Social Cognitive and Affective Neuroscience Advance Access published November 19, 2012

doi:10.1093/scan/nss079 SCAN (2012) 1 of 10

Specific neural correlates of successful learning and adaptation during social exchanges

Adam P.R. Smith-Collins, Chiara Fiorentini, Esther Kessler, Harriet Boyd, Fiona Roberts, and David H. Skuse
Behavioural and Brain Sciences Unit, UCL Institute of Child Health, University College London, 30 Guildford Street, London WC1N 1EH, UK

Cooperation and betrayal are universal features of social interactions, and knowing who to trust is vital in human society. Previous studies have identified brain regions engaged by decision making during social encounters, but the mechanisms supporting modification of future behaviour by utilizing social experience are not well characterized. Using functional magnetic resonance imaging (fMRI), we show that cooperation and betrayal during social exchanges elicit specific patterns of neural activity associated with future behaviour. Unanticipated cooperation leads to greater behavioural adaptation than unexpected betrayal, and is signalled by specific neural responses in the striatum and midbrain. Neural responses to betrayal and willingness to trust novel partners both decrease as the number of individuals encountered during repeated social encounters increases. We propose that, as social groups increase in size, uncooperative or untrustworthy behaviour becomes progressively less surprising, with cooperation becoming increasingly important as a stimulus for social learning. Effects on reputation of non-trusting decisions may also act to drive pro-social behaviour. Our findings characterize the dynamic neural processes underlying social adaptation, and suggest that the brain is optimized to cooperate with trustworthy partners, rather than avoiding those who might betray us.

Keywords: cooperation; trust; learning; neuroeconomics; fMRI

INTRODUCTION

Social interaction is among the most complex and fundamental of human behaviours. Optimal decision making in social contexts is crucially dependent on the actions of others (Rilling *et al.*, 2008). Cooperation typically benefits all parties in a social interaction, and there is evidence for evolutionary advantages to pro-social preferences (Axelrod, 1984; Gintis, 2000; Hay, 2009). However, cooperation with social partners who act selfishly may result in gains for the uncooperative partner at a cost to the cooperator. Therefore, appropriate social decision making requires us to build a dynamic representation of the mental state and likely actions of those with both partners with whom we are interacting for the first time, and those who have previously been encountered. Successfully selecting those social partners who will reciprocate trust and cooperate is crucial to success.

A widely used model of decision making in the course of social interaction is the 'Trust Game', a sequential reciprocal exchange paradigm (McCabe *et al.*, 2001). One partner acts as the investor, choosing whether or not to transfer some of their money to another (the trustee). If a transfer is made, the amount is multiplied and the trustee chooses either to cooperate and return a proportion of the money, or to betray the investor and keep everything.

Previous imaging studies have used the trust game as a paradigm to investigate neural activity that is associated with decision making by investors (Delgado *et al.*, 2005; King-Casas *et al.*, 2005; Singer *et al.*, 2006; Tomlin *et al.*, 2006) and trustees (King-Casas *et al.*, 2005; Tomlin *et al.*, 2006), anticipation of trustee responses by investors (McCabe *et al.*, 2001), feedback of the trustee responses to investors (De Quervain *et al.*, 2004; Phan *et al.*, 2010) and decisions to punish uncooperative partners (De Quervain *et al.*, 2004). Decision making during trust games leads to engagement of brain regions associated with reward processing, conflict resolution and representation of mental states. Neural responses to partners may be modulated by expectations of a partner's trustworthiness, based on explicit information about reputation (Delgado *et al.*, 2005), or previous experience of their actual behaviour (Singer *et al.*, 2006; Krueger *et al.*, 2007; Chang *et al.*, 2010; Phan *et al.*, 2010).

Decision making by investors in the trust game relies upon modelling the expected actions of a trustee based on the available information. In the absence of information about a trustee's previous behaviour or reputation, trust decisions are influenced by implicit expectations. Partners whose faces are judged as 'trustworthy' attract greater investment, even if they are no more likely to cooperate (van't Wout and Sanfey, 2008). Experience gained during encounters with partners interacts with expectations of their behaviour, such that there is modulation of neural processing when experience differs from the expected outcome. These modulatory effects may reflect signalling of prediction errors, directly or indirectly, neural signals which support learning and modification of future behaviour in response to outcomes (Schultz *et al.*, 1997; O'Doherty *et al.*, 2003a, 2004). Regional neural activity during trust exchanges may also reflect strategic effects. For example, pairs of subjects alternating as investors and trustees in a trust game may show regional neural activity reflecting whether subjects develop a strategy based upon 'unconditional' or 'conditional' trust, dependent on initial expectations of mutual cooperation or self-interest, respectively (Krueger *et al.*, 2007).

While previous studies have examined neural activity associated with generating and updating representations of a social partner, they have not addressed how that neural activity translates to subsequent modifications in behaviour. The ability to learn from social experience, to dynamically update representations of partners, and adjust behaviour appropriately during future encounters, is fundamental to successful social interaction.

We were interested in determining how the brain engages in the process of updating expectations and partner representations during social exchanges. We typically lack explicit information about whether to trust another individual when we first encounter them. To optimize

Received 25 May 2011; Accepted 20 July 2012

The authors wish to thank Oliver Josephs and the Birkbeck/UCL Centre for Neuroimaging for technical assistance. A.P.R.S.-C, E.K. and D.H.S. conceived the study. A.P.R.S.-C, H.B. and F.R. collected and analysed pilot data. A.P.R.S.-C designed the fMRI experiment. A.P.R.S.-C, C.F. and H.B. ran the experiment. A.P.R.S.-C analysed the data. A.P.R.S.-C and D.H.S. wrote the paper, with comments from C.F., E.K. and H.B. This work was supported by funding from the European Commission's Sixth Framework Programme (FP6/2002-2006) under GEBACO [grant number 28 696]. A.P.R.S.-C. is supported by an Academic Clinical Fellowship from the National Institute of Health Research.

Correspondence should be addressed to Adam P. R. Smith-Collins, Behavioural and Brain Sciences Unit, UCL Institute of Child Health, University College London, 30 Guildford Street, London WC1N 1EH, UK. E-mail: adamprsmith@gmail.com

our long-term outcomes, we must learn from making incorrect decisions and update our representations of individuals to guide choices made on future encounters. While the standard trust game paradigm results in an outcome which can allow learning when the investor partner does trust the other partner (i.e. trust is reciprocated or betrayed), one cannot learn from the trials when the investor decided not to trust. However, not engaging with trustworthy partners can lead to opportunities lost as may be evident, for example, when a rival takes advantage by trusting the partner we rejected and realizing the rewards which we have passed up. The value of investing trust in others, even when such a choice may appear risky, is reflected in the enduring fables of such encounters, such as that of Androcles, who removed a thorn from a lion's paw, ultimately resulting in him being saved by the lion from execution in the arena in Rome. We aimed to determine first how the brain learns from errors made during social decision making, and applies this learning to future encounters with particular social partners, and secondly whether this learning differed when decisions resulted in betrayal and loss compared with missed opportunities with trustworthy partners.

Our study employed a simultaneous move, repeated trust game with a large number of different social partners. During their first encounter with each partner, subjects could use only implicit information to judge whether or not to trust each partner. Following this, the intentions of the trustees were revealed, such that subjects could learn from all types of encounter, those with positive (gain, or avoidance of loss) and negative (betrayed trust and loss, or missed opportunity) outcomes. We predicted that effective learning from social encounters would rely upon brain regions associated with reward processing and associative learning, and that there would be valence-specific differences in neural activity, reflecting partially dissociable mechanisms for learning from actual and potential losses during social encounters.

