (Group x Agent x Belief interaction shown in Fig 2 and Fig S1a in red) separated for Positive outcome trials (A) and Negative outcome trials (B). The labels 1st, 3rd, and Cmp refer to 1st person, 3rd person and Computer trials respectively. The labels PO and PE refer to Predicted Outcome and Prediction Error trials respectively. This bar plot highlights that trails where another person unexpectedly wins (positive social PEs; 3rd PE+) were driving the Belief x Agent interaction and the Group x Belief X Agent interaction in the ACCg. Error bars indicate standard error.

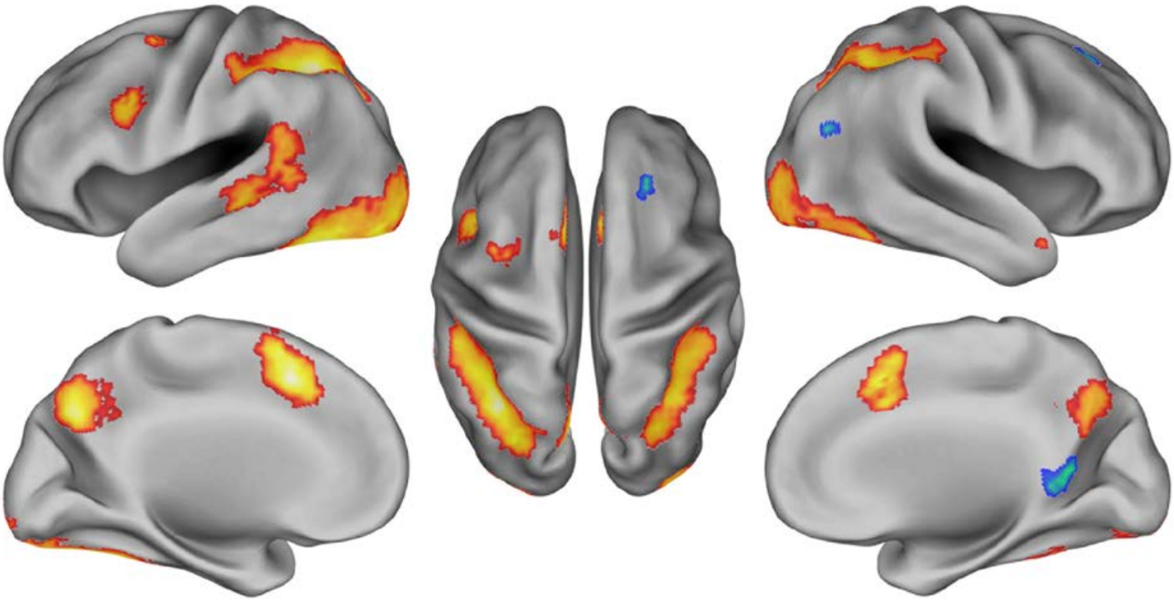

Figure S3: Brain regions showing a significant increase in BOLD activity on PE trials compared to PO trials (red-yellow) across all Agents. Regions in blue showed a significantly larger negative BOLD response for PE compared to PO trials. All these regions are described in the Table S1.

