ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 22 (2004) 1694–1703

The neural correlates of theory of mind within interpersonal interactions

James K. Rilling,[a,*,1] Alan G. Sanfey,[a,b,1] Jessica A. Aronson,[a,c] Leigh E. Nystrom,[a,c] and Jonathan D. Cohen[a,c,d]

[a] *Center for the Study of Brain, Mind and Behavior, Princeton University, Princeton, NJ 08544, USA*
[b] *Center for Health and Well-Being, Princeton University, Princeton, NJ 08544, USA*
[c] *Department of Psychology, Princeton University, Princeton, NJ 08544, USA*
[d] *Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213, USA*

Received 18 December 2003; revised 2 April 2004; accepted 2 April 2004

Tasks that engage a theory of mind seem to activate a consistent set of brain areas. In this study, we sought to determine whether two different interactive tasks, both of which involve receiving consequential feedback from social partners that can be used to infer intent, similarly engaged the putative theory of mind neural network. Participants were scanned using fMRI as they played the Ultimatum Game (UG) and the Prisoner's Dilemma Game (PDG) with both alleged human and computer partners who were outside the scanner. We observed a remarkable degree of overlap in brain areas that activated to partner decisions in the two games, including commonly observed theory of mind areas, as well as several brain areas that have not been reported previously and may relate to immersion of participants in real social interactions that have personally meaningful consequences. Although computer partners elicited activation in some of the same areas activated by human partners, most of these activations were stronger for human partners.

Keywords: Theory of mind; PDG; Computer partners

Introduction

One of the distinctive attributes of human social cognition is our propensity to build models of other minds: to make inferences about the mental states of others. This ability has become widely known as theory of mind (Premack and Woodruff, 1978). Several neuroimaging studies have attempted to elucidate the neural substrates that support this distinctively human ability. At least four separate studies have asked participants to make inferences about the mental states of characters in stories or cartoons (Brunet et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; Vogeley et al., 2001). Others have asked participants to infer mental states from expressions in photographs (Baron-Cohen et al., 1999), or to attribute mental states to animations of geometric shapes (Castelli et al., 2000). Collectively, these studies have implicated a consistent network of brain areas in theory of mind, including anterior paracingulate cortex, the posterior superior temporal sulcus at the temporo-parietal junction, and the temporal pole (Gallagher and Frith, 2003). Two more recent studies have probed the neural correlates of theory of mind in participants who are actually immersed in a social interaction with a partner who is outside the scanner. Both reported activation in anterior paracingulate cortex (Gallagher et al., 2002; McCabe et al., 2001), but not posterior superior temporal sulcus or temporal pole.

As part of an effort to explore the ''social brain'' with interactive games, we scanned a group of participants using fMRI as they played two different social games, the Ultimatum (UG) and Prisoner's Dilemma (PDG) games, each of which assesses cooperative intent in a different way. In the Ultimatum Game, two players are asked to split a sum of money. One player proposes how the sum should be divided and the other player either accepts or rejects the offer. If the offer is accepted, it is divided as proposed. However, if it is rejected, both players receive nothing. Although the game theoretic prediction is for the proposer to offer the smallest possible amount based on the assumption that any monetary amount will be accepted by a rational responder, cooperative proposers will offer an even split. In the present study, scanned participants were always in the role of responder in the UG game. In contrast to the Ultimatum Game, the Prisoner's Dilemma Game confronts each of two players with the same decision: cooperate or defect. Each player is awarded a sum of money that depends upon the interaction of their two choices, such that individual earnings are maximized by defection but collective earnings are maximized by cooperation. In both tasks, participants witness a decision by their partner, an offer in the UG game or a choice in the PDG game, which reveals something about the partner's intentions. Are they generous or greedy in the UG game? Are they cooperative or selfish in the PDG game? To the extent

[*] Corresponding author. Department of Anthropology, Emory University, Geosciences Building Room 207, 1557 Pierce Drive, Atlanta, GA 30322.

E-mail address: jrillin@emory.edu (J.K. Rilling).

[1] These two authors contributed equally to this manuscript; authorship was determined by an arbitrary selection process.

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.04.015

that this feedback provokes inferences about the partners' intentions, it is expected to engage theory of mind neural systems.

Our goals in this study were threefold. First, to determine whether learning about the intentions of others activates the putative theory of mind neural network outlined above. Second, to assess the generality and reproducibility of these findings by comparing the neural correlates of two different tasks conducted within the same session that both involve learning about the mental states of others. Third, to utilize a task with high face validity in which participants are engaged in actual social interactions and compare our results with those of other studies in which participants were not similarly engaged. As noted above, this is not the first theory of mind imaging study to immerse participants in genuine social interactions. In both the Gallagher et al. (2002) and McCabe et al. (2001) studies, participants received feedback from human partners in interactive games. However, McCabe et al. did not focus their analysis on epochs in which partner decisions are revealed, and the Gallagher et al. (2002) PET study lacked the temporal resolution to distinguish epochs involving feedback from others. Here, we use fMRI to specifically focus on the blood oxygen level dependent (BOLD) response to receiving feedback from a partner that reveals something about the partner's intent.

Materials and methods

We describe here aspects of the methods relevant to the current interests. Detailed methods pertaining to the Ultimatum Game are reported in a separate article (Sanfey et al., 2003), in which we focused on the BOLD response to receiving fair vs. unfair offers in the UG game, in preparation for responding to the offer. Here, we instead examine the main effect of receiving feedback from a human partner vs. a computer, irrespective of the valence of that feedback (positive or negative), in both the UG and PDG games.

Participants

Participants were 11 females and 8 males recruited from the Princeton University campus, with a mean age of 21.8 years (SD = 7.8 years).

The Ultimatum Game

The Ultimatum Game, as described above, is a two-player game in which one player, the responder, learns whether the other player, the proposer, is generous or greedy. In our version of the game, participants split a US$10 sum. We used the single-shot version of the game in which responders receive a single offer from each proposer.