METHODS

Participants

Twenty-seven healthy, right-handed females participated in the experiment. Two were excluded due to excessive movement artefact, and one due to error in recording behavioural data. The remaining twenty-four participants had a mean age of 20.8 years (range 18–25 years). We used female participants, and photographs of females representing playing partners, to avoid potential inter-gender effects on social interaction. All participants gave informed consent, according to the Declaration of Helsinki, and the study was approved by the Research Ethics Committee at University College London.

Experimental design

We used functional magnetic resonance imaging (fMRI) to determine how neural responses to partners' actions related to changes in future behaviour. Subjects acted as the investor partner in two series of trust games with different trustees represented by full-face photographs (Figure 1). In the first series, all trustee partners were novel, while in the second two-thirds of partners had been encountered previously, allowing subjects to use past experience of those individuals to guide decision making. In each game, investors could choose either to trust their partner, risking reward or loss, or to not trust them and avoid these risks. Whether or not subjects decided to invest, they were always informed of the other partner's actual or intended response.

We focussed on neural responses to the four possible outcomes in the first series of trust games (Figure 2). These outcomes were: subject trusts, partner cooperates (expected cooperation, EC); subject trusts, partner betrays (unexpected betrayal, UB); subject does not trust but partner would have cooperated (unexpected cooperation, UC); does not trust but partner would have betrayed (expected betrayal, EB). We examined how neural activity elicited by particular outcomes related to future trust decisions with individual partners. We also investigated changes in neural responses over the course of the experiment.

Participants played the role of the investor partner. They were given an initial endowment of 200 points to use during the games, which were subsequently converted to a monetary reward. In each game, they would have the option to invest 10 points with their partner, or to keep their own points. If they did invest, the trustee partner would either return 20 points, or keep the invested points. If participants kept their points, they could not gain or lose, but were informed of what their partners actions would have been had they invested.

The trustee partners were represented by colour photographs of emotionally neutral female faces (adapted from the Aberdeen set from the Psychological Image Collection at Stirling http://pics.psych.stir.ac.uk). These were scored for emotionality and trustworthiness during a pilot study, and photographs whose ratings deviated significantly from neutral on either scale were removed from the stimulus set. Each photograph represented either a cooperative or uncooperative partner, while across subjects, the assignment of a particular photograph to trustee type was counterbalanced, to obviate specific influences of particular faces on behaviour or neural responses. Overall, each participant encountered 50% cooperative and 50% uncooperative partners.

Stimulus presentation and acquisition of behavioural responses were implemented using the Cogent 2000 (Wellcome Department of Imaging Neuroscience, UCL, London, UK) toolbox for Matlab (The Mathworks Inc.). Stimuli were viewed via a mirror mounted on the MRI head coil, allowing visualization of a projector screen, and subjects responded via a custom-built button box held in the right hand.

The sequence of events during each trust game is shown in Figure 1. Participants first saw the photograph representing the trustee partner, and had 4 s to indicate whether or not they would invest. After a brief jittered delay of 0.5–2.5 s, they were shown the same partner and informed of their response. This outcome image was shown for 7 s, before presentation of a fixation cross for 0.5–2.5 s.

During the first scanning session, participants played 48 trust games with novel partners. In a second series 72 trust games were played, one-third (24 partners) of the trustees were novel, and two-thirds (48 partners) had been seen previously. The previously encountered partners always responded in the same manner as they had on first encounter, although this was not explicitly stated to the participants. A third scanning session consisted of a control task during which participants did not make trust judgements, but performed a recognition memory task on photographs of female faces, some of which were new and some of which had been seen during the trust games. During this recognition task, subjects were randomly assigned to invest or not with partners, independent of their recognition judgement.

After the experiment, participants points were totalled and converted at a rate of 20 points = £1. This was added to a £5 'show up fee' to determine their total payout.

fMRI data acquisition

Data were acquired using a Siemens 1.5T Avanto MR scanner and a 32-channel head coil (Siemens, Erlangen, Germany). We acquired high-resolution T_1-weighted structural images (1 mm × 1 mm × 1 mm) and T_2*-weighted echo-planar images (40 slices, 2 mm × 2 mm × 2 mm voxels, repetition time (TR) = 3.6 s, field of view (FOV) = 192 mm × 192 mm, slice gap = 1 mm). In order to reduce signal dropout in anterior medial temporal lobe, orbitofrontal cortex (OFC) and ventro-medial prefrontal cortex (VMPFC), each image was acquired with an oblique orientation of 30° relative to the anterior–posterior commissure axis, and with a reversed pre-pulse. This followed comparison of regional signal recovery on our MRI

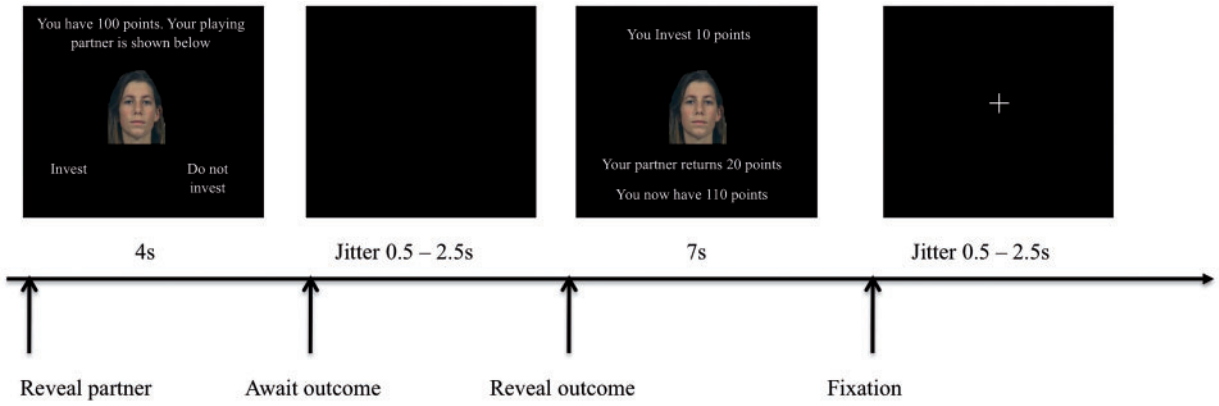

Fig. 1 Stimuli and timing in the trust game.