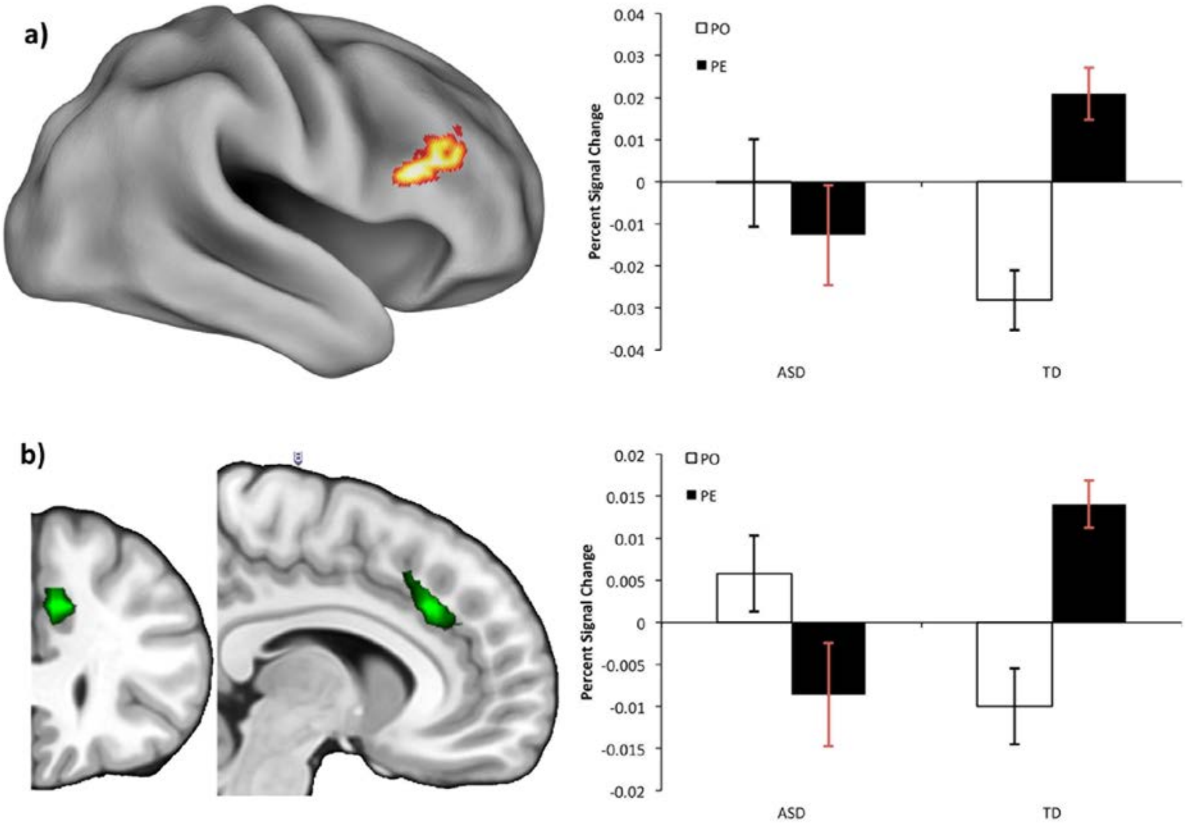

Figure S4: Group differences in Belief (PE <> PO) were present in the right middle frontal gyrus (A) and ACCs (B). Bar plots right of each brain image show percent signal change values for PE and PO trials per group. Error bars indicate standard error.

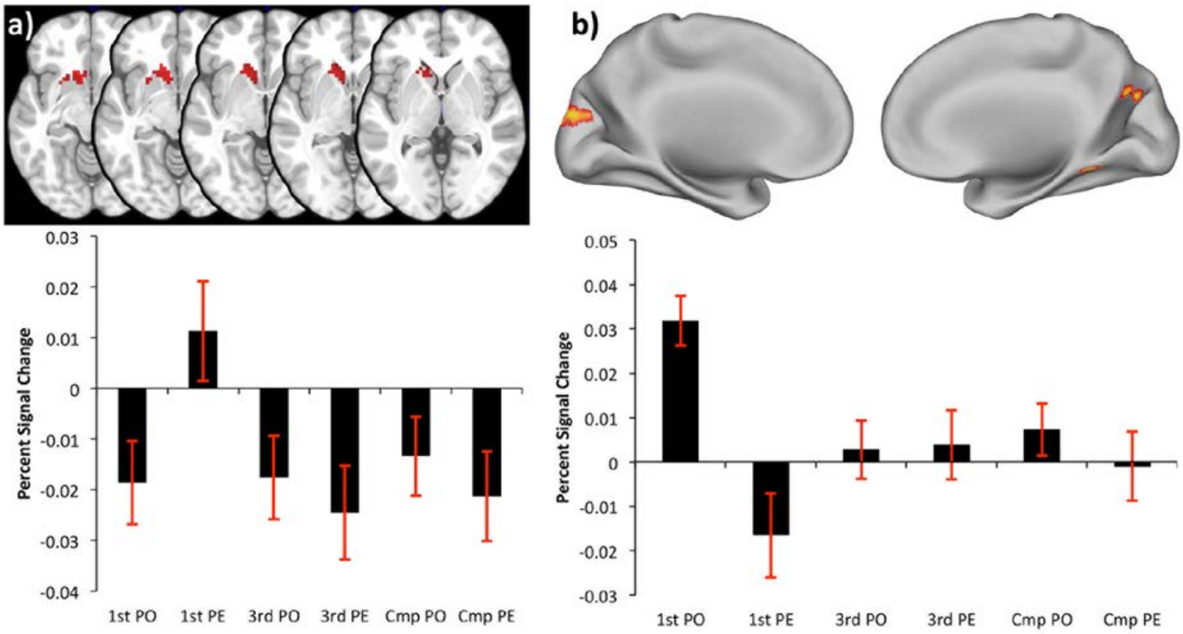

Figure S5: The left caudate nucleus (A) and occipital lobe (B) showed a strong

difference between 1st person PE and PO trials, although it did not show any differences between PO and PE trials for other agents or any group differences. Bar plots underneath each brain image show percent signal change values for each condition. The labels 1st, 3rd, and Cmp refer to 1st person, 3rd person and Computer trials respectively. The labels PO and PE refer to Predicted Outcome and Prediction Error trials respectively. Error bars indicate standard error.