The Prisoner's Dilemma Game

The Prisoner's Dilemma Game (PDG) is a two-player game in which both players learn whether their partner is cooperative or selfish. We used a single-shot version of the Prisoner's Dilemma Game to examine reciprocated and unreciprocated altruism (Axelrod, 1984; Rapoport and Chammah, 1965). In this game, two players choose to either cooperate with each other or not, and each is awarded a sum of money that depends upon the interaction of both players' choices. There are four possible outcomes: both player A and player B cooperate (CC), player A cooperates and player B defects (CD), player A defects and player B cooperates (DC), or both player A and player B defect (DD). The payoffs for the outcomes are arranged such that DC > CC > DD > CD, and CC > (CD + DC)/2. Each cell of the payoff matrix (Fig. 1a) corresponds to a different outcome of a social interaction. DC represents the situation where player A opts for noncooperation and player B cooperates so that player A benefits at player B's expense. CD is the converse. CC involves mutual cooperation and DD involves mutual noncooperation. In the version of the game we use here (i.e., the single-shot version), the game is played a single time with each partner. Participants were informed that they would choose first in each game and that their partner would witness that choice before making their own decision. Although the decision to reciprocate cooperation is irrational in this version of the game (Nash, 1950), people are sometimes altruistic and cooperation is not uncommon in the single-shot PDG (Sally, 1995). For example, in one recent sequential, single-shot PDG study, second-movers reciprocated cooperation by first-movers at a rate of 42% (Clark and Sefton, 2001).

Behavioral procedures

Before being scanned, each participant completed a tutorial that explained the rules of the two games. After the tutorial, each participant met a group of 10 partners (confederates). Investigators first introduced the participant to the partners by stating his/her name. Partners were informed that they would each play one round of the Ultimatum Game (UG) and one round of the Prisoner's Dilemma (PDG) game with the participant. Each partner then stated his/her name for the participant. The participant was told that the partners could not confer with each other, so that each trial was independent. Afterwards, the participant was escorted to the scanner room and digital photographs were taken of the partners for use in the experiment. Following acquisition of anatomical scans, functional images were acquired as participants played the two games with what they believed were the partners they had met previously. For each trial with a putative human partner, the participant was shown a different photograph of 1 of the 10 confederates. In actuality, for both games, putative partners' choices were actually generated and administered by a computer algorithm. The Ultimatum Game was played first. PDG trials began after all UG trials had finished. Timelines for UG and PDG trials are shown in Figs. 1b and c.

For the UG game, each participant completed 30 rounds in all, 10 playing the game with a putative human partner (once with each of the 10 partners), 10 with an avowed computer partner, and a further 10 control rounds in which they simply received money for a button press (for simplicity, "putative human partners" will henceforth be referred to simply as human partners, and "avowed computer partners" as computer partners). The rounds were presented randomly, and all involved splitting US$10. Each round began with a 12-s preparation interval. For trials with human partners, the participant then saw the photograph and name of their partner for that trial for 6 s. Otherwise, they saw a picture of a computer or a roulette wheel (on control trials). Next, participants saw the offer proposed by the partner for a further 6 s, following which they indicated whether they accepted or rejected the offer by pressing one of two buttons on a button box. Partner choices were determined in advance by a computer algorithm so that all participants saw the same set of offers in the Ultimatum Game. Half of the 10 offers the participants saw were fair, that is, a

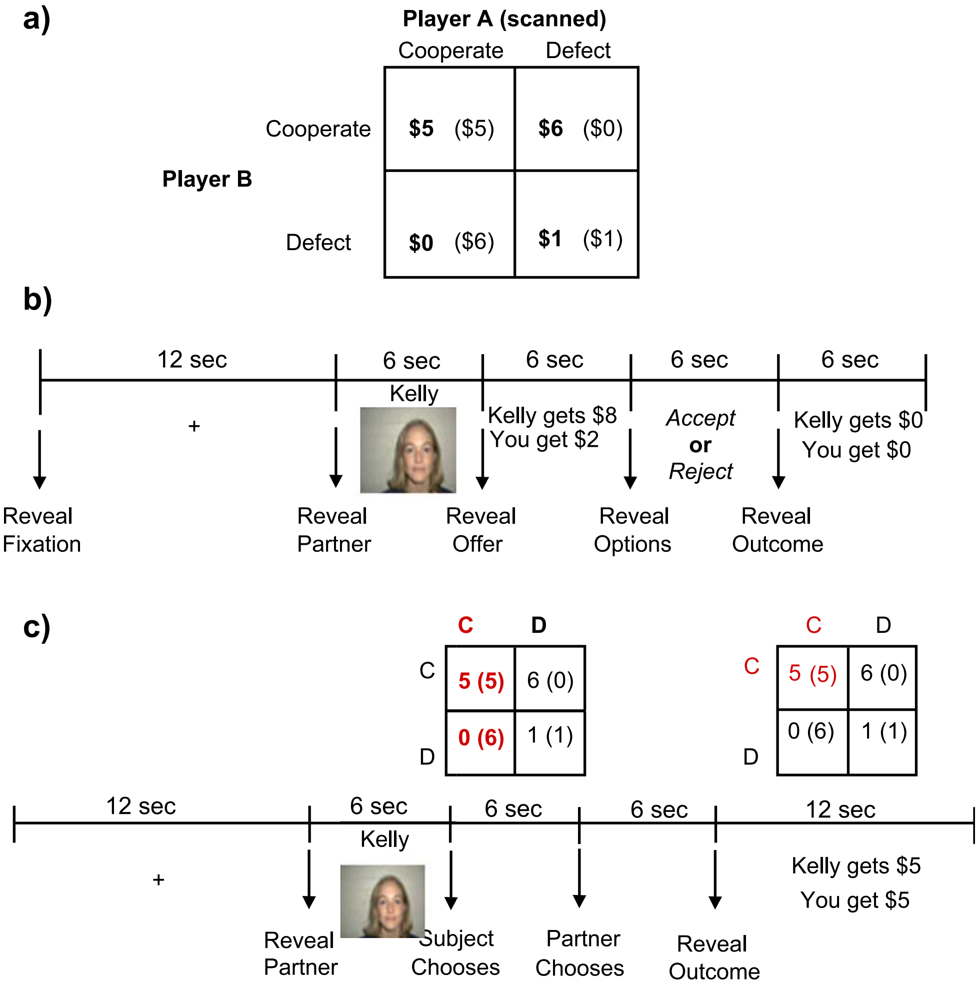

Fig. 1. Illustration of Prisoner's Dilemma and Ultimatum Game tasks. (a) Payoff matrix for Prisoner's Dilemma Game. Scanned participant's choices (player A) are listed atop columns and non-scanned participant's choices (player B) are listed aside rows. Dollar amounts in bold are awarded to player A. Amounts in parentheses are awarded to player B; (b) timeline for a single UG trial. Each trial lasted 36 s; (c) timeline for a single PDG trial. Each trial lasted 42 s.