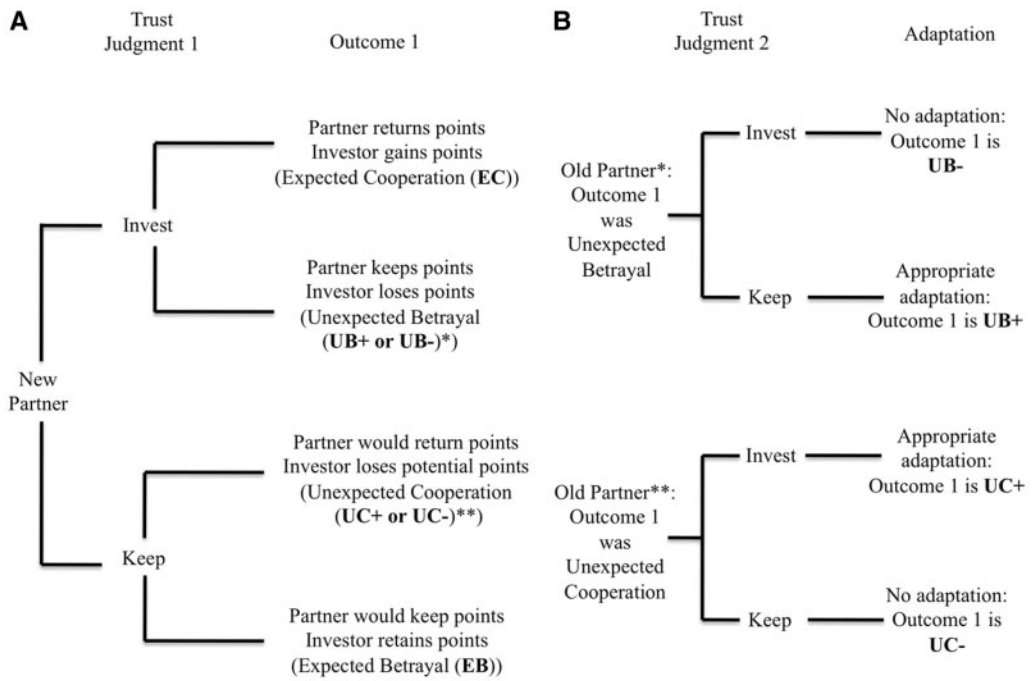

Fig. 2 Decision tree showing the structure for defining outcomes in each trust game. Deciding to trust a partner and invest points could lead either to reciprocation (EC) or betrayal (UB). After a decision to keep points it could be revealed that partners would have cooperated (UC) or would have betrayed (EB) (**A**). For partners where the first trust judgement led to an unexpected outcome (UB or UC), trials were subdivided according to whether investors subsequently adapted their behaviour when playing a second trust game with the same partner (**B**).

system in a pilot group of subjects using the approach described by Weiskopf *et al.* (2006).

Behavioural data analysis

For each subject, data were analysed according to the outcomes experienced by participants when interacting with a particular trustee (see Figure 2 for decision tree). A linear mixed effects model was used to investigate the impact of round, previous choice with a partner, previous response from a partner and the interaction of previous choice and response on the likelihood of investing with old partners when encountering them again in round two. This was supplemented by *post hoc* paired *t*-tests between investment probabilities for different subgroups of old partners according to round one outcome and with new partners. Identical analysis methods were used to compare reaction times for trust judgements, and accuracy of recognition memory.

fMRI analysis

fMRI data were pre-processed and analysed using SPM8 (Wellcome Department of Imaging Neuroscience, UCL, London, UK). After the first two volumes were discarded to account for equilibration effects, images were realigned to the first volume to correct for subject motion, normalized to the Montreal Neurological Institute EPI template with a voxel size of 3 mm × 3 mm × 3 mm and smoothed with a Gaussian kernel (8 mm full-width at half maximum). A high pass filter of 128 s was applied.

Following pre-processing, parameters of a general-linear model in each subject were estimated to generate voxel-wise statistical parametric maps. For each subject we modelled activity associated with the four possible outcomes in the first series of trust games—EC, EB, UC and UB (Figure 2). EC, EB, UC and UB parameters were subdivided according to whether subjects subsequently adapted their trust

behaviour appropriately (+) or did not (−), following feedback from their decision on their first encounter with a social partner. Comparison of trials subsequently eliciting adaptation or not was taken as an index of encoding of social information, analogous to the approach used in subsequent memory paradigms (Fernandez *et al.*, 1999; Otten *et al.*, 2001).

The main regressors of interest were modelled as mini-boxcar functions of 7 s duration, aiming to capture both immediate and delayed neural activity associating the outcome with individual partners. Additional regressors were estimated for the trust decision phases (modelled as event 'stick' functions) and parametric modulators of the change in outcome responses over time (utilizing trial number as the parametric modulator and convolving these function with the outcome mini-boxcar functions). These regressors were convolved with a canonical haemodynamic response function and were included in a design matrix, together with regressors generated from the realignment parameters, to correct for residual subject motion.

Parameter estimates for each subject were taken from this analysis and entered into a random effects (between-subjects) second-level analysis, and linear contrasts used to identify brain regions with differential responses according to outcome, those associated with changes in future behaviour and change in response over time. Statistical parametric maps were reported at a significance threshold of $P < 0.001$, with cluster threshold correction identifying only clusters of at least five contiguous voxels. These were plotted on the average of subjects' anatomical images for structural localization. Contrasts were inclusively masked with the averaged grey matter masks from each subjects, acquired from segmentation during image pre-processing. Where significant effects were revealed, these were quantified by extracting average parameter estimates from the activated cluster in each individual and plotting the mean of these across subjects. To assess whether patterns of neural activity were common between contrasts, we used inclusive masking, which reveals effects that are significant across multiple contrasts, at a threshold of $P < 0.005$.

RESULTS

Behaviour

During the first series of trust games, subjects played only with novel partners. They chose to invest in 56% of the encounters with partners who would subsequently reciprocate trust and 53% of partners who would subsequently betray them (Table 1, panel a). The overall proportion of outcomes experienced in the first series of trust games were: EC 0.28, UC 0.22, EB 0.24 and UB 0.26. The difference in proportion of outcomes experienced did not reach significance across subjects [$F(2.08, 47.8) = 2.74$, $P = 0.073$].

Social learning was indexed by changes in behavioural responses during the trust games, and how these varied according to the nature of previous encounters with playing partners.

In the second round, decisions to invest with particular partners were significantly influenced by previous partner response [$F(1,138) = 7.38$; $P = 0.007$], indicating that subjects were more likely to invest with those previously encountered partners who had indicated willingness to cooperate. Subjects were more likely to repeat their investment choice from series one than to change it in series two {probability 0.56 *vs* 0.44 [$F(1,138) = 26.7$; $P < 0.001$}. There was also a significant interaction between previous choice and previous response [$F(1,138) = 6.01$, $P = 0.015$]. This interaction reflected a difference in modifying investment behaviour with partners who had previously been cooperative or uncooperative (Table 1, panel a; Figure 3). Partners who had previously been cooperative were more likely to attract investment in the second series of trust games than were new partners, regardless of whether the previous cooperation had been expected [EC; investment probability 0.66 *vs* new partners in second series 0.42; $t(23) = 3.00$, $P = 0.006$] or unexpected [UC; investment probability 0.64 *vs* new 0.42; $t(23) = 3.07$; $P = 0.005$].

Similarly, where partners had previously shown that they would have betrayed the investor, and this was consistent with expectations in series one (EB) they were less likely to attract investment in series two [investment probability 0.31 *vs* new 0.42; $t(23) = 2.15$; $P = 0.042$]. In contrast, partners who had previously unexpectedly betrayed trust (UB) were no less likely to attract investment than new partners [investment probability 0.54 *vs* new 0.42; $t(23) = 1.48$; $P = 0.15$], and indeed the trend was towards a higher rate of investment with partners who had previously unexpectedly betrayed trust than with new partners.

The finding that subjects appeared to show less appropriate adaption of behaviour with partners who unexpectedly betrayed them in series one than they did for other outcomes, was not explained by differences in facial recognition memory. There was no significant effect of series one outcome on recognition memory for the faces of the partners in a subsequent recognition memory test [$F(2.68, 61.6) = 0.26$, $P = 0.835$] (Table 1, panel b). There were no significant differences in reaction times for either trust judgements or recognition memory according to the player's prior decision, or the outcome of that decision in round one.

