doi:10.1093/brain/aww287 BRAIN 2017: 140; 235–246 | 235

BRAIN
A JOURNAL OF NEUROLOGY

Disrupted prediction errors index social deficits in autism spectrum disorder

Joshua H. Balsters,[1,2] Matthew A. J. Apps,[3] Dimitris Bolis,[1] Rea Lehner,[1] Louise Gallagher[4] and Nicole Wenderoth[1,5]

Social deficits are a core symptom of autism spectrum disorder; however, the perturbed neural mechanisms underpinning these deficits remain unclear. It has been suggested that social prediction errors—coding discrepancies between the predicted and actual outcome of another's decisions—might play a crucial role in processing social information. While the gyral surface of the anterior cingulate cortex signalled social prediction errors in typically developing individuals, this crucial social signal was altered in individuals with autism spectrum disorder. Importantly, the degree to which social prediction error signalling was aberrant correlated with diagnostic measures of social deficits. Effective connectivity analyses further revealed that, in typically developing individuals but not in autism spectrum disorder, the magnitude of social prediction errors was driven by input from the ventromedial prefrontal cortex. These data provide a novel insight into the neural substrates underlying autism spectrum disorder social symptom severity, and further research into the gyral surface of the anterior cingulate cortex and ventromedial prefrontal cortex could provide more targeted therapies to help ameliorate social deficits in autism spectrum disorder.

1 Neural Control of Movement Laboratory, Department of Health Sciences and Technology, ETH Zurich, Switzerland
2 Department of Psychology, Royal Holloway University of London, Egham, Surrey, UK
3 Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK
4 Department of Psychiatry and Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland
5 Movement Control and Neuroplasticity Research Group, Department of Kinesiology, KU Leuven, Belgium

Correspondence to: Dr Joshua Henk Balsters,
Neural Control of Movement Lab,
Department of Health Sciences and Technology,
ETH Zurich, Switzerland
E-mail: Joshua.balsters@hest.ethz.ch

Keywords: autism spectrum disorder; social cognition; anterior cingulate cortex; social prediction errors

Abbreviations: ACCg/s = anterior cingulate cortex gyrus/sulcus; ASD = autism spectrum disorder; BOLD = blood oxygen level-dependent; TD = typically developing; vmPFC = ventromedial prefrontal cortex

Introduction

One of the cardinal characteristics of autism spectrum disorder (ASD) is a deficit in social interaction along with an inability to understand the beliefs and intentions of others. However, the computational mechanisms that underpin this social deficit are currently unclear. Recent theoretical accounts of ASD (Lawson *et al.*, 2014; Van de Cruys *et al.*, 2014) propose that deficits in understanding others may arise due to aberrant computation of socially-specific prediction errors. Traditionally, prediction errors signal the discrepancy between our own expectations and the actual outcomes of an action (e.g. earning a reward); however, social prediction errors shift the frame of reference from the first person to the third

Received July 25, 2016. Revised September 10, 2016. Accepted September 23, 2016

person perspective by comparing actual outcomes with the perceived expectations of another person. One candidate brain area for signalling social prediction errors is the gyral surface of the anterior cingulate cortex (ACCg) as indicated by converging evidence from both human and non-human primates (Apps *et al.*, 2013*b*, 2016). First, the ACCg has a unique connectivity fingerprint compared to other regions of the anterior cingulate cortex such as the adjacent sulcus of the anterior cingulate cortex (ACCs). While both the ACCs and ACCg are interconnected with brain regions dedicated to processing reward-related information (Yeterian and Pandya, 1991; Lynd-Balta and Haber, 1994; Haber *et al.*, 1995), i.e. indicating a positive outcome of an action, the ACCg has additional unique anatomical connections with brain regions that process social information (Vogt and Pandya, 1987; Seltzer and Pandya, 1989; Barbas *et al.*, 1999). Specifically, the ACCg has stronger connections to regions within the temporoparietal junction and dorsomedial prefrontal cortex—regions often implicated in social cognition and mentalizing processes—that do not overlap with connections from the adjacent regions of the cingulate cortex (Apps *et al.*, 2013*b*, 2016). These unique connections facilitate the processing of reward-related information for others in the ACCg. Second, lesions specifically to the ACCg that leave the ACCs intact have been shown to impair the processing of social stimuli and social behaviours in non-human primates (Rudebeck *et al.*, 2006). Third, single-unit recordings in non-human primates show that a larger proportion of ACCg neurons, compared with those in orbitofrontal cortex or ACCs, track the rewards and outcomes of others (Chang *et al.*, 2013). Hill *et al.* (2016) recently replicated this in humans by demonstrating that neurons in the rostral ACC (putatively ACCg) encode both the expected outcome and discrepancies between the expected and actual outcome of another person's decision (i.e. social prediction errors) when learning from another through observation. This coding was present only in the ACCg and not the amygdala or ventromedial prefrontal cortex (vmPFC) suggesting a certain degree of specificity for social prediction error coding in the ACCg. Finally, there is a large body of neuroimaging evidence demonstrating that the ACCg encodes reward-related information for other individuals such as the probability that another individual will receive a reward (Lockwood *et al.*, 2015), the net-value (cost-benefit) of another individual's decision (Apps and Ramnani, 2014), and the discrepancy between another person's expectations and the actual outcomes, i.e. social prediction errors (Behrens *et al.*, 2008; Apps *et al.*, 2012, 2013*a*, 2015). Crucially, in all these studies the ACCg responded exclusively to outcome-related information about others (i.e. only the third person perspective), but not to outcome-related information for themselves or non-biological agents (i.e. computer players).

Interestingly, the ACCg has been shown to be affected by ASD pathology (Torta and Cauda, 2011) as indicated by post-mortem studies showing decreased neuron size and density in the ACCg of individuals with ASD (Simms *et al.*, 2009). Neuroimaging studies have shown aberrant ACCg resting connectivity in a large sample of ASD individuals (Balsters *et al.*, 2016), and a meta-analysis of functional MRI studies in ASD requiring social information processing showed consistently reduced ACCg activity during social tasks (Di Martino *et al.*, 2009). However, it is unclear what mechanism links together structural and functional changes in the ACCg in ASD and deficits in social behaviour. Here, we test whether individuals with ASD have a deficit in correctly representing social prediction errors when tracking the expectations of others and whether this putative social deficit can be linked to abnormal activity in the ACCg.

Materials and methods

Participants

Twenty-eight ASD and 28 typically developing (TD) individuals participated in this study. All participants were male and self-reported as right-handed. Twelve ASD and eight TD individuals were excluded due to poor task performance ($<55\%$ correct responses in any condition: three TD, nine ASD), excessive head movement [>2 standard deviations (SD) of the group specific mean framewise displacement: one TD, one ASD], or technical difficulties during scanning (signal dropout: three TD, two ASD; missing responses: one TD). All analyses included 16 ASD (age: 20.97 years $\pm$ 3.17 years; IQ 115.5 $\pm$ 10.61) and 20 TD individuals (age: 22.17 years $\pm$ 5.2 years; IQ: 118 $\pm$ 11) who were matched for age, IQ, gender, and handedness. ASD participants were recruited through an associated genetics research programme, clinical services, schools, and advocacy groups. TD individuals were recruited through schools, the university (Trinity College Dublin), and volunteer websites. Ethical approval was obtained from St. James's Hospital/AMNCH (ref: 2013/08/09) and the Linn Dara CAMHS Ethics Committees (ref: 2013/01/15). Written informed consents/assents were obtained from all participants and their parents (where under 18 years of age).

Exclusion criteria included a Full Scale IQ (FSIQ) <80, known psychiatric, neurological, or genetic disorders, a history of a loss of consciousness for >5 min. TD individuals were excluded if they had a first-degree relative with ASD or scored >50 on the Social Responsiveness Scale (SRS; Constantino *et al.*, 2003) or >10 on the Social Communication Questionnaire (SCQ; Rutter *et al.*, 2013). The adult prepublication version of the SRS was used with permission in cases 18 years or older (Constantino and Todd, 2005). All participants had normal, or corrected to normal, vision.