proposal to split the US\$10 evenly (US\$5:US\$5), with the remaining offers proposing unequal splits (two offers of US\$9:US\$1, two offers of US\$8:US\$2, and one offer of US\$7:US\$3). The 10 offers from the computer partner were identical to those from the human partners (half fair, half unfair). However, participants were told that computer offers were randomly generated. The remaining 10 trials were designed to control for the response to monetary reinforcement, independent of the social interaction. Participants were offered a sum of money and simply had to press a button to accept the money. The amounts were equal to the proposals in the game conditions (i.e., 5 × US\$5, 1 × US\$3, 2 × US\$2, and 2 × US\$1). The distribution of offers generally mimics the range of offers made in the typical uncontrolled version of the game, in which actual human partners make proposals.

For the PDG game, each trial began with a 12-s preparation interval. For trials involving human partners, participants were then shown a photograph of one of the confederates for that trial for 6 s. During the next 6-s epoch, participants chose to cooperate or defect by pressing one of two buttons on a button box. Because we were interested in reciprocated vs. unreciprocated cooperation,

Fig. 2. Activations related to receiving feedback on trials with putative human partners. Colored maps of the t statistic for the contrast between trials with human partners and control trials for (a) the Ultimatum Game and (b) the Prisoner's Dilemma Game. Maps are thresholded at $P < 0.001$. APC = anterior paracingulate cortex; STS = posterior superior temporal sulcus; PCC = posterior cingulate cortex and precuneus.

Fig. 3. Activation in anterior paracingulate cortex with human playing partners. For both PDG and UG games, statistical maps were calculated for the contrast between witnessing a human partner's decisions and control outcomes. Panel a shows the intersection (overlap) of PDG and UG maps in anterior paracingulate cortex; (b) event-related plot of Blood Oxygen Level-Dependent (BOLD) activation for the anterior paracingulate cortex ROI in the PDG game. The partner's choice to cooperate or defect is revealed at time = 0; (c) event-related plot for the anterior paracingulate cortex in the UG game. The partner's offer is revealed at time = 0. For both plots, the y-axis indicates the percentage of signal change relative to a baseline, computed as the average response during the preceding 18 s. Note that the BOLD response exhibits a characteristic delay, reaching its maximum several seconds after stimulus onset.

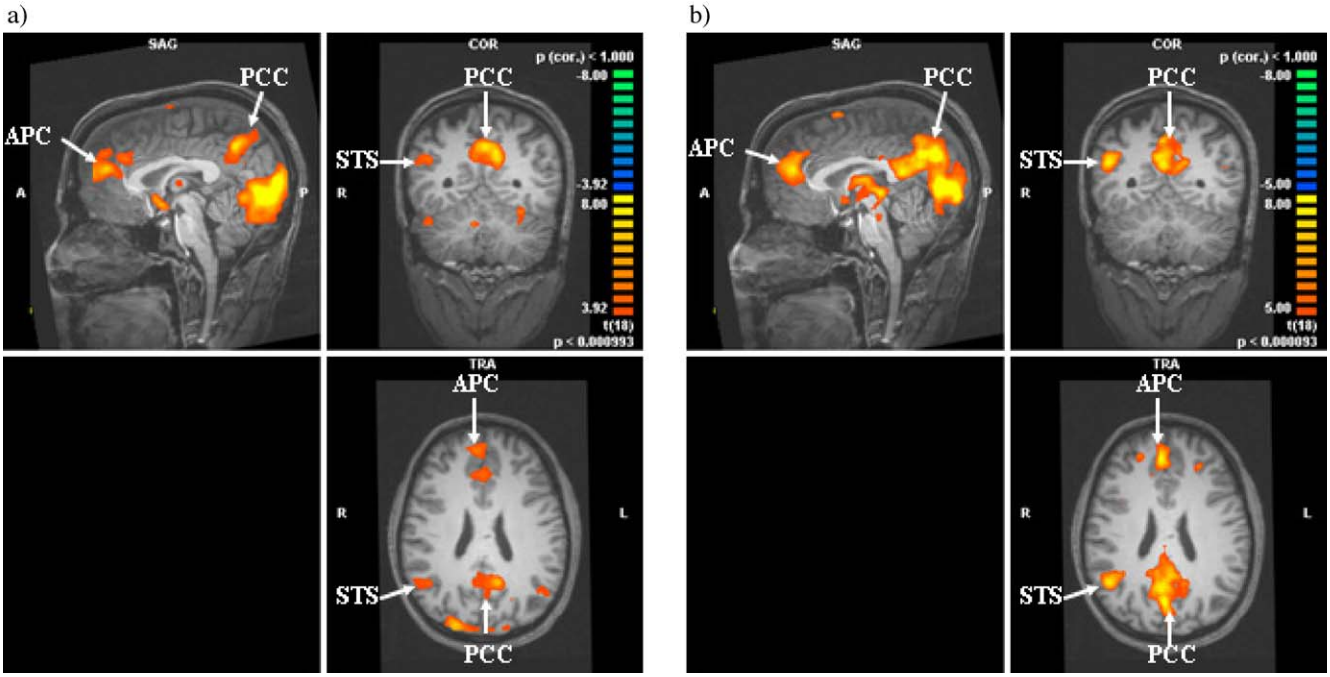
a)
SAG
COR
APC
PCC
STS
A
P
R
p (cor.) < 1.000
-8.00
-3.92
8.00
3.92
t(18)
p < 0.000993
TRA
APC
R
L
STS
PCC
b)
SAG
COR
APC
PCC
STS
A
P
R
p (cor.) < 1.000
-8.00
-5.00
8.00
5.00
t(18)
p < 0.000093
TRA
APC
R
L
STS
PCC

Fig. 2.