The differences in subsequent investment behaviour, which appeared to show reduced adaptation following UB in round one, compared with other outcomes, implies that investors learned less effectively from UB than from cooperation. There are a number of possible explanations for this finding. The apparent difference in adaptation might simply result from a response bias towards investing during the second series of trust games. However, the overall proportion of 'invest' *vs* 'keep' decisions in the second series of trust games showed no such bias. The respective proportions with old partners did not differ significantly [0.52 invest *vs* 0.48 keep, $t(23) = 0.84$, $P = 0.41$] although there was a distinct bias towards untrusting behaviour with new partners [0.38 *vs* 0.62, $t(23) = 2.75$, $P = 0.01$], arguing against a simple response bias explanation. Interestingly, during the first round of encounters, subjects were more likely than not to invest with partners about whom they had no prior information, while this tendency decreased over time and there was a bias towards untrusting behaviour with new partners by the end of the experiment (Figure 3B). This suggests that there was an initial bias towards investing but that subjects became progressively less trusting over the course of the experiment as group size increased, with new partner investment rates falling substantially below the group reciprocation rate of 0.5. The observed differences in subsequent behaviour may then either reflect differences in initial expectations of cooperative or uncooperative behaviour, or some intrinsic difference in how subjects learned from different types of partner responses.

fMRI

Neural responses to outcomes, following investment decisions in series one, differentiated between expected *vs* unexpected partner responses (i.e. between EC and EB, contrasted with UC, and UB). This contrast highlighted greater activity in the right ventral striatum [12 voxels, peak MNI coordinates (18, 17, −5) $Z = 3.39$, $P < 0.001$ uncorrected; Figure 4A] in response to trustee actions corresponding to investor expectations. No brain region showed significantly greater overall activity following unexpected *vs* expected outcomes.

Our primary aim was to determine how neural activity encoded successful social learning and behavioural adaptation. We focussed on how neural activity in response to the outcome of a trust decision with a particular partner in series one related to what trust

Table 1. Probability of investment with old and new partners in the trust game (top line), and accuracy of recognition memory in the control task (bottom line)

	Old partners				New partners			
	EC	UB	UC	EB	Round 1		Round 2	
					C	B	C	B
Likelihood of investment, Mean (s.e.)	0.66 (0.05)	0.54 (0.05)	0.64 (0.04)	0.31 (0.04)	0.56 (0.02)	0.53 (0.03)	0.43 (0.04)	0.40 (0.04)
Recognition, Mean (s.e.)	0.80 (0.04)	0.78 (0.04)	0.79 (0.05)	0.82 (0.04)	0.90 (0.09)			

Responses with old partners are divided according to previous outcome. Responses with new partners are divided according to round and subsequent partner response (C, cooperation, B, betrayal).

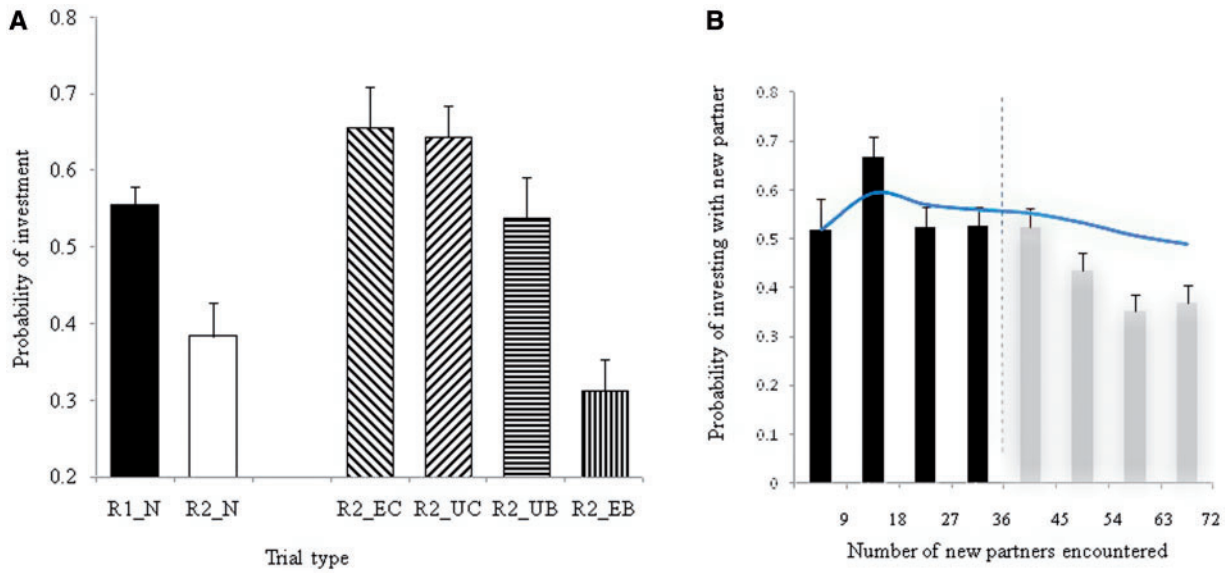

Fig. 3 Behavioural outcomes in the trust game. (**A**) Probability (±s.e.m.) of investing with old and new partners during the first (R1_N) and second (R2_N) series of trust games. Old partners are divided according to their outcome in the first round (EC, UB, UC, EB). EC and UC partners attracted a higher level of investment, and EB a lower level of investment than new partners. (**B**) The change in probability of investing with new partners (black bars: round 1, grey bars: round 2) and the mean cumulative probability (solid line) of investing with 'New' partners as the number of partners increased over the experiment. The dashed line shows the division between the first and second round of trust games.

decision would be made when encountering that partner again in series two. Social learning could be indexed by either repeating a correct response (e.g. investing again with a partner who had previously reciprocated), suggesting that the initial response had reinforced the behaviour, or by changing responses following an unexpected outcome. To determine regional activity associated with reinforcement of responses, we subdivided trials where there was EC in series one into EC+ (series two, participant invested again, suggesting reinforcement of the investment response) and EC− (series two, participant did not invest, no evidence of reinforcement) (Figure 2). Similarly, EB trials were subdivided into EB+ (series two, participant again chose not to invest, suggesting reinforcement) and EB− (series two, participant invested, no evidence of reinforcement). Overall contrast of reinforced *vs* non-reinforced trials did not show any significant differences in regional neural responses. However, contrasting only EC+ with EC− trials showed that successful reinforcement following EC was associated with increased regional activity in the right ventral striatum [5 voxels, peak MNI coordinates (9, 20, −8) $Z = 3.50$, $P < 0.001$ uncorrected] and mid-frontal gyrus [17 voxels, peak MNI coordinates (39, 23, 31) $Z = 4.86$]. Inclusive masking of this 'reinforcement effect' from reciprocated cooperation and the previously reported contrast of effects of positive *vs* negative outcomes, thresholded at $P < 0.005$, revealed an overlap between effects in the right ventral striatum [4 voxels, peak MNI coordinates (18, 14, −5) $Z = 3.17$; Figure 4B]. There were no significant effects revealed by the contrast of EB+ *vs* EB−.

We used a similar approach to examine neural responses associated with successful *vs* unsuccessful behavioural adaptation in series two following unexpected outcomes in series one (Figure 2). UC in round one was subdivided into UC+ (series two, participant invested, reflecting an appropriate adaptive response to UC) and UC− (series two, participant again chose not to invest; no evidence of adaptation). Similarly, UB was subdivided into UB+ trials (series two, did not invest, implying successful adaptation) and UB− (invested during series two despite previous betrayal; no evidence of adaptation).