Diagnostic assessments and cognitive measures

Clinical diagnosis of ASD, which was established prior to recruitment to the study for all ASD participants, was confirmed using the Autism Diagnostic Observation Schedule (ADOS; Lord *et al.*, 2000) and the Autism Diagnostic Interview Revised (ADI-R; Lord *et al.*, 1994), and clinical consensus diagnosis carried out by an expert clinician (L.G.) in

accordance with DSM-IV-TR criteria. FSIQ was measured using the four-subtest version of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) or the Wechsler Intelligence scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003).

Task

Here we developed a novel false belief paradigm that is more amenable to event-related functional MRI than traditional Theory of Mind (ToM) paradigms (Fig. 1). This paradigm had three players: 'Player 1' (the participant inside the MRI scanner; first person perspective), 'Player 2' (the stooge outside the MRI scanner; third person perspective), and 'Computer' (a computer generated player; non-biological control). Before the participant went into the MRI scanner they met the stooge (Player 2) outside the MRI scanner and practised the task together for ~15 min. This gave the participant the feeling that Player 2 was really playing outside the MRI scanner, even though Player 2's responses were computer-generated and identical for each participant. After the functional MRI experiment all participants confirmed that they believed Player 2's responses were made by the stooge outside the scanner.

A trial began with the presentation of two white 'Doors' (duration up to 1 s). Participants were instructed that there was a prize behind one of the doors (€1) and if they chose the correct door they would win €1 on that round. If a door was not selected after 1 s the trial was marked as missed and any reward-related content for that trial was replaced with the word 'Missed' in red letters. Printed underneath the 'Doors' was the name of the Agent who had to respond on that trial (Fig. 1A). For example, if 'Player 1' was printed under the 'Doors' then the participant in the MRI scanner could choose a door in order to win a prize for themselves. If 'Player 2' or 'Computer' was printed underneath the 'Doors' then the participant in the MRI scanner would watch what they believed to be the person outside the MRI scanner or the Computer choose a door to try and win a prize. As soon as the left or right door was chosen it would change colour to red or green (duration 500 ms). When a door turned green it indicated that the Agent playing that trial had likely won €1, if the door turned red then whoever was playing that trial had likely not picked the winning door and would not win anything on that trial. The green-win/red-neutral contingency was true for the majority of trials (66%). For the remaining 34% of trials this contingency was reversed (green-neutral/red-win). For Players 1 and 2 the onset of the colour change varied from

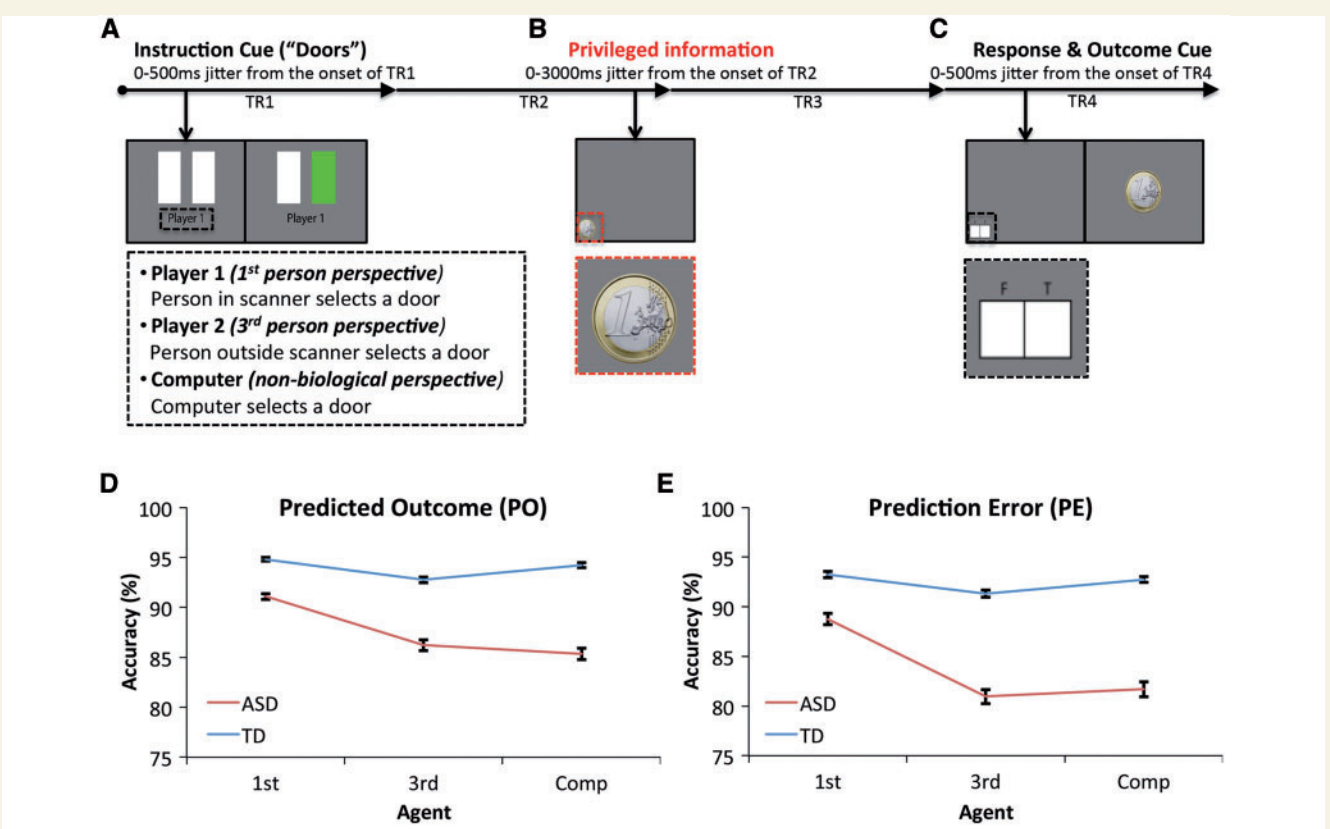

Figure 1 Task schematic and behavioural results. (**A**) A trial begins with the presentation of two white rectangles (doors) and the label of the Agent playing that trial. Once a door was chosen it changed colour indicating if the Agent has probably chosen the door with a prize (green) or the empty door (red). (**B**) The participant in the MRI scanner was shown the actual outcome before the other Agents. At this point the participant can determine if the outcome is expected (green-win/red-neutral) or unexpected (green-neutral/red-win). (**C**) The participant indicated if the outcome was expected (T) or unexpected (F) for all Agents and trials, after which the outcome is revealed to all Agents. Behavioural results showing accuracy [i.e. if participants correctly recognized that the outcome of the trial was expected (**D**) or unexpected (**E**)] shown for each Agent.

trial-to-trial highlighting variability in response times, whereas the change in colour for the Computer always occurred 500 ms after the two white doors appeared. This manipulation reinforced the feeling that Player 2 was a real person outside the MRI scanner.

The participant inside the MRI scanner was informed that they would always learn the outcome of a trial (i.e. what is behind the door) before the other Agents. We refer to this as privileged information (Fig. 1B). Once the participant in the MRI scanner learned the actual outcome they could determine whether the expectation was true (green-win/red-neutral) or false (green-neutral/red-win). While expected outcome trials should not elicit a prediction error (expected outcome = actual outcome), unexpected outcome trials should elicit a prediction error given that the expected outcome was not equal to the actual outcome. Depending on which Agent was performing that trial it would either elicit a first person prediction error, a third person prediction error, or a Computer prediction error.

After the privileged information cue the participant in the MRI scanner was prompted to respond to indicate if the outcome of the trial was expected (T) or unexpected (F) (Fig. 1C). The participant had to indicate this for all trials, keeping the participant engaged in trials for other Agents as well as their own. Finally, the outcome is revealed to all participants.

Conditions

The 12 trial types were embedded in a $2 \times 3 \times 2$ factorial design (three factors with 2/3 levels).