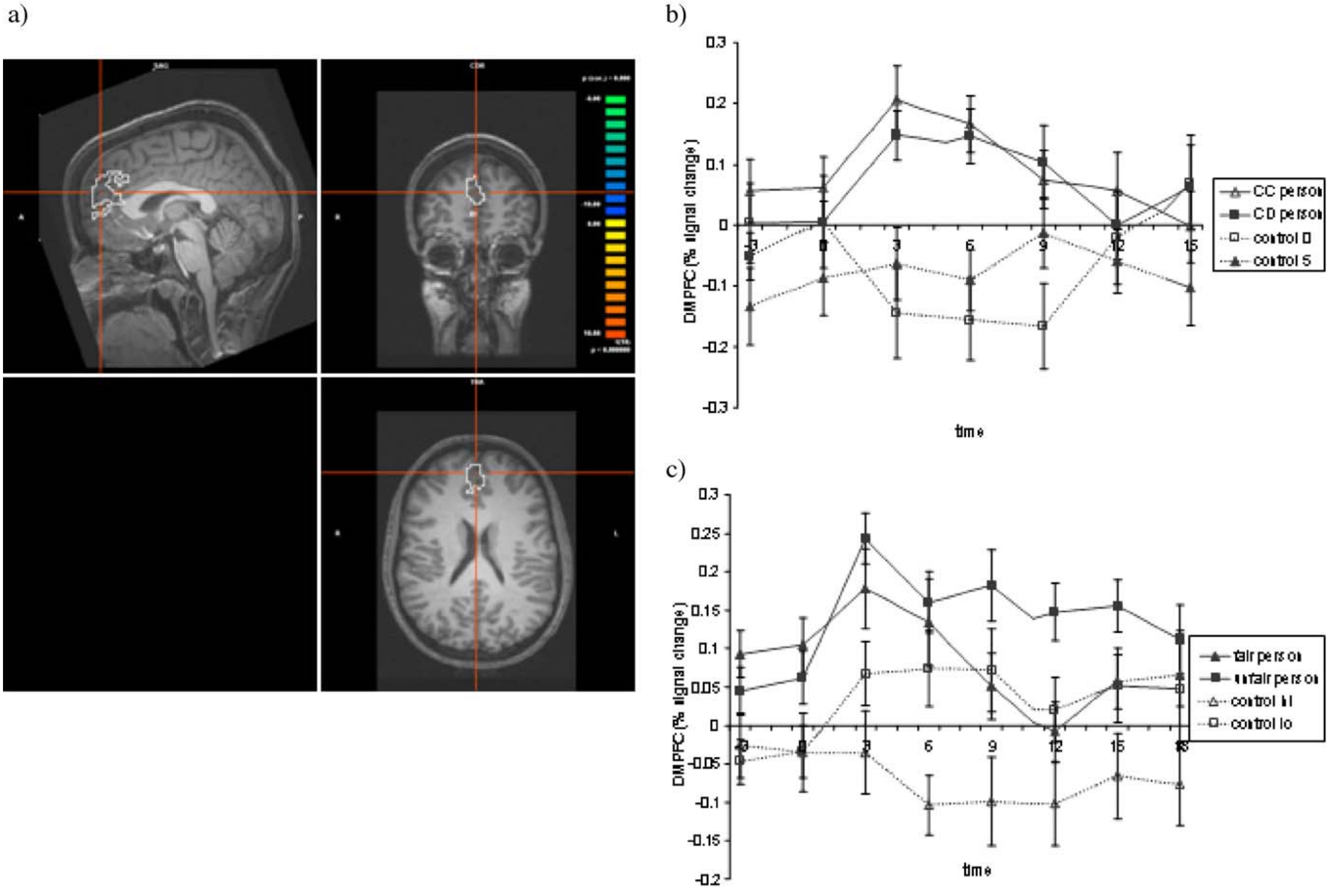
a)
SAG
COR
TRA
b)
DMPFC (% signal change)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-3
0
3
6
9
12
15
time
CC person
CD person
control 0
control 5
c)
DMPFC (% signal change)
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-3
0
3
6
9
12
15
18
time
fair person
unfair person
control hi
control lo

Fig. 3.

our payoff matrix was designed to bias participants toward choosing cooperation over defection. We did so by making the payoff for CC many times larger than that for DD (Rapoport and Chammah, 1965). During the subsequent 6-s epoch, participants were told that their partner was considering their choice and would respond by cooperating or defecting. As with the UG game, partner choices were actually determined by a computer algorithm. The partner's choice and the trial outcome were then displayed for 12 s. Each participant completed 28 PDG trials while in the scanner. Ten were with human partners, as described above. For half of these trials, cooperation was reciprocated and for the other half it was not. This approximates the 42% rate of reciprocation to cooperation observed in second movers in a sequential PDG (Clark and Sefton, 2001). For all trials with human partners, defection by the participant was reciprocated. Another 10 trials were with avowed computer partners. For these trials, participants saw a picture of a computer rather than a partner photograph. They were told only that the computer would make choices according to programmed probabilities. Half of all computer partners reciprocated cooperation and, in contrast to human partners, only half of all computer partners reciprocated defection. That is, participants could realize DC (US$3) outcomes with computer partners. This was deliberately included so that participants could perceive a difference between human and computer partners' strategies. The remaining eight trials were designed to control for the response to monetary reinforcement, independent of the social interaction. In these trials, participants saw a picture of a roulette wheel rather than a partner's photograph. They were then presented with one of two payoff matrices that were matched to those for CC (US$5) and CD (US$0) outcomes. The first had US$0 in all four squares. The second had US$5 in both squares of one column and US$0 in both squares of the other. Participants were simply asked to select either column whereby they would receive the payoff for that column. Participants chose one of the two US$0 columns in the first matrix and overwhelmingly chose the US$5 column in the second (i.e., on 99% of trials). The final 12 s of the control trial indicated how much money the participant had earned (either US$5 or US$0).

Image acquisition

Anatomical scans were acquired with a Siemens 3.0-T head-dedicated MRI scanner using a T1-weighted MP-RAGE protocol (256 × 256 matrix, FOV = 256 mm, 128 1.33 mm sagittal slices). Functional images were acquired using T2*-weighted EPI (TR = 3000 ms, TE = 22 ms, 64 × 64 matrix, FOV = 192 mm, thirty 2.5-mm axial slices with 1.5-mm gap). Four functional runs [2 of 185 scans (UG) and 2 of 200 scans (PDG)] were collected.