Appropriate behavioural adaptation in the second series of trust games was associated with specific neural responses to *unexpected outcomes* in the first series of games. Brain regions in which these *adaptation* effects were evident included dorsal striatum, anterior cingulate cortex (ACC), right dorsolateral prefrontal cortex (DLPFC), left OFC and the midbrain in the region of the substantia nigra (Table 2). Regional analysis of parameter estimates revealed that activity in some brain areas, notably the midbrain, was only associated with successful *vs* unsuccessful adaptation following UC trials. Other regional neural activity, such as that in dorsal ACC, was associated with successful adaptation following both UC and UB by trustee partners

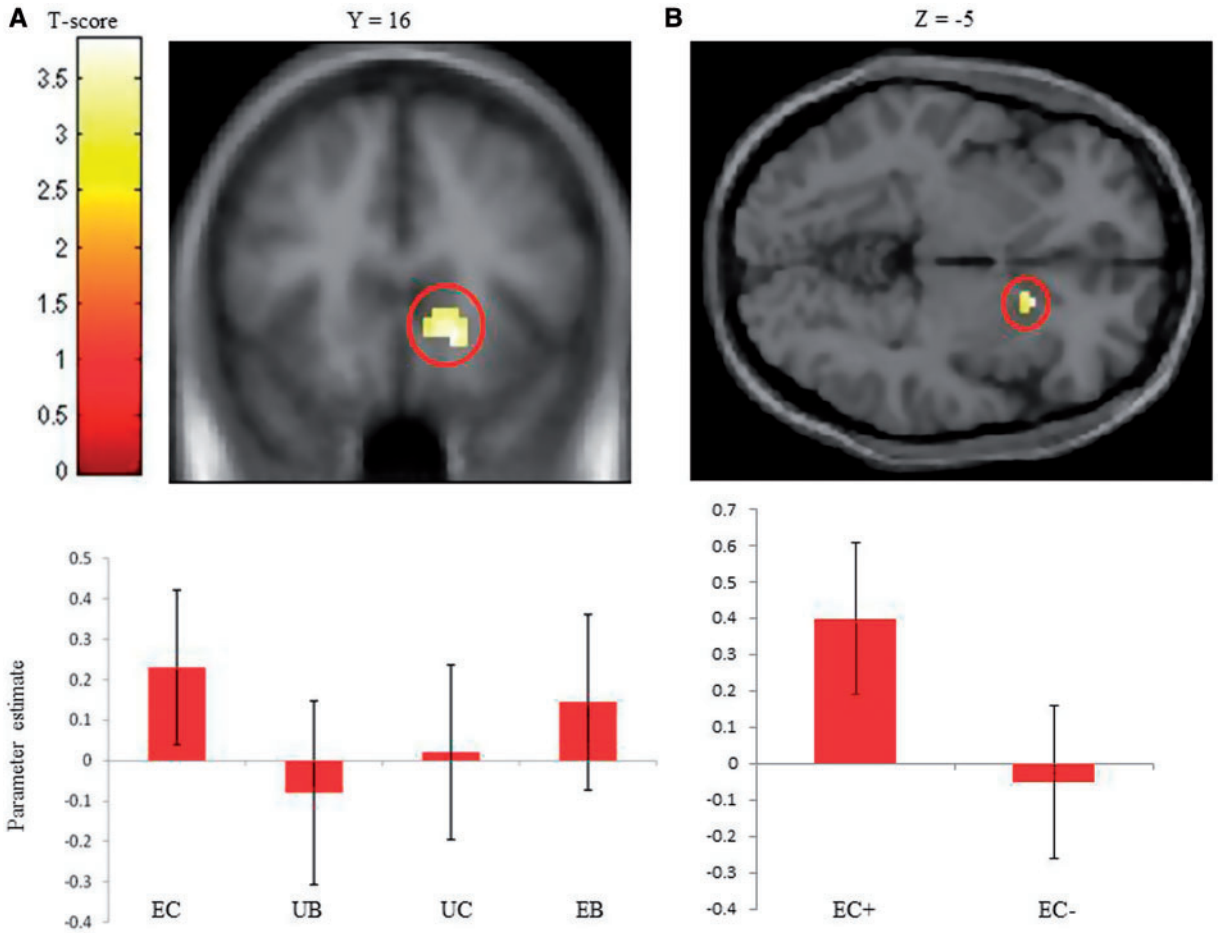

Fig. 4 Activity in ventral striatum showing increased bold signal following successful outcomes during social encounters in the first round of trust games (**A**) and reinforcement of investment behaviour following reciprocation (**B**). Parameter estimates (±s.e.m.) are shown for EC, UB, UC, EB. For (**B**), EC is shown according to whether there was evidence of subsequent reinforcement (EC+) or not (EC−) in the second round of games. Images shown at uncorrected threshold $P < 0.001$.

(Figure 5). Crucially, these 'adaptation effects' demonstrate differential activity during a first social encounter associated with particular behavioural choices made subsequently with that same social partner.

In order to determine whether the adaptation effects in some brain regions were elicited only by either cooperative or uncooperative partners, we compared overall neural responses with UC and UB outcomes from the first series of trust games.

UC *vs* UB elicited greater neural activity in several areas (Table 2). Inclusive masking revealed that these valence-specific cooperation effects overlapped with adaptation effects in dorsal striatum and the midbrain. No brain regions showed significantly greater responses to UB compared with cooperation.

We also analysed whether the neural responses to UC or betrayal changed over the course of the experiment. Given the behavioural evidence for decreasing trust in new partners as the group size increased, we hypothesized that betrayal by a new partner might be progressively less surprising, and cooperation more surprising, and that this would be reflected in differences in neural responses to these outcomes. We predicted that neural responses in some brain regions supporting adaptation might show modulation as more new partners were encountered.

As group size increased, anterior and posterior cingulate, right parahippocampal gyrus, right lateral OFC and left superior frontal gyrus showed relatively decreased responses to UB compared with UC outcomes (Table 2, Figure 6). In parahippocampal gyrus and OFC, these modulatory effects were co-localized with the regional neural activity associated with successful adaptation. We suggest that decreasing neural responses to betrayal, in regions associated with effective adaptation of behaviour, may contribute to an impaired ability to learn from UB. One possibility is that as enlarging group size predisposes investors to be less trusting of new partners, the degree of expectation violation caused by betrayal decreases, with a corresponding reduction in learning. No brain regions showed significant increases in activity over the experiment for UB compared with cooperation. We also directly tested whether restricting the comparison of activity elicited by UB *vs* UC revealed significant effects when only the first 50% of trials were included in the analysis, again without significant regional increases in activity associated with UB.

DISCUSSION

How do we learn and remember who to trust among a huge number of potential social partners? Our findings reveal how effective social learning and behavioural adaptation in a repeated social reciprocation paradigm is associated with specific patterns of neural activity. In this task, we show evidence for greater effectiveness of learning and adaptation following a social interaction that elicits UC than one which results in UB, and we identify corresponding dissociations in patterns of neural activity.

In this experiment, we found that subjects were equally likely to recognize faces of partners who had reciprocated trust during their

first encounter as they were the faces of those who had betrayed them. However, when the same partners were encountered during a second trust game, the likelihood that the decision to trust them would change appropriately from the first encounter depended critically on whether that partner had been unexpectedly trustworthy or untrustworthy. Appropriate behavioural change was significantly more likely following UC.