Factor 1: Belief (predicted outcome or prediction error). For 66% of trials the outcomes were predictable (predictable outcome trials; i.e. expected outcome = actual outcome; green door = win/red door = no win). For the remaining 34% of trials the outcome was unexpected eliciting a prediction error (prediction error trials; i.e. expected outcome $\neq$ actual outcome; green door = no win/red door = win).

Factor 2: Agent (Player 1; Player 2; Computer). A trial could be performed by one of three agents: Player 1 (first person perspective), Player 2 (third person perspective), or the Computer (non-biological control).

Factor 3: Reward (Positive or neutral). Participants could either win money on a trial (positive) or not win money on a trial (neutral).

This $2 \times 3 \times 2$ factorial resulted in 12 conditions. Condition 1: first person positive predictable outcome. Player 1 (person inside the MRI scanner) expects to win €1 (green door) and they do win €1 (39 trials); Condition 2: first person neutral predictable outcome. Player 1 (person inside the MRI scanner) does not expect to win anything on this trial (red door) and they do not win anything (39 trials); Condition 3: first person positive prediction error. Player 1 (person inside the MRI scanner) does not expect to win anything on this trial (red door) but they actually win €1 (21 trials); Condition 4: first person neutral prediction error. Player 1 (person inside the MRI scanner) expects to win €1 (green door) but they do not win anything (21 trials); Conditions 5–8 are identical to conditions 1–4; however, they are performed by Player 2 (third person perspective); and Conditions 9–12 are identical to conditions 1–4, however, they are performed by the Computer.

All participants completed three sessions of the task, performing a total of 360 trials (120 per session). Each session included 40 trials per agent [26 predictable outcome trials (13 positive and 13 neutral) and 14 prediction error trials (seven positive and seven neutral)]. Trials were blocked so that the agent was constant for 10 trials, i.e. 10 first person trials, followed by 10 Computer trials, etc. The order of agents was pseudo-randomized so that the agent changed every 10 trials.

The aim of this investigation was to examine activity related to group differences in social prediction errors. Regions showing a social prediction error would be established through an Agent × Belief contrast, specifically a larger magnitude response to prediction error compared to predictable outcome trials that is greater for the third person perspective compared to first person and Computer. Reward could be included as an additional factor in this contrast, given that there may be differences in the sign of the prediction error signal for positive compared to negative outcomes (O'Doherty *et al.*, 2003; Schultz, 2007). We therefore focused on Group × Agent × Belief, and Group × Agent × Belief × Reward interactions.

Behavioural analyses

Behavioural data were analysed using SPSSv23. Two sample *t*-tests were used to examine group differences in age, IQ measures, SRS, and SCQ scores. Mixed model (between/within subjects) ANOVAs were used to examine accuracy and reaction time data. Pearson's correlations were conducted to examine the relationship between the blood oxygen level-dependent (BOLD) response and SRS score and accuracy. Correlations between BOLD response and ADOS/ADI scores were calculated using Spearman's rho, as ADOS/ADI scores are ranked/ordinal. Correlations were corrected for multiple comparisons using false discovery rate (FDR) correction.

Functional imaging and analyses

Apparatus

Subjects lay supine in an MRI scanner with the fingers of their right hand positioned on a 2-button MRI-compatible response box. Stimuli were projected onto a screen behind the subject and viewed in a mirror positioned above the subject's face. Presentation software (Neurobehavioral Systems, Inc., USA) was used for stimulus presentation both inside and outside the scanner. Transistor-Transistor Logic (TTL) pulses were used to drive the visual stimuli in Presentation.

Data acquisition

A high-resolution T_1-weighted anatomic magnetization-prepared rapid gradient echo image (field of view = 230 mm, thickness = 0.9 mm, voxel size = 0.9 mm × 0.9 mm × 0.9 mm) was acquired first. Then each participant performed three echo planar imaging (EPI) sessions containing 494 volumes lasting 16.5 min. The field of view covered the whole brain, 240 mm × 240 mm (80 × 80 voxels), and 40 axial slices were acquired with a voxel size of 3 mm × 3 mm × 3 mm (0.3 mm slice gap), repetition time = 2 s, echo time = 25 ms, flip angle = 90°. In between each EPI session, subjects had a 2-min rest while T_1- and T_2-weighted clinical scans were acquired. These images were not used for any analytical purposes. All MRI data were collected on a Philips 3 T Achieva

MRI Scanner using a 32 channel head coil (Trinity College Dublin).

Preprocessing

Scans were preprocessed using SPM12 (www.fil.ion.ucl.ac.uk/spm). First, structural and functional images were coregistered to the T_1 template, then functional images were coregistered to structural images. After this, functional images were realigned, slice-time corrected, normalized to MNI space using the unified segmentation approach (Ashburner and Friston, 2005), resliced to $3 \times 3 \times 3$ mm, and smoothed with an 8 mm full-width half-maximum kernel. Structural images were segmented to generate grey matter, white matter, and CSF images for each subject. These would be used later to generate additional regressors of no interest (CompCorr; Chai *et al.*, 2012). Two sample *t*-tests comparing all six head motion directions did not show any significant differences between ASD and TD groups ($P > 0.2$).

Statistical analyses

First-level single-subject analyses

Eight conditions were modelled at the first level. The onsets of the instruction cue (Doors) were modelled as one event, including an additional parametric modulator coding for any response made the participant [i.e. 1 if the participant made a response to select a Door (including accidental responses made by the participant to Player 2 or Computer trials), and 0 if the participant did not make a response]. The onset of the privileged information cue was modelled as six event types. Rather than modelling all 12 conditions, the factor of Reward was excluded reducing the model to six conditions: first person predictable outcome and prediction error trials; third person predictable outcome and prediction error trials; Computer predictable outcome and prediction error trials. To account for signed prediction errors [i.e. positive haemodynamic response function (HRF) for positive outcomes and negative HRF for neutral outcomes] and unsigned prediction errors (i.e. positive HRF for positive and neutral outcomes) each of the six events included a parametric modulator coding the outcome of the trial (1 for positive outcomes and −1 for neutral outcomes). Finally, the onset of the belief response cue was also modelled as the eighth event. Only correct trials were included in the analysis, incorrect and missed trials were not modelled. Trials where the participant made an incorrect response could be made for a number of reasons, such as missing the initial instruction cue, failing to hold the predicted value of the cue in working memory, etc. Examining trials where the participant could not identify whether the outcome was predicted or not would therefore be invalid because the mechanisms underlying the processing of the outcomes are impossible to interpret. It is not clear whether expected outcome trials that were reported as unexpected would also elicit a prediction error, or whether unexpected outcome trials reported as expected would fail to elicit a prediction error signal. Due to this ambiguity these trials were ignored to avoid any potentially confounding effects in the analyses. All events were convolved with the canonical HRF. The residual effects of head motion were modelled as covariates of no interest in the analysis by including the six head motion parameters estimated during the realignment stage of the preprocessing. We additionally included the first five principle components derived from white matter and CSF masks as additional regressors of no interest (Chai *et al.*, 2012). Prior to the study, a set of planned experimental timings was carefully checked so that they resulted in an estimable general linear model, in which the events of interest were uncorrelated with other event types ($r < 0.2$). The regression coefficients were then estimated using robust linear regression (Diedrichsen and Shadmehr, 2005), correcting for movement artefacts not accounted for by the head movement, white matter, or CSF regressors, by down-weighting noisy images.

Second-level random-effects group analyses

Second-level analyses were performed using the permutation testing (10 000 permutations) in Randomise (Jenkinson *et al.*, 2012; Winkler *et al.*, 2014) (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Contrasts for main effects and interactions were defined at the first level and input into second level two-sample *t*-test design matrices. Results are reported at $P < 0.05$, familywise error (FWE) corrected at the voxel level across the whole brain using Threshold-Free Cluster Enhancement (TFCE; Smith and Nichols, 2009). Given our anatomically specific hypothesis about the ACCg, a small volume correction was used to correct for multiple comparisons in some cases (ACC mask from the Harvard-Oxford cortical atlas thresholded at >50%). Anatomical localization was guided by the Anatomy toolbox (Eickhoff *et al.*, 2005, 2006, 2007), along with more detailed connectivity-based parcellation atlases of the frontal lobe (Sallet *et al.*, 2013; Neubert *et al.*, 2015).