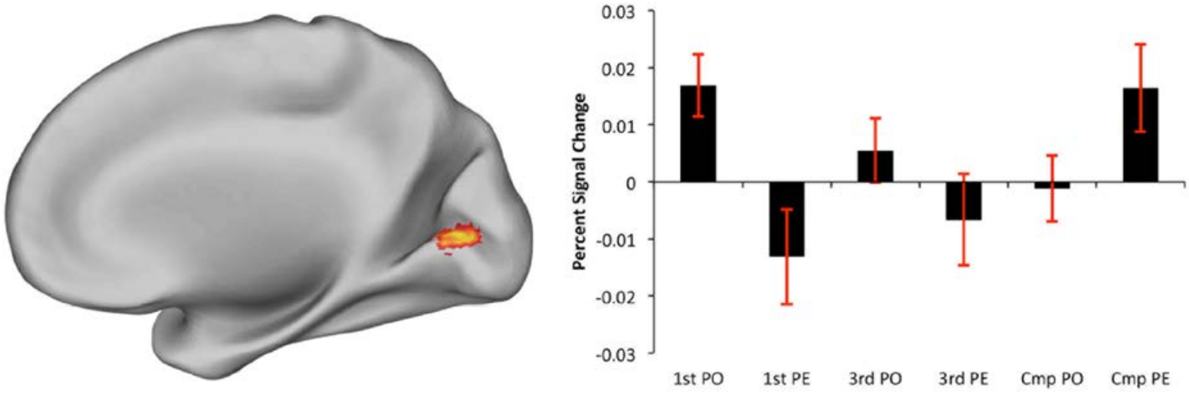

Figure S6: Activation in the left occipital lobe showing a strong difference between 1st person PE and PO trials, as well as Computer PE and PO trials. This region did not show any differences between PO and PE trials for other agents or any group differences. Bar plots underneath each brain image show percent signal change values for each condition. The labels 1st, 3rd, and Cmp refer to 1st person, 3rd person and Computer trials respectively. The labels PO and PE refer to Predicted Outcome and Prediction Error trials respectively. Error bars indicate standard error.

Table S1: Main effect of Belief (PE <> PO) timelocked to the presentation of privileged information (Fig. 1b), p<0.05 FWE corrected using TFCE. Cluster size indicates the number of voxels active in each cluster. X-coordinates with a negative value represent activity in the left hemisphere. Anatomical localization was guided by the Anatomy toolbox (5-7), along with more detailed frontal lobe atlases (8, 9). Abbreviations: RCZp: posterior rostral cingulate zone; IFJ: Inferior Frontal Junction; Ant PMd: anterior dorsal premotor cortex; hIP3: human intraparietal area 3; FG2: area 2 of the fusiform gyrus.

Anatomical Label	cluster size	co-ordinates			t	Cytoarchitectonic or connectivity based parcellation %
		x	y	z		
Frontal Lobe						
Posterior Medial Frontal	303	-6	13	52	6.97	Pre-SMA (54%), RCZp (37%)
L Inferior Frontal Gyrus (pars Opercularis)	96	-39	4	28	5.75	IFJ (96%)
L Middle Frontal Gyrus	40	-30	-2	55	5.61	Ant PMd (36%), Area 8A (35%)
Temporal Lobe						
L Middle Temporal Gyrus	54	-51	-38	1	4.48	
L Superior Temporal Gyrus	43	-54	-50	19	3.93	
R Medial Temporal Pole	3	48	10	-23	5.5	
Parietal Lobe						
L Inferior Parietal Lobule	1172	-27	-62	46	7.62	hIP3 67%
R Angular Gyrus	"	30	-62	55	6.38	hIP3 39%
Precuneus	"	-6	-74	46	6.24	
Occipital Lobe						
L Fusiform Gyrus	719	-39	-68	-8	7.42	FG2 55%
R Inferior Occipital Gyrus	203	39	-74	-2	4.94	
Subcortical						
R Cerebellum	252	30	-80	-26	4.75	Crus I (90%)

Table S2: Neural correlates of ASD social symptom severity. Table showing correlations (r values) between BOLD fMRI signals and neuropsychological measures of social symptom severity. The only significant correlation (ACCg social PE+ and ADOS social symptom severity) is highlighted in bold (P=0.007, FDR corrected p<0.05). Abbreviations: ADOS: Autism Diagnostic Observation Schedule; ADI-R: Autism Diagnostic Interview Revised; SRS: Social Responsiveness Scale; SCQ: Social Communication Questionnaire; ACCg: Anterior Cingulate Gyrus; ACCs: Anterior Cingulate Sulcus; R MFG: right middle frontal gyrus; vmPFC: ventromedial prefrontal cortex;

	ADOS Social	ADI Social	SRS Total	SCQ	Behaviour
ACCg Social PE+	**0.664**	0.412	0.280	0.033	-0.196
vmPFC Interaction effect	-0.262	-0.231	-0.220	-0.192	-0.518
vmPFC to ACCg DCM	-0.060	0.222	0.082	-0.022	0.385
ACCs PE>PO	-0.071	-0.005	0.368	0.039	0.176
R MFG PE>PO	0.200	-0.165	-0.137	-0.022	0.404