Table 2. Brain regions (and Brodmann areas) associated with successful social adaptation following unexpected outcomes in a repeated trust game; differentiating UC and UB; and showing differential changes in response to UC and betrayal over time

L/R	BA	Voxels	MNI coordinates	Z-score
Successful > unsuccessful adaptation (UB+ & UC+ > UB− & UC−)				
L	Lingual gyrus (17)	8	(−9, −97, 7)	4.08
L	Middle temporal gyrus (39)	8	(−36, −64, 37)	4.01
L	Superior parietal lobe (7)	13	(−30, −52, 43)	4.15
L	Fusiform cortex (37)	8	(−51, −43, −8)	3.98
R	Midbrain (SN)	5	(9, −13, −20)	4.01
R	Caudate	15	(18, −7, 31)	4.5
R	DLPFC (9)	48	(36, 11, 31)	4.72
L	Anterior cingulate (32)	5	(−12, 23, 37)	4.43
L	Middle frontal gyrus (9)	46	(−39, 29, 19)	4.35
L	Middle frontal gyrus (10)	7	(39, 29, 28)	4.3
L	Lateral OFC (10)	6	(−39, 59, −8)	4.4
UC > UB				
R	Lingual gyrus (18)	9	(12, −73, −5)	3.33
R	Hippocampus	13	(30, −28, −5)	3.84
L/R	Thalamus	5	(0, −22, 1)	3.33
L/R	Midbrain (SN)*	13	(3, −16, −20)	3.52
R	Caudate*	5	(18, −10, 28)	3.19
R	Anterior insula	7	(39, 5, 13)	3.41
Relative increase in UC *vs* UB over time				
L/R	Posterior cingulate (29)	6	(0, −52, 7)	3.33
L	Anterior cingulate (24)	16	(−6, 26, 13)	3.38
R	Anterior cingulate (32)	8	(3, 44, 1)	3.18
L	Superior frontal gyrus (9)	21	(−24, 38, 1)	3.41
R	Lateral OFC (11)**	13	(33, 47, −17)	3.71

Main effects thresholded at $P < 0.001$, uncorrected for multiple comparisons. Extent threshold = 5 voxels.
*Regions showing overlap of A and B in conjunction at $P < 0.005$.
**Regions showing overlap of A and C in conjunction at $P < 0.005$.

A number of possible mechanisms to explain the differences in observed adaptation exist, and previous studies have identified better explicit memory for cooperative than uncooperative partners' actions (Singer *et al.*, 2004). However, other studies (e.g. Mealey *et al.*, 1996) have suggested that uncooperative 'cheat' partners are more salient than 'fair' players, and there is evidence that partners do tend to decrease their rate of investment with uncooperative players in repeated trust games (e.g. Baumgartner *et al.*, 2008). It is notable that recent study of memory for 'fair' and 'unfair' players in a social 'ultimatum game' showed no specific advantage for either group, but revealed that subjects who expected their partner to be fair had improved memory for unfair players, while those who had low expectations remembered fair or generous players better (Chang and Sanfey, 2009). These findings suggest that expectation violation may be a key influence upon learning from social encounters.

Over the course of the present experiment, we observed that subjects had a decreasing willingness to trust new 'stranger' partners, despite the fact that there was no change in the likelihood of trustees engaging in untrustworthy behaviour. Previous studies have predicted that cooperative behaviour tends to develop more easily in small, compared with large, groups (Boyd and Richerson, 1988; Barta *et al.*, 2011; Takezawa and Price, 2010). We hypothesize that, as the experiment progressed, the increasing group size of potential investment partners may have served to promote a prior expectation that 'strangers are likely to be untrustworthy', and resulted in a strategic shift in behaviour to avoid investment with them. Under these circumstances, reciprocated trust by a novel partner becomes less and less expected, and therefore is more salient, as the degree of expectation violation increases. This increased salience could be the explanation for the observed difference in adaptive behaviour, which focused on investing with those who had reciprocated trust, rather than avoiding those who were known to be untrustworthy.

An alternate possibility is that the initial tendency to investment in round one reflects an expected investment probability of greater than

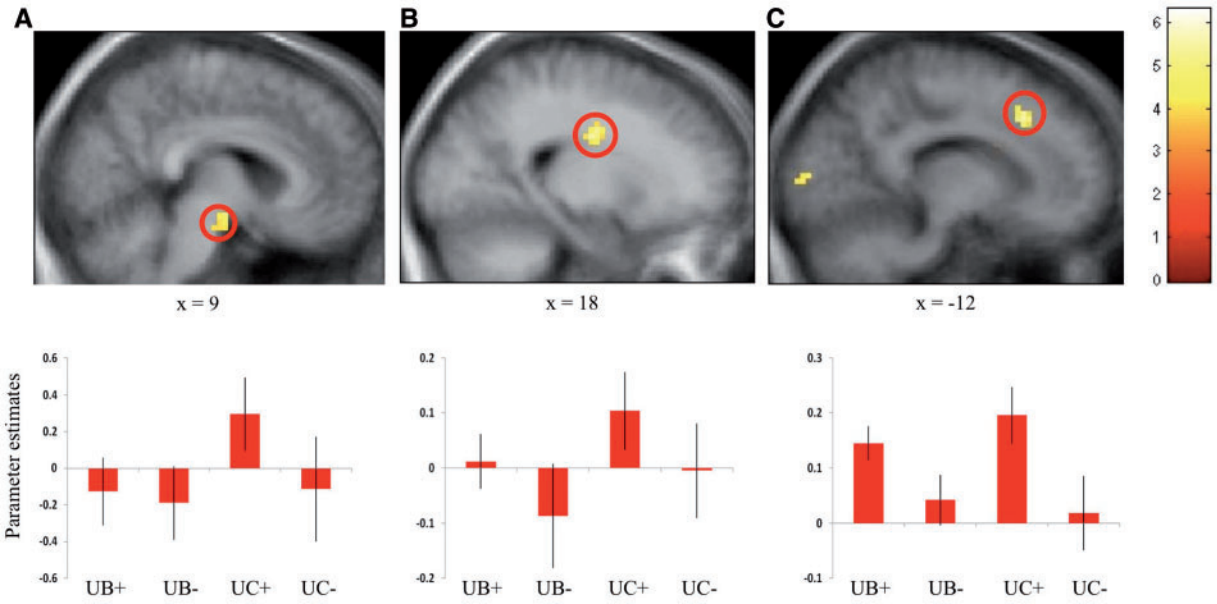

Fig. 5 Brain regions associated with successful adaptation of behaviour following unexpected outcome in a trust game. Contrast of outcomes leading to successful adaptation compared with no adaptation revealed effects in the brainstem (circled, **A**), dorsal caudate (**B**) and anterior cingulate cortex (**C**). Parameter estimates (±s.e.m.) are shown in each region for outcomes of UB leading to adaptation (UB+) or not (UB−) and UC leading to adaptation (UC+) or not (UC−). Images shown at uncorrected threshold $P < 0.001$.

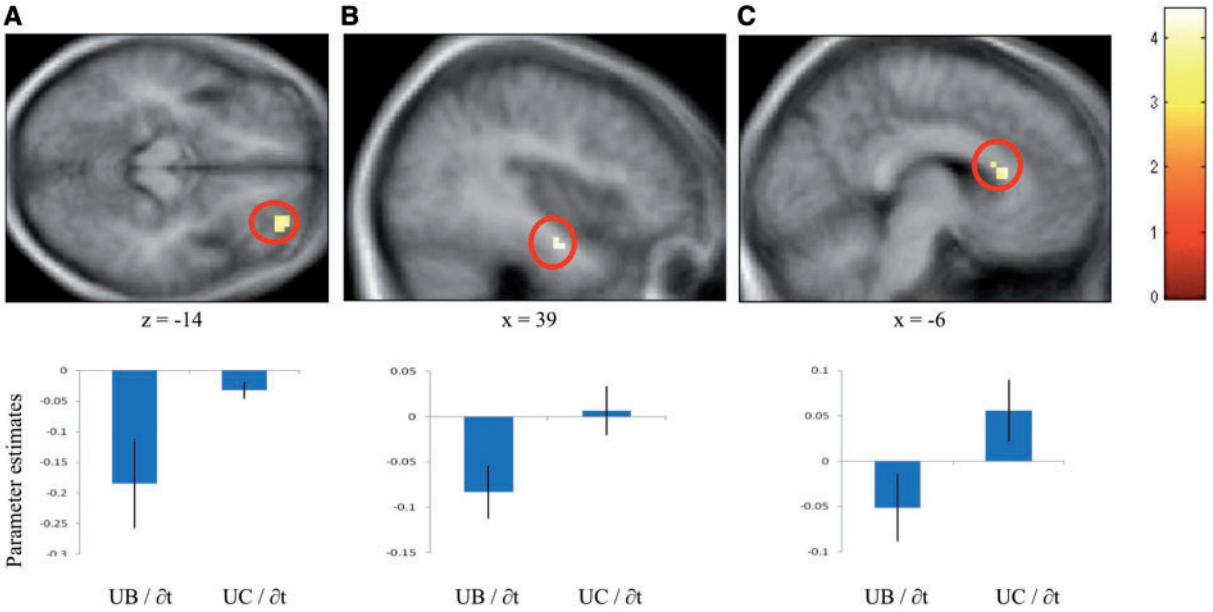

Fig. 6 Change in neural responses to unexpected betrayal and cooperation over time. There were greater decreases in responses to UB (UB/dt) relative to UC (UC/dt) in lateral OFC (**A**, circled), parahippocampal cortex (**B**) and anterior cingulate cortex (**C**). Parameter estimates (±s.e.m.) for the depicted regions are plotted in the lower part of each panel. Images shown at uncorrected threshold $P < 0.001$.