Image analysis

Data were preprocessed and analyzed using BrainVoyager software (Maastricht, The Netherlands). Image preprocessing included: six-parameter, 3D motion correction, slice scan time correction using linear interpolation, spatial smoothing with a 6-mm FWHM Gaussian kernel, voxel-wise linear detrending, and high pass filtering of frequencies below three cycles per functional run. Spatial normalization was performed using the standard nine-parameter landmark method of Talairach and Tournoux (1988). A separate general linear model (GLM) was defined for each participant (Friston et al., 1995) in both games that compared the neural response to human, computer, and control trials. In the UG game, we analyzed the epoch when the offer was revealed, and in the PDG game, we analyzed the epoch when the partner's choice to cooperate or defect was revealed. Thus, in both games, we modeled epochs when the participant receives information about the partner that can be used to infer social motivation or intent. For the UG game, six regressors were defined: fair offers from human partners, unfair offers from human partners, fair offers from computer partners, unfair offers from computer partners, high control offers (US$5), and low control offers (<US$5). Six regressors were also defined for the PDG game: CC outcomes with human partners, CD outcomes with human partners, CC outcomes with computer partners, CD outcomes with computer partners, control trials in which participants earn US$0, control trials in which participants earn US$5. Each regressor was convolved with a standard gamma model of the hemodynamic impulse-response function. The resulting general linear model was corrected for temporal autocorrelation using a first-order autoregressive model. For each participant, contrasts were calculated at every voxel in the brain between regression coefficients of interest. A one-sample t test was then used to determine where the average contrast value for the group as a whole (n = 19 participants) differed significantly from zero (a random-effects analysis). The resulting map of the t statistic was thresholded to display only those voxels where the t statistic reached a P value less than 0.001, except for comparisons involving computer partners in the PDG game, where $P < 0.005$ was used. This less stringent threshold was adopted because five participants defected on every computer trial and therefore lacked the necessary outcomes (CC and CD) for comparisons with humans (CC and CD) and control trials (US$5 and US$0). Hence, our statistical power was reduced for computer trials in the PDG game. In considering the imaging results below, we focus on results that replicated across the two games.

Results

Behavior

In the UG game, participants accepted all fair offers, with decreasing acceptance rates as the offers became less fair (Sanfey et al., 2003). In summary, rejection rates for 7:3, 8:2, and 9:1 offers from human partners were 5%, 47%, and 61%, respectively. Rejection rates for 7:3, 8:2, and 9:1 offers from computer partners were 5%, 16%, and 34%, respectively. Unfair offers of US$2 and US$1 on human trials were rejected at a significantly higher rate than those offers on computer trials (US$9:US$1 offer: χ^2 = 5.28, 1 *df*, P = 0.02; US$8:US$2 offer: χ^2 = 8.77, 1 *df*, P = 0.003). In the PDG game, participants cooperated on 81% of trials with human partners and on 66% of computer trials (Fig. 2; χ^2 = 11.3, 1 *df*, P < 0.001). Thus, in terms of behavior, participants distinguished between human and computer partners in both games. Lower rates of cooperation with computer partners in the PDG game may have resulted from participants learning that DC (US$3) outcomes were attainable in this condition.

Imaging

Several brain areas were activated (P < 0.001) in both games (UG and PDG) in response to witnessing a human partner's

Table 1
Areas activated by human partners in both games

Brain region	Talairach coordinates	# voxels	UG	PDG	UG	PDG	UG	PDG
			P-CT	P-CT	CP-CT	CP-CT	P-CP	P-CP
Right mid STS/MTG (BA 21)	56 −10 −13	6	6.69*	8.17*	0.28	−0.02	6.25*	6.92*
Right posterior STS/STG (BA 39)	48 −55 27	34	7.35*	9.42*	0.43	3.95*	6.75*	4.02*
Lingual gyrus (BA 18)	−1 −80 0	1383	13.04*	10.61*	12.03*	7.03*	0.88	1.65
Hypothalamus/midbrain/thalamus	8 −11 −4	75	7.41*	7.71*	1.02	1.55	6.03*	4.71*
Right superior frontal gyrus (BA 8)	20 27 45	4	5.63*	5.78*	1.37	2.98	4.14*	2.19
DMPFC/rostral ACC (BA 9, 32)	3 44 20	243	8.38*	10.00*	2.46	4.40*	5.75*	4.21*
Post cingulate/precuneus (BA 7, 31)	−1 −58 34	271	8.83*	10.73*	2.8	4.55*	5.86*	4.52*
Thalamus	0 −11 8	27	6.75*	8.37*	2.24	4.81*	4.38*	2.32
L hippocampus	−16 −26 −2	45	6.8*	6.35*	1.72	1.94	4.94*	3.63*
L putamen	−22 15 −3	17	5.42*	7.22*	1.61	2.32	3.71*	2.13

These regions are the intersection of (1) activations in response to offers from human partners in the UG game, and (2) activations in response to a human partner's decision to cooperate or defect in the PDG game. The numbers in the cells are *t* statistics for tests that compare the mean contrast value across all participants with zero (*df* = 18). Voxel dimensions are 3 × 3 × 3 mm.
P = person; CT = control; CP = computer.
* $P < 0.001$.

decision compared to receiving an equivalent monetary offer or payoff in the control condition (Table 1). The reproducibility of these activations is illustrated in Fig. 2. Fig. 3 shows the location and time course of activation in one of these areas which spans both dorsomedial prefrontal cortex (DMPFC) and rostral anterior cingulate cortex (i.e., anterior paracingulate cortex, BA 9/32).

To determine whether the activations in Table 1 and Fig. 2 related specifically to interacting with another person or whether they could also be elicited by a computer partner, we tested for a difference between computer and control trials in each of the regions of interest (ROI) from Table 1. In the Ultimatum Game, computer partners were largely ineffective at activating these regions. Only 1 of the 10 regions in Table 1 was activated at $P < 0.001$ for the contrast between computer and control trials. This was a large occipital lobe activation, centered on the left lingual gyrus. On the other hand, several regions from Table 1 were activated by computer partners in the Prisoner's Dilemma Game, including classic theory of mind areas such as anterior paracingulate cortex and right posterior STS, as well as posterior cingulate cortex/precuneus, thalamus, and left lingual gyrus.