Dynamic causal modelling

To investigate effective connectivity, Dynamic Causal Modelling (DCM) (Friston *et al.*, 2003; Stephan *et al.*, 2010) was performed using SPM12 (r6591). Subject-specific time series were extracted from specific regions of interest that were selected on the basis of the second-level random-effects (RFX) analysis. As in previous studies (Ouden *et al.*, 2010; Vossel *et al.*, 2012), time series were extracted from the nearest supra-threshold voxel within a radius of 8 mm from the group maximum (ensuring no overlap between regions of interest). The first eigenvariate was then computed across all voxels within 4 mm of the subject-specific coordinates. The resulting time series were adjusted for effects of no interest (i.e. instruction cues, true belief trials, and belief responses) and physiological confounds (head movement, white matter, and CSF regressors). A bilinear DCM model was constructed assuming recursive connections between both regions of interest. Regressors from all prediction error trials (but not the parametric modulator of prediction error trials) were used driving inputs on both regions of interest. Given that group level regions of interest were driven by signed prediction errors we only used parametric modulators of all prediction error trials as modulatory inputs on both connections. DCM parameters were extracted from each subject and input into a Group × Agent ANOVA for each connection.

Results

Individuals with ASD were generally less accurate at identifying both expected and unexpected outcomes compared to typically developing (TD; $n = 20$) matched controls [main effect of Group: predictable outcome trials: $F(1,34) = 11.88$,

$P = 0.002$, $\eta p^2 = 0.259$, Fig. 1D; prediction error trials: $F(1,34) = 11.86$, $P = 0.002$, $\eta p^2 = 0.259$, Fig. 1E]. However, the difference between ASD and TD accuracy was significantly larger for third person and Computer trials compared to first person trials [Group × Agent interaction: predictable outcome trials: $F(2,68) = 4.5$, $P = 0.015$, $\eta p^2 = 0.117$; prediction error trials: $F(2,68) = 3.42$, $P = 0.038$, $\eta p^2 = 0.091$]. The calculation of accuracy excluded missed trials where the participant did not respond ($4.18 \pm 6.77\%$ in ASD; $1.47 \pm 1.78\%$ in TD; see Supplementary material), thus this drop in accuracy reflected significantly more incorrect responses in ASD highlighting that they had greater difficulty monitoring the expectations and outcomes for other Agents than they do for themselves. Both ASD and TD individuals were more accurate for positive outcome trials ($90.262\% \pm 0.953$) compared to neutral outcomes ($88.581\% \pm 1.255$) [main effect of Reward: $F(1,34) = 4.76$, $P = 0.036$, $\eta p^2 = 0.123$]. However, the specific deficit in ASD in tracking the outcomes of others emerged irrespective of the outcome [Group × Agent interaction: positive outcome: $F(2,68) = 4.61$, $P = 0.013$, $\eta p^2 = 0.12$; neutral outcome: $F(2,68) = 3.29$, $P = 0.043$, $\eta p^2 = 0.09$].

Social prediction error signals in the gyral surface of the anterior cingulate cortex

To investigate the neural correlates of agent-specific prediction errors in the neurotypical brain we first restricted our analysis to the TD group and identified regions where activity evoked by prediction error trials differed from predictable outcome trials depending on which agent performed the task (Belief × Agent interaction time-locked to the privileged information cue). Given our hypothesis regarding social prediction errors our first analysis set out to identify brain regions where the difference between third person prediction error and predictable outcome trials was greater than the difference between first person prediction error and predictable outcome trials or Computer prediction error and predictable outcome trials. In agreement with previous work (Apps *et al.*, 2013*a*; Lockwood *et al.*, 2015), this contrast exclusively highlighted activity in the ACCg [Fig. 2B, Supplementary Figs 1 and 2; MNI coordinates ($x = 3$, $y = 29$, $z = 17$), $t = 3.57$, $k = 19$, $P < 0.05$ FWE corrected using TFCE (Smith and Nichols, 2009)

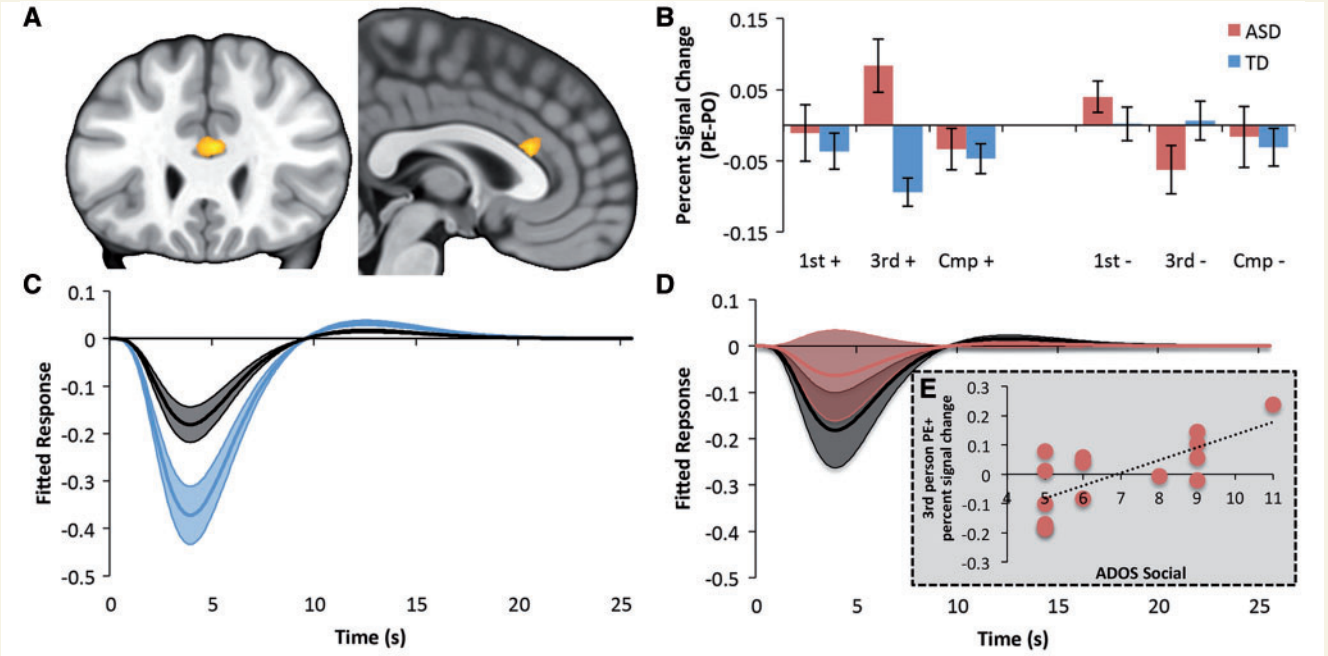

Figure 2 Group differences in social prediction error signalling. (**A**) Group × Belief × Agent interaction in the ACCg [Area 24 (78%; Neubert *et al.*, 2015); MNI: 3 26 20, thresholded at $P < 0.001$ uncorrected] time-locked to the privileged information cue (Fig. 1B). (**B**) Per cent signal change values from the ACCg showing Agent specific prediction errors (prediction error − predictable outcome). first+, third+ and Cmp+ refer to positive outcome trials for first person, third person and Computer trials, respectively. first−, third− and Cmp− refer to negative outcome trials for first person, third person and Computer trials, respectively. This highlights that BOLD differences in the ACCg were driven by third person unexpected rewards (third prediction error positive) compared to all other trials. (**C**) Fitted responses from the ACCg for third person prediction error positive (blue) and third person predictable outcome positive (black) in TD subjects. (**D**) Fitted responses from the ACCg for third person prediction error positive (red) and third person predictable outcome positive (black) in ASD. This highlights that the third+ response in ASD in **B** was driven by larger responses to third predictable outcome positive trials rather than third prediction error positive trials (also see Supplementary Fig. 2). (**E**) Scatter plot showing the significant correlation (r = 0.66) between third person prediction error positive per cent signal change values and social symptom severity in ASD.

and small volume correction, assigned to Area 24 (91%; Neubert *et al.*, 2015)]. Importantly, when distinguishing between trials where the Agent was unexpectedly rewarded (prediction error positive, Fig. 2B) versus those where the expected reward was not received (prediction error negative, Fig. 2B) we found that this effect was driven by prediction error positive trials only (Fig. 2B, C and Supplementary Fig. 2). This finding is consistent with Apps *et al.* (2013), who also showed a strong negative BOLD response in the ACCg was elicited when another person was unexpectedly rewarded, but not when another person failed to receive an expected reward.