0.5. In such a situation, equivalent learning effects from cooperation or betrayal would result in different observed outcomes. This could result in the probability of subsequent cooperation being regarded as high following an 'UC' outcome, but only being downgraded to intermediate following 'UB'. However, whilst this effect may have contributed to the observed differences, comparing the initial investment rates in round one with the subsequent probabilities of investment following UC and UB trials suggests that an initial positive bias is unlikely to be the only explanation.

A third factor may be that the negative consequences of UC and betrayal, and indeed of choosing to trust a partner or not, differ in this study. Although the points value which was lost in UB encounters was equivalent to the potential loss (or lack of gain) in UC trials was equivalent, choices in social encounters may have additional non-monetary costs (see Fehr and Camerer, 2007). Being betrayed by a partner who you hoped would reciprocate is hurtful, and the 'cost' may exceed the money lost. However, not trusting a potentially cooperative partner results not only in a monetary loss (or lack of gain) for the investor, but also for the social partner, and may result in a loss of reputation—the investor may therefore lose potential future opportunities for collaboration. If in the present experiment, the non-monetary cost of being the 'bad' partner who does not cooperate outweighed the non-monetary cost of being betrayed, that could form a stronger stimulus for social learning and subsequent social adaptation.

Social learning requires the brain to monitor choices and outcomes, to bind these with contextual factors and to use this information to guide future decision making. In the present experiment, social adaptation implies that during a second encounter a subject would be less likely to invest with a trustee by whom they had been betrayed when previously encountered, but would be more likely to invest following cooperation, while social reinforcement suggests implies increased likelihood of investing again following reciprocation and of not investing again with partners who would not have reciprocated if given the opportunity.

One key brain region identified as being involved in social learning is the striatum, which has been widely implicated in reward processing (Delgado *et al.*, 2000; Elliot *et al.*, 2000; Kampe *et al.*, 2001), and responds differentially to trust game outcomes (Delgado et al, 2005; King-Casas *et al.*, 2005; Phan *et al.*, 2010). There is prior evidence of a striatal ventral–dorsal dissociation in responses to unexpectedly positive *vs* unexpectedly negative outcomes (Seymour *et al.*, 2007; Robinson *et al.*, 2010).

In the present experiment, when a social partner responded in accordance with a subject's prediction, i.e. reciprocation from those invested with (EC) or not from those who were not trusted (EB), there is greater engagement of the ventral striatum than when they act contrary to expectations. In a social reciprocal exchange paradigm, both outcomes which result in gains and those which avoid losses can be construed as beneficial to the investor and thus are 'positive'. Although these outcomes are described as 'expected', in that they match the investing subject's prediction, there is a degree of uncertainty in that prediction. Ventral striatal activity may reflect the updating of the predicted likelihood of a partner cooperating, reflecting strengthening of subjects beliefs about a particular partner following 'positive' feedback. This interpretation is supported by the finding that there was greater ventral striatal activity elicited by EC from partners who again attracted investment on subsequent encounters (EC+) compared with those partners who were not trusted during future social interactions (EC−). Further supporting evidence for a role in updating values based on positive feedback comes from a Prisoner's Dilemma paradigm, which also utilized only female subjects, showing that activity in ventral striatum was associated with development of mutual cooperation between players (Rilling *et al.*, 2002).

In contrast, when social outcomes with novel partners were opposed to those predicted, resulting in either an actual (UB) or virtual (UC) loss, activity in the dorsal striatum was associated with subsequent behavioural adaptation, supporting the previously described ventral–dorsal distinction for processing 'positive' and 'negative' feedback and implying distinct functional roles is supporting social learning.

Social adaptation following unexpected outcomes was also associated with neural activity in lateral OFC, which processes information about reward contingencies (O'Doherty *et al.*, 2003b; Tobler *et al.*, 2007) and supports behavioural adaptation following unexpected

negative outcomes during reversal learning (Ghahremani *et al.*, 2010). Similar effects were observed in the dorsolateral prefrontal cortex, a brain region associated with strategic thinking in social interaction games (Yoshida *et al.*, 2010), and with linking reward signals to future behaviour (Wallis and Kennerley, 2010).

ACC activity was also associated with successful social adaptation. This brain region has been implicated in processing of uncertainty and response selection (Critchley *et al.*, 2001; Stern *et al.*, 2010). This role has previously been shown to extend to social learning, with regional ACC activity correlated with reward prediction error in a decision-making task in which subjects could make choices based on both their own experience, and the advice of a social 'confederate' (Behrens *et al.*, 2008). Learning the value of the advice of a social partner was associated with differential regional activity in ACC compared with learning from individual experience in that task. In the present study, the association between increased ACC signal and likelihood of subsequently adapting behaviour supports the concept of this region detecting prediction error in social contexts and supporting behavioural change appropriately.

Three distinct patterns of activity were observed when we compared neural activity following first round encounters that led to UC, compared with UB. The first effect was 'valence specific' and differentiated successful from unsuccessful adaptation in round two following UC, but not betrayal. This pattern was most notable in the midbrain/substantia nigra region. Activity in this region following UC by partners during round one was associated with an adaptive change in behaviour, with increased signal elicited by UC in round one associated with subsequent investment *vs* non-investment decisions during round two.

The substantia nigra has been strongly implicated in reward processing (Herberg *et al.*, 1976; Wittmann *et al.*, 2005), and projects dopaminergic neurons to many brain regions including the striatum (Smith and Kieval, 2000). The ventral tegmental area, which is closely related to substantia nigra, has previously been implicated in making trust decisions conditional on a partner's expected response (Krueger *et al.*, 2007). Although it must be acknowledged that exact anatomical localization of midbrain nuclei is difficult with fMRI, the current findings are consistent with a possible role for dopaminergic projections from the midbrain in the region of the substantia nigra playing a key role in learning about potentially rewarding outcomes and adapting behaviour appropriately.

The second pattern of activity was observed in dorsal striatum. Following both UC and UB during round one, such activity was associated with successful adaptation during round two. As was observed with the effects in the midbrain, striatal activity was greater following UC than UB. Neural activity in the striatum and the midbrain has been extensively linked with signalling of 'prediction errors' (e.g. Schultz *et al.*, 1997; O'Doherty *et al.*, 2003a; 2004). It may be that those encounters that are more surprising, and thus have greater 'prediction error' are more likely to be associated with subsequent adaptations in behaviour. Although the difference in proportions was not significant, there were fewer cases of UC than of UB during the first round of trust games in the present experiment. This may have increased the prediction error and hence the subsequent learning from those encounters. However, the present findings may also be interpreted as showing that striatal activity more closely reflects effective learning and adaptation effects, which are a downstream consequence of prediction error signalling, rather than reflecting the size of the prediction error *per se*. In this case, previous studies may have found striatal effects to be larger with increasing prediction error because those trials were more likely to generate learning or adaptation in response.