We also tested for a difference between human and computer trials in the ROIs from Table 1. This revealed three areas that were more significantly activated by human partners than computer partners, and did not activate at all for the computer-control contrast in both games: (1) the right mid STS, (2) a region spanning the hypothalamus, ventral thalamus, and midbrain, and (3) a region centered on the hippocampus. Additionally, some areas that were significantly activated by computer partners (vs. control) in the PDG game showed even stronger activation to human partners in the PDG game and were also activated by human partners in the UG game: right posterior STS, anterior paracingulate cortex, and posterior cingulate/precuneus.

Table 2
Areas activated by computer partners in the Prisoner's Dilemma Game

Brain region	Talairach coordinates	# voxels
R inferior parietal lobule (BA 40)	43 −46 43	15
R middle frontal gyrus (BA 10)	40 42 19	69
R postcentral gyrus (BA 2)	28 −36 61	92
Right middle frontal gyrus (BA 8)	32 38 38	14
R precentral gyrus (BA 6)	11 −21 68	38
L inferior parietal lobe (BA 40)	−41 −40 43	18

The areas listed showed greater activation for computers compared with people at $P < 0.005$ (*df* = 18). Voxel dimensions are 3 × 3 × 3 mm.

Table 3
Areas activated by human faces in the PDG game

Brain region	Talairach coordinates	# voxels	P-CT	CP-CT	P-CP
R DLPFC (BA 46)	47 23 24	68	4.58*	−1.36	6.64*
R STG (BA 22)	52 −59 15	10	4.27*	0.87	3.60*
R fusiform gyrus (BA 37)	40 −58 −14	302	9.65*	−1.62	11.93*
R precentral gyrus (BA 6)	41 5 35	38	3.93*	−0.08	4.24*
R STS (BA 39)	40 −55 32	40	3.38	0.98	2.55
R inferior frontal gyrus (BA 47)	30 23 −8	20	5.42*	−0.59	6.36*
R superior frontal gyrus (BA 6)	31 −4 59	10	5.79*	2.56	3.43*
R posterior cingulate (BA 31)	2 −59 28	176	6.23*	0.80	6.89*
R frontal pole (BA 10)	8 50 −8	28	5.90*	1.01	6.74*
R caudate	11 1 12	13	3.78*	−0.44	4.46*
L cerebellum	−16 −65 −23	11	3.13	−1.09	4.46*
L fusiform gyrus (BA 37)	−38 −49 −17	71	6.42*	−1.75	8.65*
Mid frontal gyrus (BA 8)	−43 9 37	23	3.42*	1.39	2.15

The areas listed showed greater activation for human faces compared with the picture of the roulette wheel in the control trials. The numbers in the cells are *t* statistics for tests that compare the mean contrast value across all participants with zero (*df* =18). Voxel dimensions are 3 × 3 × 3 mm.
P = person; CT = control; CP = computer.
* $P < 0.001$.

We were also interested in areas that activated more to computer than human partners. Although there were no such areas for the UG game, there were several such areas for the PDG game (Table 2), including right dorsolateral prefrontal cortex (DLPFC) and right parietal lobe.

In addition to engaging theory of mind processes when the game outcomes were revealed, we considered the possibility that participants might attempt to infer intentionality upon viewing their partner’s photograph earlier in PDG trials, when they were making their decision. Indeed, during this decision epoch, both the right posterior STS and the posterior cingulate were activated by photographs of human partners more so than by photographs of a computer or roulette wheel in the computer and control trials, respectively (Table 3). Event-related plots for both of these ROIs reveal an increase in activation in response to the partner’s face that remains elevated until the game outcome is revealed, at which time there is a secondary increase that rapidly resolves (Fig. 4). In contrast to the later outcome epoch, during the earlier decision epoch, activation was not observed in anterior paracingulate cortex, mid STS, or the regions including the hypothalamus and hippocampus.

Discussion

Our objective in this study was to determine whether inferring the intentions of others activated the putative theory of mind neural network and whether activated areas would replicate across our two games. Indeed, for both games, we detected activation in two of the three classic TOM areas: anterior paracingulate cortex and posterior STS.

Both of these areas responded to decisions from both human and computer partners, but showed stronger responses to human partners in both games. The stronger response to human partners is consistent with the behavioral data showing that participants distinguished between human and computer partners, rejecting unfair offers from human partners more frequently in the UG and cooperating more often with human partners