To identify whether the ASD group exhibited a different neural response to social prediction errors we calculated a Group × Belief × Agent interaction across the whole brain using the same contrast defined above [i.e. (third prediction error – predictable outcome) > (first prediction error – predictable outcome and Computer prediction error – predictable outcome)]. This Group × Belief × Agent interaction highlighted the same region of the ACCg identified above [Fig. 2A, MNI coordinates ($x = 3$, $y = 26$, $z = 20$), t = 3.78, k = 34, $P < 0.05$ FWE corrected using TFCE (Smith and Nichols, 2009), assigned to Area 24 (78%; Neubert *et al.*, 2015)]. While the TD group showed a strong negative ACCg response when others receive an unexpected reward (third person prediction error positive; Fig. 2B and C), there was no difference between third person prediction error positive and predictable outcome positive trials in the ASD group (Fig. 2B and D). No such effect was observed for prediction error negative trials, i.e. when the Agent did not receive an expected reward (Fig. 2B and Supplementary Fig. 2). TD and ASD groups did not differ in ACCg's responses to first person or Computer trials. No other brain region responded differentially to social prediction error trials in the ASD or the TD group. These analyses highlight that the ACCg of TD individuals exclusively signalled when another person was unexpectedly rewarded; however, this signal is not found in the ACCg of ASD individuals, or in any other brain region.

To investigate whether the absence of BOLD signals in the ACCg was linked to behaviour or ASD social symptom severity, we correlated third person prediction error positive per cent signal change values with behaviour on third person prediction error positive trials and measures of ASD social symptom severity (ADOS social subscale, ADI social subscale, SRS total, and SCQ). There was no significant relationship between third person prediction error positive accuracy and the strength of social prediction errors in the ACCg (r = −0.19, $P = 0.47$). Individuals with a larger social interaction deficit as measured by the ADOS (higher values on the ADOS social subscale) tended to exhibit increasingly positive activity for social prediction error positive trials (Fig. 2E; r = 0.66, $P = 0.007$, FDR corrected $P < 0.05$; all other subscales $P > 0.127$ even though a similar trend was revealed for the ADI social subscale; Supplementary Table 2). A multiple regression analysis confirmed that only the third person prediction error positive BOLD responses (compared to third person predictable outcome positive, predictable outcome negative, prediction error negative BOLD responses) indexed ADOS social values (B = 2.1, SEM = 3.91, $P = 0.02$; all other third person outcomes $P > 0.69$). This demonstrates that the aberrant coding of social prediction errors is linked to deficits in social interaction in ASD as measured by diagnostic instruments.

Do other brain areas signal prediction errors irrespective of whether they are tracked for the first person, third person or a computer and does activity in these areas differ between TD and ASD? A large network of brain regions showed significantly greater activity for prediction error trials compared to predictable outcome trials regardless of the agent, including dorsomedial prefrontal cortex, left posterior superior temporal sulcus (bordering on the temporo-parietal junction), right temporal pole, bilateral parietal lobules and the precuneus (Supplementary Fig. 3 and Supplementary Table 1). However, these regions did not show any differences between agents or groups. Significant group differences in encoding prediction error versus predictable outcome trials were found in the ACCs [MNI coordinates ($x = 9$, $y = 29$, $z = 32$), t = 4.32, k = 54, $P < 0.05$ FWE corrected using TFCE (Smith and Nichols, 2009), assigned to anterior rostral cingulate zone (36%); Neubert *et al.*, 2015)] and 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 (Smith and Nichols, 2009), assigned Area 9/46v (31%; Sallet *et al.*, 2013)], driven by a reduced differentiation between prediction error and predictable outcome trials in individuals with ASD in these regions (Supplementary Fig. 4). Neither of these regions showed a significant relationship with task accuracy for the corresponding contrast (i.e. difference between prediction error and predictable outcome trials), or ASD social symptom severity ($P > 0.12$; Supplementary Table 2); however, this is not surprising given that these group differences were independent of Agent. We also identified the left caudate nucleus and occipital lobe as regions that specifically signalled prediction errors related to first person or Computer trials (Supplementary Figs 5 and 6). Importantly, there were no significant group differences in BOLD activity in these regions between TD and ASD individuals during first person or Computer prediction error trials.

Aberrant ventromedial prefrontal cortex coding of agency in autism spectrum disorders and connectivity to the anterior cingulate gyrus

Given our behavioural effect (Group × Agent interaction), we also investigated whether any brain region encoded differences between Agents, regardless of whether the outcome was expected (predictable outcome) or unexpected (prediction error), and whether this differed between the ASD and

TD groups. Only the vmPFC [MNI coordinates ($x = 3$, $y = 44$, $z = 5$), t = 4.21, k = 38, $P < 0.05$ FWE corrected using TFCE (Smith and Nichols, 2009), assigned to area 32PL (Neubert *et al.*, 2015)] showed a Group × Agent interaction time-locked to the presentation of the privileged information cue (Fig. 3A). BOLD activity in the vmPFC was higher for first person trials compared to third person and Computer trials in TD while there were no differences between agents in ASD (Fig. 3B). The vmPFC interaction effect showed a significant correlation with task accuracy ($r = -0.52$, $P = 0.04$), however this did not survive correction for multiple comparisons ($P > 0.05$, FDR corrected). The vmPFC interaction effect did not correlate with any neuropsychological measures of ASD social symptom severity ($P > 0.346$; see Supplementary Table 2), suggesting that the ACCg may be a more reliable indicator of ASD social symptom severity.

Given that the vmPFC and ACCg are anatomically connected (Yeterian *et al.*, 2012), and both showed group differences in the processing of Agents, we used dynamic causal modelling (DCM12) to investigate whether differences in effective connectivity between the vmPFC and ACCg might contribute to deficits in representing social prediction errors in ASD versus TD. In the TD group, there was a significant increase in connectivity from the vmPFC to the ACCg for third person but not for first person or Computer prediction error trials, while ASD individuals showed no difference between the first and third person Agent but exhibited a boost in connectivity for the Computer prediction error trials [Fig. 3C; Group × Agent interaction $F(2,68) = 7.13$, $P = 0.002$, $\eta p^2 = 0.173$]. TD individuals who showed a greater increase in connectivity from the vmPFC to the ACCg during third person prediction error positive trials also showed a larger social prediction error response in the ACCg ($r = -0.46$, $P = 0.04$; Fig. 3D) but this was not the case in ASD ($r = 0.15$, $P = 0.59$; Fig. 3E). There was also no relationship between vmPFC to ACCg connectivity and ACCg activity for Computer prediction error trials in either group (ASD $r = -0.046$, $P = 0.87$; TD: $r = -0.047$, $P = 0.84$). The strength of connectivity from the vmPFC to the ACCg did not correlate with any neuropsychological measures of ASD social symptom severity or behaviour ($P > 0.426$; Supplementary Table 2). To establish whether this pattern