A third group of brain regions, including the right lateral OFC, also showed patterns of activity that discriminated between successful and unsuccessful adaptation to outcomes of UC or UB during the first series of trust games, but additionally showed a modulation of activity over the experiment. UB elicited decreasing activity in these brain regions as the experiment progressed and the number of trustee partners encountered increased. Reducing expectations of trustworthy behaviour by partners during the course of the experiment paralleled this modulation of regional neural activity.

We propose that, as expectations of partners' actions are lowered, betrayal becomes less surprising and the degree of expectation violation reduces. The corresponding decrease in regional neural responses associated with successful behavioural adaptation implies that learning from betrayal may decrease as a result of altered expectations with increasing group size. Taken together, the changes in regional neural responses, adaptation effects and response bias imply that different strategic approaches to reciprocal exchange may optimize individual outcomes in groups of different sizes, and that the social brain adapts to facilitate this.

As noted in the methods section, the current study only compared behaviour and neural activity in female participants interacting with other females. While this avoided the risk of confounding by cross-gender behavioural differences, it remains to be determined whether the pattern of behaviour and neural activity would be similar in male participants. In particular, the non-monetary costs associated with adverse outcomes may differ with individual social preferences which might be influenced by gender. Additionally, inter-gender social interactions may be associated with differences in the neural processes associated with social learning and subsequent behaviour.

Our findings provide evidence for a network of brain regions, involved in reward processing and decision making, associated with reinforcement and adaptation of behaviour during social interactions. We observed distinct effects sensitive to expectation violation and others apparently specific to particular outcomes of social encounters, and show modulation of these effects over time. These results suggest that the social brain reinforces and adapts neural activity to drive cooperative behaviour with trustworthy partners.

CONFLICT OF INTEREST

None declared.

REFERENCES

Axelrod, R.M. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Barta, Z., McNamara, J.M., Huszar, D.B., Taborsky, M. (2011). Cooperation among non-relatives evolves by state-dependent generalized reciprocity. *Proceedings of the Royal Society B: Biology*, *278*(1707), 843–8.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, *58*(4), 639–50.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–9.

Boyd, R., Richerson, P.J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, *132*(3), 337–56.

Chang, L.J., Sanfey, A.G. (2009). Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining. *Frontiers Behavioral Neuroscience*, *3*, 36.

Chang, L.J., Doll, B.B., van't Wout, M., Frank, M.J., Sanfey, A.G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology*, *61*(2), 87–105.

Critchley, H.D., Mathias, C.J., Dolan, R.J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Journal of Neuroscience*, *20*(8), 3033–40.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*(6), 3072–7.

Delgado, M.R., Frank, R.H., Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–8.

De Quervain, D.J., Fischbacher, U., Treyer, V., et al. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–8.

Elliott, R., Friston, K.J., Dolan, R.J. (2000). Dissociable neural responses in human reward systems. *Journal of Neuroscience*, *20*(16), 6159–65.

Fehr, E., Camerer, C.F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Science*, *11*(10), 419–27.

Fernandez, G., Brewer, J.B., Zhao, Z., Glover, G.H., Gabrieli, J.D. (1999). Level of sustained entorhinal activity at study correlates with subsequent cued-recall performance: a functional magnetic resonance imaging study with high acquisition rate. *Hippocampus*, *9*(1), 35–44.

Ghahremani, D.G., Monterosso, J., Jentsch, J.D., Bilder, R.M., Poldrack, R.A. (2010). Neural components underlying behavioural flexibility in human reversal learning. *Cerebral Cortex*, *20*(8), 1843–52.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–79.

Hay, D.F. (2009). The roots and branches of human altruism. *British Journal of Psychology*, *100*(3), 473–9.

Herberg, L.J., Stephens, D.N., Franklin, K.B. (1976). Catecholamines and self-stimulation: evidence suggesting a reinforcing role for noradrenaline and a motivating role for dopamine. *Pharmacology Biochemistry and Behavior*, *4*(5), 575–82.

Kampe, K.K., Frith, C.D., Dolan, R.J., Frith, U. (2001). Reward value of attractiveness and gaze. *Nature*, *413*(6856), 589.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83.

Krueger, F., McCabe, K., Moll, J., et al. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(50), 20084–9.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11832–5.

Mealey, L., Daood, C., Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology*, *17*(2), 119–28.

O'Doherty, J., Dayan, P., Friston, K., Critchley, H., Dolan, R.J. (2003a). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–37.

O'Doherty, J., Critchley, H.D., Deichmann, R., Dolan, R.J. (2003b). Dissociating valence of outcome from behavioural control in human orbital and ventral prefrontal cortices. *Journal of Neuroscience*, *23*(21), 7931–9 .

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–4.

Otten, L.J., Henson, R.N., Rugg, M.D. (2001). Depth of processing effects on neural correlates of memory encoding: relationship between findings from across- and within-task comparisons. *Brain*, *124*(2), 399–412.

Phan, K.L., Sripada, C.S., Angstadt, M., McCabe, K. (2010). Reputation for reciprocity engages the brain reward centre. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 13099–104.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rilling, J.K., King-Casas, B., Sanfey, A.G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*(2), 159–65.

Robinson, O.J., Frank, M.J., Sahakian, B.J., Cools, R. (2010). Dissociable responses to punishment in distinct striatal regions during reversal learning. *Neuroimage*, *51*(4), 1459–67 .

Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9.

Seymour, B., Daw, N., Dayan, P., Singer, T., Dolan, R.J. (2007). Differential encoding of gains and losses in the human striatum. *Journal of Neuroscience*, *27*(18), 4826–31.

Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., Frith, C.D. (2004). Brain responses to the acquired moral status of faces. *Neuron*, *41*(4), 653–62.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466–9.

Smith, Y., Kieval, J.Z. (2000). Anatomy of the dopamine system in the basal ganglia. *Trends in Neuroscience*, *23*(10 Suppl.), S28–33.

Stern, E.R., Gonzalez, R., Welsh, R.C., Taylor, S.F. (2010). Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. *Journal of Neuroscience*, *30*(23), 8032–41.

Takezawa, M., Price, M.E. (2010). Revisiting 'The evolution of reciprocity in sizable groups': continuous reciprocity in the repeated n-person prisoner's dilemma. *Journal of Theoretical Biology*, *264*(2), 188–96.

Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W. (2007). Reward value coding distinct from risk attitude- related uncertainty coding in human reward systems. *Journal of Neurophysiology*, *97*(2), 1621–32.

Tomlin, D., Kayali, M.A., King-Casas, B., et al. (2006). Agent-specific responses in the cingulate cortex during economic exchanges. *Science*, *312*(5776), 1047–50.

van't Wout, M., Sanfey, A.G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision making. *Cognition*, *108*(3), 796–803.

Wallis, J.D., Kennerley, S.W. (2010). Heterogeneous reward signals in prefrontal cortex. *Current Opinion in Neurobiology*, *20*(2), 191–8.

Weiskopf, N., Hutton, C., Josephs, O., Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3T and 1.5T. *Neuroimage*, *33*(2), 493–504.

Wittmann, B.C., Schott, B.H., Guderian, S., Frey, J.U., Heinze, H.J., Duzel, E. (2005). Reward-related fMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long term memory formation. *Neuron*, *45*(3), 459–67.

Yoshida, W., Seymour, B., Friston, K.J., Dolan, R.J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience*, *30*(32), 10744–51.