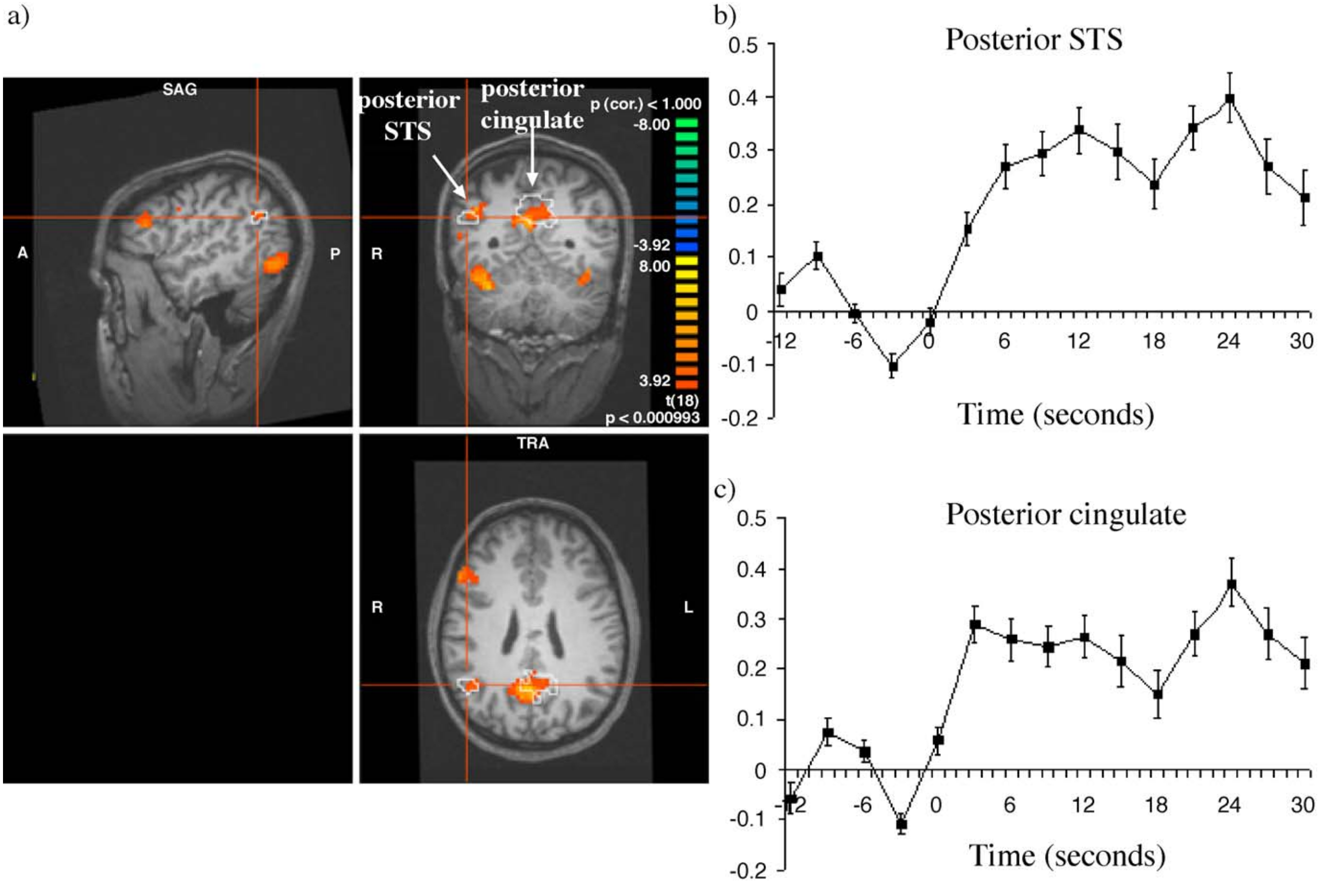

Fig. 4. Activation in posterior STS and posterior cingulate in the PDG game in response to seeing a human partner’s face. (a) Colored regions were more active when viewing human partners’ faces compared to the roulette wheel in the control condition (the early epoch; 0–6 s on timeline in b and c). Regions outlined in white are those that showed a main effect of receiving feedback from human partners compared with control trials in both games (the later epoch; 18–24 s on timeline in b and c); (b) event-related plot showing the time course of Blood Oxygen Level-Dependent (BOLD) activation in the posterior STS ROI (colored region in a), when playing with human partners. The partner’s face appears at $t = 0$ s and the game outcome is revealed at $t = 18$ s; (c) event-related plot showing the time course of BOLD activation in the posterior cingulate ROI (colored region in a) when playing with human partners. The partner’s face appears at $t = 0$ s and the game outcome is revealed at $t = 18$ s. The baseline for both plots is the epoch from −12 to 0 s on the timeline. The BOLD response exhibits a characteristic delay, reaching its maximum several seconds after stimulus onset.

in the PDG game. The fact that computer partners are able to activate this network, albeit to a lesser extent than human partners, suggests that either this neural system can also be activated by reasoning about the unobservable states of nonhuman systems, or that participants imbue their computer partners with human attributes.

Beyond areas that we found in common with earlier studies, we also observed activation in several areas that have not previously been reported for theory of mind tasks. As already mentioned, most of the earlier studies have not involved actual social interactions with real outcomes; therefore, these previously unreported areas may represent brain regions that are uniquely recruited when participants are inferring the intentions of real social partners with whom they are directly interacting and whose behavior has consequences for their material well-being. The areas include: (1) posterior cingulate/precuneus, (2) mid STS, (3) an activation spanning hypothalamus, midbrain, and thalamus, and (4) an activation centered on the left hippocampus.

Activation in posterior cingulate has consistently been associated with emotionally salient stimuli (Maddock, 1999). This activation and the one including the hypothalamus may relate to emotional arousal in response to receiving feedback from human partners, that is diminished (posterior cingulate) or absent (hypothalamus) when interacting with computer partners (see Table 1). Mid STS has been implicated in biographical memory retrieval (Gorno-Tempini et al., 1998; Haxby et al., 2000; Leveroni et al., 2000; Mitchell et al., 2002), a process that is unlikely to be engaged by our task. However, participants are learning new information about others. Thus, it is possible that mid STS is involved in biographical memory encoding, in addition to retrieval. To further assess the reproducibility of this activation, we used the earlier data of Rilling et al. (2002) to conduct a new analysis that examined the main effect of receiving feedback from a real human partner in the iterated PDG. Indeed, the contrast (CChuman + CDhuman) − (CCcomputer + CDcomputer) activated mid STS (Fig. 5). The involvement of episodic memory processes is also suggested by the hippocampal activation, which replicated across tasks (Squire and Zola, 1996). Like mid STS, hippocampal activation could relate to encoding specific subject's game behavior and intentions. Are they fair or unfair, cooperative or selfish?

Why were these activations not observed in earlier studies that immersed participants in real social interactions (Gallagher et al., 2002; McCabe et al., 2001)? The study by Gallagher et al. (2002) is a PET study and may lack the temporal resolution to capture the effects we document here, and McCabe et al. (2001) focused their analysis on decision-making epochs that did not involve receiving feedback from a partner. In contrast to these two studies, our analysis focused specifically on the BOLD response to receiving feedback from a partner that reveals something about the partner's social motivation. Differences between studies may also relate to details of the experimental set-up, such as amount of interaction with partners before scanning, formality of introduction procedures, tone of experimenters, and so on. In our games, participants met a group of 10 human partners to whom they were formally introduced. Post-scan debriefings reveal that the interactions felt real to the participants. Furthermore, the type of games utilized in the present study are quite effective at arousing strong feelings of unfairness and untrustworthiness, which may elicit neural responses over and above those found by other games such as 'stone/paper/scissors', where partner decisions are less personal and perhaps less emotionally arousing. We suspect that differences in these factors contribute to discrepancies between the pattern of activation found here and those reported by Gallagher et al. (2002) and McCabe et al. (2001). Finally, differences in statistical power may also contribute, as our study included 19 participants, compared with $n = 9$ for Gallagher et al. (2002) and $n = 12$ for McCabe et al. (2001).