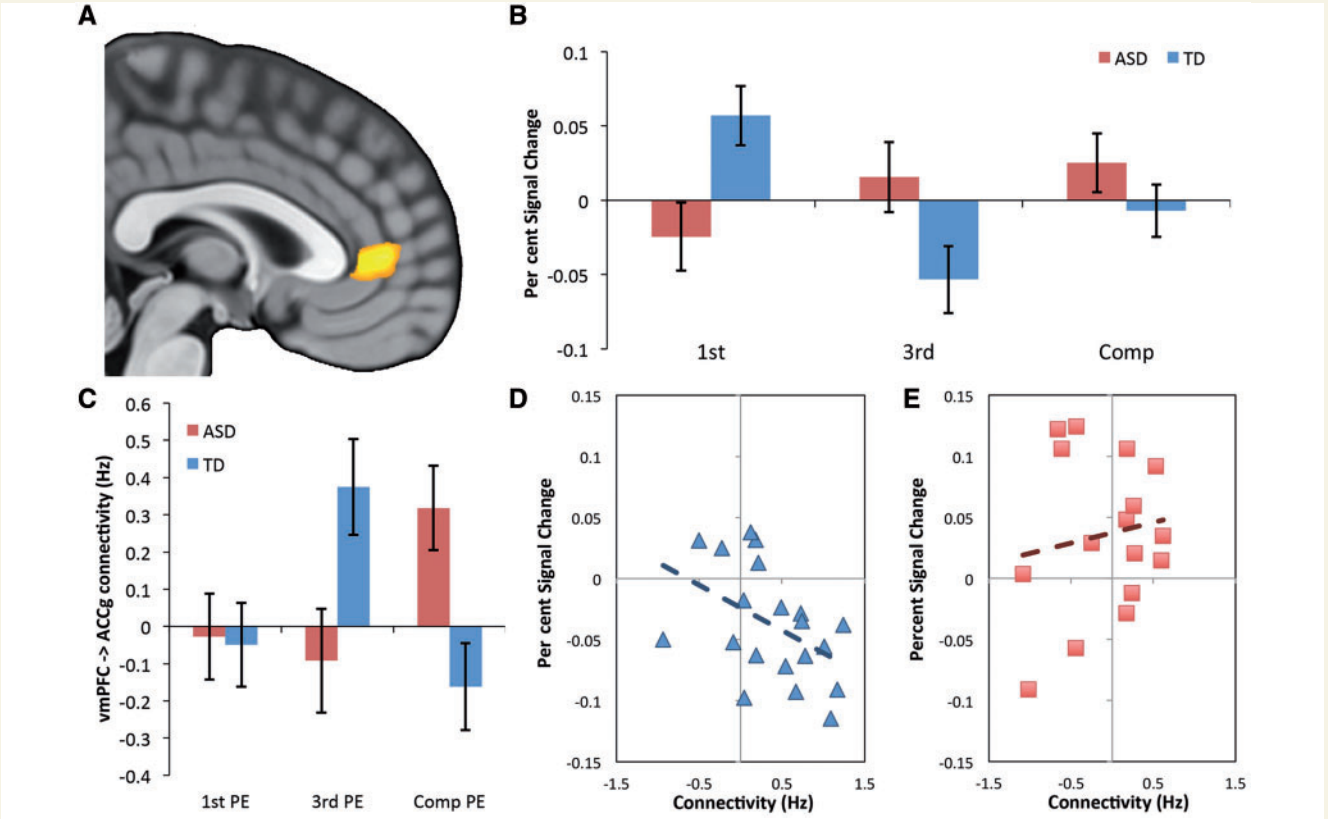

Figure 3 Group differences encoding Agency and connectivity with the ACCg. (**A**) Group × Agent interaction in vmPFC [Area 32PL (Neubert *et al.*, 2015); MNI: 3 44 5, thresholded at $P < 0.001$ uncorrected] time-locked to privileged information (Fig. 1B). (**B**) Bar plots illustrating percent signal change [prediction error (PE) + predictable outcome] in vmPFC for TD and ASD. The activation in vmPFC was driven by an Agent effect in TD (first person > third person and Computer) that is not present in ASD. Error bars indicate standard error. (**C**) Bar plots illustrating that a boost in connectivity from the vmPFC to the ACCg specifically for third person (social) prediction errors in the TD group. (**D** and **E**) Correlations between effective connectivity strength social prediction errors in TD (**D**) and ASD (**E**). This shows a significant correlation between connectivity from the vmPFC to the ACCg and the strength of social prediction errors in TD ($r = -0.46$), which is absent in ASD ($r = 0.15$).

of connectivity between the vmPFC–ACCg could underlie the group difference in social prediction errors seen in the ACCg, we regressed out vmPFC–ACCg connectivity from ACCg activity profile. Initially, there was a significant group difference in ACCg activity [$t(34) = 3.65$, $P = 0.0009$]; however, this group difference was no longer present after regressing out vmPFC–ACCg connectivity [$t(34) = -1.86$, $P = 0.07$]. These results suggest that the absence of social prediction error signals in the ACCg may be due to a lack of input from the vmPFC specifically related to prediction errors arising from the outcomes of others behaviour.

Discussion

Computational approaches to social cognition are providing novel perspectives for understanding deficits in social interaction (Chiu *et al.*, 2008; Behrens *et al.*, 2009; Diaconescu *et al.*, 2014; Sevgi *et al.*, 2015). For example, Sevgi *et al.* (2015) recently showed that neurotypical individuals with higher autistic traits failed to utilize social information during a reward-based learning task. Our paradigm was designed to isolate a different computational mechanism, social prediction errors, which may play a crucial role in understanding the perspectives of others (Apps *et al.*, 2013*b*, 2016). A cardinal characteristic of ASD is the inability to understand the perspectives of others (Baron-Cohen *et al.*, 1985; Baron-Cohen, 1997; Peterson *et al.*, 2013; but also see Senju *et al.*, 2009), as demonstrated by false belief paradigms such as the Sally-Anne task (Wimmer and Perner, 1983). The Sally-Anne task consists of two main elements: (i) establishing a prediction or belief about Sally's expected outcome (i.e. Sally believes the ball is in the basket); and (ii) seeing Anne move the ball in Sally's absence so that Sally's expectation no longer matches the actual outcome. The corner stone of the Sally-Anne task is the ability to recognize that Sally's predicted outcome is not the same as the actual outcome, i.e. a social prediction error. As in the Sally-Anne task, we show that individuals with ASD are less accurate at monitoring the expectations and outcomes of other Agents (Baron-Cohen *et al.*, 1985; Baron-Cohen, 1997; Peterson *et al.*, 2013). We therefore suggest that our paradigm is tapping into the same mechanism that has been repeatedly shown to be perturbed in ASD. However, the design of our task allowed us to directly compare expected and unexpected outcomes, and when outcomes were better or worse than predicted, across multiple Agents. Using this approach we could specifically isolate BOLD activity time-locked to the unexpected outcomes of others' decisions (i.e. social prediction errors) with greater precision than classical paradigms.

A number of brain regions showed greater activity for prediction errors across all Agents (e.g. dorsomedial prefrontal cortex, inferior parietal lobules, and the posterior superior temporal sulcus), or prediction errors specific for first person and Computer trials (caudate nucleus and occipital lobe). However, none of these regions showed differences between TD and ASD individuals. Group differences were found in the ACCs and right inferior frontal gyrus (putatively Area 9/46v; Sallet *et al.*, 2013), specifically showing a reduced differential response between prediction error and predictable outcome trials in ASD. However, these differences could not explain variation in ASD social symptom severity and likely reflect a general cognitive deficit. This is in keeping with studies that have shown that the ACCs encodes prediction errors for all agents in neurotypical cohorts (Behrens *et al.*, 2007, 2008; Matsumoto *et al.*, 2007; Apps *et al.*, 2013*b*; Chang *et al.*, 2013), while the ACCg exclusively encodes social prediction errors.

There has recently been some debate about whether it is even necessary to distinguish between the first person and third person perspective, and whether more valuable information might be gained by focusing on the interaction between agents, the so-called second person perspective (Schilbach *et al.*, 2013; Schilbach, 2016). Here, in line with previous studies (Behrens *et al.*, 2008; Apps *et al.*, 2012, 2013*a*; Lockwood *et al.*, 2015), we found that the ACCg exclusively encoded the unexpected outcomes of another agent's decision (i.e. social prediction errors) in TD individuals, and crucially, we demonstrated for the first time that individuals with ASD do not elicit social prediction errors in the ACCg or in any other brain region. These findings suggest that individuals with ASD have a deficit in understanding the perspectives of others in the absence of social interaction, and that understanding the first and third person perspectives in isolation could help to further inform impairments in social interaction and the second person perspective.

Although there were fewer ASD individuals in this study then TD individuals, we do not believe the absence of social prediction errors in the ASD group was due to lack of statistical power. First, previous neuroimaging studies have shown evidence of social prediction errors consistently in the same ACCg region using smaller samples ($n = 12$, 14 and 15) of TD individuals (Apps *et al.*, 2012, 2013*a*, 2015). In addition, in a recent single-unit recording study in humans (Hill *et al.*, 2016), social prediction errors were identified in the ACCg when only 10 patients were present. Although the method is different, this suggests that social prediction errors may be a reliable property of the response in this region, and only a small number of TD individuals are required to elicit this response. Second, a power calculation on our data confirmed that in TD individuals only 11 participants were required to achieve 80% statistical power with alpha equal to 0.05. Third, we show that in the ASD population there is a strong correlation between social symptom severity and ACCg activity. This highlights that ASD individuals who showed a 'TD-like' ACCg response profile for social prediction errors, were also the ones with lower social symptom severity. We believe this correlation gives more weight to the conclusion

that the absence of an ACCg response in ASD was related to social deficits rather than statistical power.

Interestingly, variability in the strength of social prediction errors was the strongest index of social symptom severity in ASD: ASD individuals who showed increasingly negative ACCg BOLD responses (similar to the TD individuals) showed reduced social symptom severity while those who showed an absent or a slightly positive ACCg signal showed greater social symptom severity. This corresponds to studies in neurotypical cohorts showing that variability in the strength of ACCg signals correlated with social value (Behrens *et al.*, 2008), social group size (Sallet *et al.*, 2011), and the ability to track another individuals actions (Zhu *et al.*, 2012). However, magnitude of the ACCg response is only one aspect relevant for social cognition. Another is the selectivity of the ACCg signal, i.e. whether ACCg activity is exclusively elicited by third person prediction errors or whether first and third person prediction errors can both evoke activity in the ACCg. In a cohort of neurotypical individuals, Lockwood *et al.* (2015) showed that the ACCg signal in individuals high in emotion contagion (individuals high in empathy) was specialized for processing others' rewards exclusively, but for those low in emotion contagion, activity in the ACCg was evoked by information about the others' rewards as well as the subject's own rewards. There is no evidence in our data to suggest that Agent selectivity was perturbed in ASD (i.e. neither ASD or TD individuals showed first person or Computer prediction errors in the ACCg; Fig. 2B, Supplementary Figs 1 and 2), instead ASD individuals presented with an absence of social prediction error signals from the ACCg. This suggests that the specific response pattern elicited by the ACCg (i.e. low magnitude of the ACCg signal but no abnormal Agent selectivity) might distinguish social deficits present in ASD pathology from social impairments in other disorders or personality traits.

While signals from the ACCg specifically reflect the unexpected outcomes of others (social prediction errors), signals from the vmPFC appeared to distinguish between outcomes relevant for the self (first person) and those relevant for others (third person and Computer) regardless of whether they were expected or unexpected, which is in line with previous findings (Murray *et al.*, 2012, 2015). Importantly, our results show that BOLD activity in the vmPFC distinguished between Agents (first person > third and Computer) in the TD group but not in the ASD group. Lombardo *et al.* (2010) similarly showed significantly different BOLD responses to self-relevant judgments compared to judgments about others in the vmPFC of TD individuals but not ASD. Studies of reward processing in ASD have also shown decreased activation in the vmPFC during the receipt of social and monetary rewards (Dichter *et al.*, 2012; Kohls *et al.*, 2013), suggesting that this region may be hypoactive in ASD. It has been suggested that signal from the vmPFC may modulate activity with other brain regions in order to enhance perception, memory, and decision making that is self-relevant rather than socially relevant (Sui and Humphreys, 2015). Here we found that connectivity from the vmPFC to the ACCg was increased in TD individuals but only for third person outcomes, and that the strength of connectivity from the vmPFC to the ACCg correlated with the strength of social prediction error signals in the ACCg. This would therefore suggest that even though the vmPFC differentiates between first person trials compared to other agents, signals from this region promote activity within interconnected regions, which process information for other Agents such as the ACCg. Consistent with this, Behrens *et al.* (2008) found that the vmPFC encoded both the probability of the outcome based on reinforcement history, and the probability of the outcome based on another person's advice. However, individuals who placed greater weight on social information showed increased coupling between the vmPFC and ACCg, whilst those who placed greater weight on reinforcement history showed greater coupling between the vmPFC and the ACCs. We therefore propose that activity in the vmPFC promotes activity in connected regions in order to enhance their related functions, and that an absence of social prediction errors in the ACCg may be a consequence of decreased coupling between the vmPFC and other connected regions.

Although a number of previous studies have investigated social deficits in ASD, the disrupted neural mechanisms underpinning poor social interaction have remained unclear. Remarkably consistent evidence across species (Apps *et al.*, 2013*b*, 2016) has suggested that socially specific reward and prediction error signals in the ACCg might be a crucial component of social behaviour. However, no previous study had translated this research into a clinical context and examined a population with disrupted social behaviour such as individuals with ASD. Here, we showed that aberrant social prediction errors in the ACCg might indeed be a key mechanism underlying social deficits in ASD given the clear association between ACCg signal magnitude and the degree of social deficits within the disorder. Moreover, we show that the vmPFC appears to signal whether information is self-relevant or relevant for others, potentially tuning ACCg for representing socially specific prediction errors. These data provide a novel insight into the neural substrates underlying ASD social symptom severity, and further research into the ACCg, vmPFC, and connectivity between these regions could provide more targeted therapies to help ameliorate social deficits in ASD.

Acknowledgements

We would like to thank Dr Daniel Wooley, Dr Kathy Ruddy, and Dr Valerio Zerbi for their helpful advice on this manuscript. We would also like to thank Dr Kerskens and Mr Josephs for their help in establishing the functional MRI protocol and collecting data.

Funding

This work was supported by funding from the Irish Research Council's New Foundation Scheme (awarded to L.G.). M.A.J.A. is supported by a BBSRC Anniversary Future Leader Fellowship (BB/M013596/1).

Supplementary material

Supplementary material is available at *Brain* online.

References

Apps MAJ, Balsters J, Ramnani N. The anterior cingulate cortex: monitoring the outcomes of others' decisions. Soc Neurosci 2012; 7: 424–35.

Apps MAJ, Green R, Ramnani N. Reinforcement learning signals in the anterior cingulate cortex code for others' false beliefs. Neuroimage 2013a; 64: 1–9.

Apps MAJ, Lesage E, Ramnani N. Vicarious reinforcement learning signals when instructing others. J Neurosci 2015; 35: 2904–13.

Apps MAJ, Lockwood PL, Balsters J. The role of the midcingulate cortex in monitoring others' decisions. Front Neurosci 2013b; 7: 251.

Apps MAJ, Ramnani N. The anterior cingulate gyrus signals the net value of others' rewards. J Neurosci 2014; 34: 6190–200.

Apps MAJ, Rushworth MFS, Chang SWC. The anterior cingulate gyrus and social cognition: tracking the motivation of others. Neuron 2016; 90: 692–707.

Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005; 26: 839–51.

Balsters J, Mantini D, Apps MAJ, Eickhoff SB, Wenderoth N. Connectivity-based parcellation increases network detection sensitivity in resting state fMRI: an investigation into the cingulate cortex in autism. Neuroimage Clin 2016; 11: 494–507.

Barbas H, Ghashghaei H, Dombrowski SM, Rempel-Clower NL. Medial prefrontal cortices are unified by common connections with superior temporal cortices and distinguished by input from memory-related areas in the rhesus monkey. J Comp Neurol 1999; 410: 343–67.

Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? Cognition 1985; 21: 37–46.

Baron-Cohen S. Mindblindness. Cambridge, MA: MIT Press; 1997.

Behrens TEJ, Hunt LT, Rushworth MFS. The computation of social behavior. Science 2009; 324: 1160–4.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature 2008; 456: 245–9. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature07538.html

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci 2007; 10: 1214–21.

Chai XJ, Castañón AN, Ongür D, Whitfield-Gabrieli S. Anticorrelations in resting state networks without global signal regression. Neuroimage 2012; 59: 1420–8.

Chang SWC, Gariépy J-F, Platt ML. Neuronal reference frames for social decisions in primate frontal cortex. Nat Neurosci 2013; 16: 243–50.

Chiu PH, Kayali MA, Kishida KT, Tomlin D, Klinger LG, Klinger MR, et al. Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. Neuron 2008; 57: 463–73.

Constantino JN, Davis SA, Todd RD, Schindler MK, Gross MM, Brophy SL, et al. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord 2003; 33: 427–33.

Constantino JN, Todd RD. Intergenerational transmission of subthreshold autistic traits in the general population. Biol Psychiatry 2005; 57: 655–60.

Di Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, Milham MP. Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis. Biol Psychiatry 2009; 65: 63–74.

Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, et al. Inferring on the intentions of others by hierarchical Bayesian learning. PLoS Comput Biol 2014; 10: e1003810.

Dichter GS, Felder JN, Green SR, Rittenberg AM, Sasson NJ, Bodfish JW. Reward circuitry function in autism spectrum disorders. Soc Cogn Affect Neurosci 2012; 7: 160–72.

Diedrichsen J, Shadmehr R. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage 2005; 27: 624–34.

Eickhoff SB, Heim S, Zilles K, Amunts K. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. Neuroimage 2006; 32: 570–82.

Eickhoff SB, Paus T, Caspers S, Grosbras M-H, Evans AC, Zilles K, et al. Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. Neuroimage 2007; 36: 511–21.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, et al. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 2005; 25: 1325–35.

Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage 2003; 19: 1273–302.

Haber SN, Kunishio K, Mizobuchi M, Lynd-Balta E. The orbital and medial prefrontal circuit through the primate basal ganglia. J Neurosci 1995; 15: 4851–67.

Hill MR, Boorman ED, Fried I. Observational learning computations in neurons of the human anterior cingulate cortex. Nat Commun 2016; 7: 12722.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage 2012; 62: 782–90.

Kohls G, Schulte-Rüther M, Nehrkorn B, Müller K, Fink GR, Kamp-Becker I, et al. Reward system dysfunction in autism spectrum disorders. Soc Cogn Affect Neurosci 2013; 8: 565–72.

Lawson RP, Rees G, Friston KJ. An aberrant precision account of autism. Front Hum Neurosci 2014; 8: 302.

Lockwood PL, Apps MAJ, Roiser JP, Viding E. Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. J Neurosci 2015; 35: 13720–7.

Lombardo MV, Chakrabarti B, Bullmore ET, Sadek SA, Pasco G, Wheelwright SJ, et al. Atypical neural self-representation in autism. Brain 2010; 133: 611–24.

Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, et al. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 2000; 30: 205–23.

Lord C, Rutter M, Le Couteur A. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 1994; 24: 659–85.

Lynd-Balta E, Haber SN. Primate striatonigral projections: a comparison of the sensorimotor-related striatum and the ventral striatum. J Comp Neurol 1994; 345: 562–78.

Matsumoto M, Matsumoto K, Abe H, Tanaka K. Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 2007; 10: 647–56.

Murray RJ, Debbané M, Fox PT, Bzdok D, Eickhoff SB. Functional connectivity mapping of regions associated with self- and other-processing. Hum Brain Mapp 2015; 36:1304–24

Murray RJ, Schaer M, Debbané M. Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and

shared neural activation between self- and other-reflection. Neurosci Biobehav Rev 2012; 36: 1043–59.
Neubert F-X, Mars RB, Sallet J, Rushworth MFS. Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. Proc Natl Acad Sci USA 2015; 112: E2695–704.
Ouden den HE, Ouden den HEM, Daunizeau J, Roiser J, Friston KJ, Stephan KE. Striatal prediction error modulates cortical coupling. J Neurosci 2010; 30: 3210–19.
O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron 2003; 38: 329–37.
Peterson CC, Slaughter V, Peterson J, Premack D. Children with autism can track others' beliefs in a competitive game. Dev Sci 2013; 16: 443–50.
Rudebeck PH, Buckley MJ, Walton ME, Rushworth MFS. A role for the macaque anterior cingulate gyrus in social valuation. Science 2006; 313: 1310–12.
Rutter M, Bailey A, Lord C. SCQ: social communication questionnaire. Los Angeles: Western Psychological Services; 2003.
Sallet J, Mars RB, Noonan MP, Andersson JL, O'Reilly JX, Jbabdi S, et al. Social network size affects neural circuits in macaques. Science 2011; 334: 697–700.
Sallet J, Mars RB, Noonan MP, Neubert FX, Jbabdi S, O'Reilly JX, et al. The organization of dorsal frontal cortex in humans and macaques. J Neurosci 2013; 33: 12255–74.
Schilbach L, Timmermans B, Reddy V, Costall A, Bente G, Schlicht T, et al. Toward a second-person neuroscience. Behav Brain Sci 2013; 36: 393–414.
Schilbach L. Towards a second-person neuropsychiatry. Philos Trans R Soc Lond B Biol Sci 2016; 371: 20150081.
Schultz W. Behavioral dopamine signals. Trends Neurosci 2007; 30: 203–210.
Seltzer B, Pandya DN. Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. J Comp Neurol 1989; 281: 97–113.
Senju A, Southgate V, White S, Frith U. Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. Science 2009; 325: 883–5.
Sevgi M, Diaconescu AO, Tittgemeyer M, Schilbach L. Social bayes: using bayesian modeling to study autistic trait-related differences in social cognition. Biol Psychiatry 2015; 80: 112–19.
Simms ML, Kemper TL, Timbie CM, Bauman ML, Blatt GJ. The anterior cingulate cortex in autism: heterogeneity of qualitative and quantitative cytoarchitectonic features suggests possible subgroups. Acta Neuropathol 2009; 118: 673–84.
Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 2009; 44: 83–98.
Stephan KE, Stephan KE, Penny WD, Moran RJ, Ouden den HEM, Daunizeau J, et al. Ten simple rules for dynamic causal modeling. Neuroimage 2010; 49: 3099–109.
Sui J, Humphreys GW. The integrative self: how self-reference integrates perception and memory. Trends Cogn Sci 2015; 19: 719–28.
Torta DM, Cauda F. Different functions in the cingulate cortex, a meta-analytic connectivity modeling study. Neuroimage 2011; 56: 2157–72.
Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, et al. Precise minds in uncertain worlds: predictive coding in autism. Psychol Rev 2014; 121: 649–75.
Vogt BA, Pandya DN. Cingulate cortex of the rhesus monkey: II. Cortical afferents. J Comp Neurol 1987; 262: 271–89.
Vossel S, Weidner R, Driver J, Friston KJ, Fink GR. Deconstructing the architecture of dorsal and ventral attention systems with dynamic causal modeling. J Neurosci 2012; 32: 10637–48.
Wechsler D. Wechsler abbreviated scale of intelligence. New York, NY: The Psychological Corporation: Harcourt Brace & Company; 1999.
Wechsler D. Wechsler intelligence scale for children, (WISC-IV). 4th edn. San Antonio, TX: The Psychological Corporation; 2003.
Wimmer H, Perner J. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 1983; 13: 103–28.
Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage 2014; 92: 381–97.
Yeterian EH, Pandya DN, Tomaiuolo F, Petrides M. The cortical connectivity of the prefrontal cortex in the monkey brain. Cortex 2012; 48: 58–81.
Yeterian EH, Pandya DN. Prefrontostriatal connections in relation to cortical architectonic organization in rhesus monkeys. J Comp Neurol 1991; 312: 43–67.
Zhu L, Mathewson KE, Hsu M. Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proc Natl Acad Sci USA 2012; 109: 1419–24.