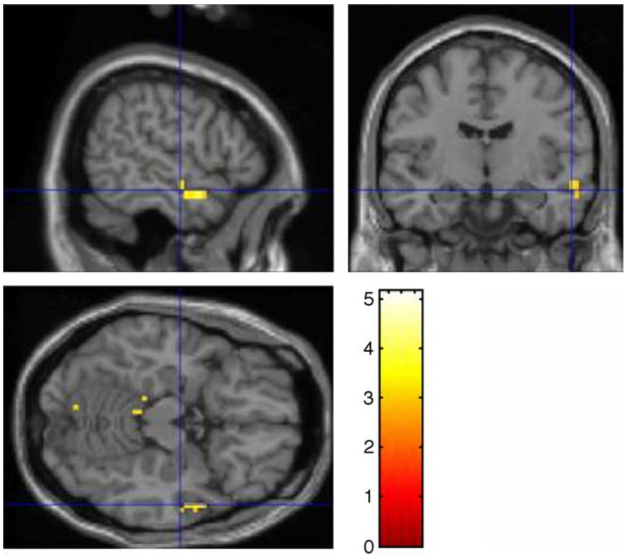

Fig. 5. Activation in right mid STS with human partners in the iterated PDG (data from Rilling et al., 2002). Areas in yellow were more active in response to CC and CD outcomes with human partners compared to the same outcomes with computer partners (at $P < 0.005$ uncorrected). The cross-hairs are centered on MNI coordinate 57, −10, −16, the equivalent of the center Talairach coordinate of our ROI in mid superior temporal sulcus (56, −10, −13). Each iterated game consisted of 20 consecutive rounds with the same playing partner (human or computer). Images are presented in radiological orientation.

As mentioned above, classic theory of mind regions like anterior paracingulate cortex and posterior STS exhibited stronger activation in response to feedback from human vs. computer partners. This raises the possibility that these regions are specialized for making inferences about other human minds. However, other explanations are possible. First, participants may simply be more cognitively engaged on trials with human partners and this may explain the greater involvement of theory of mind processes. Alternatively, differences in activation strength could also relate to differences between human and computer trials. For each trial with a human partner, participants are shown a photograph of a different person. On the other hand, for computer trials, participants see the same photograph of the same computer on every trial. Thus, participants may perceive the computer trials as an iterated interaction with a single entity, as opposed to a series of single-shot interactions, as for the human trials. If so, we might reasonably assume that participants would be more motivated to identify the computer strategy since they believe that they will continue to interact with it, and therefore participants would actually have a stronger representation of the computer's intentions than those of the 10 human partners. Despite this, the areas of interest were activated more for human than computer partners. Another difference between human and computer trials in the PDG is that DC (US$3) outcomes could only be realized on computer trials.

However, there were no differences in the behavior of human and computer partners in the UG game, and the anterior paracingulate and posterior STS activations were also observed in the UG game. Although it remains possible that these two conditions felt different because computer trials were perceived as iterated and human trials as single-shot, we note that Gallagher et al. (2002) also reported stronger activation in the anterior paracingulate region for interactions with human compared with computer partners, when both played an identical random strategy. Thus, this area may be specialized for making inferences about other human minds.

There was a notable discrepancy between the UG and PDG games in the ability of computer partners to elicit activation in areas also activated by human partners. Computer trials were very effective at eliciting activation in the PDG game but largely ineffective in the UG game. This difference may relate to the varying responsiveness of the two computer strategies. In the UG game, the computer did not respond to participant choices; it simply made the participant an offer. But in the PDG game, the computer responded to a choice by the participant and gave the impression that the computer's decision was contingent on that of the participant. Our data suggest that this element of responsiveness may be critical for computers to elicit patterns of neural activity analogous to those elicited by people. These interactive, responsive strategies may feel more human-like, and participants may be more likely to attribute human motives and attributes to them.

Finally, we note that computer partners activated right DLPFC and right parietal lobe more than human partners in the PDG game. DLPFC has traditionally been regarded as a cognitive control area (e.g., Miller and Cohen, 2001), so this activation may reflect the allocation of attentional resources to attempt to elucidate the computer strategy and the optimal response to it, especially since, as noted above, participants may have perceived that they were engaged iteratively with a single computer partner. In contrast, participants might have processed human partners' decisions in a more affective manner, without consideration of what the partner is likely to do next, especially if these were accurately perceived as single-shot interactions.

In sum, we used the PDG and UG games to determine what brain areas are engaged when participants receive consequential feedback from human partners in real-life social interactions. Consistent with earlier theory of mind neuroimaging studies that did not immerse participants in genuine social interactions, we detected activation in both games in anterior paracingulate cortex and posterior STS. Although these same areas were activated by computer partners in the PDG game, activations were significantly stronger for human partners. Additionally, we observed activation in several areas that have not been reported previously, but that may likely be engaged in everyday, real-world social interactions. These include the mid STS and hippocampus which may be involved in encoding biographical memories, as well as the posterior cingulate and an area including the hypothalamus that may relate to emotions aroused by interactions with human partners. Our data also show that people are more likely to engage theory of mind brain areas with computer partners when the latter are perceived to be responsive to their human partner's choices. In conclusion, we believe that the use of more ecologically realistic tasks like these are a useful complement to simpler, more rigorously controlled paradigms such as those presently in the literature to more fully characterize the neural basis of human theory of mind abilities.

Acknowledgments

This work was supported by a grant from the Mind Brain Body and Health Initiative, and by the Center for the Study of Brain, Mind, and Behavior and Princeton University.

References

Axelrod, R.M., 1984. The Evolution of Cooperation. Basic Books, New York.

Baron-Cohen, S., Ring, H.A., Wheelwright, S., Bullmore, E.T., Brammer, M.J., Simmons, A., Williams, S.C.R., 1999. Social intelligence in the normal and autistic brain: an fMRI study. Eur. J. Neurosci. 11, 1891–1898.

Brunet, E., Sarfati, Y., Hardy-Bayle, M.-C., Decety, J., 2000. A PET investigation of the attribution of intentions with a nonverbal task. NeuroImage 11, 157–166.

Castelli, F., Happe, F., Frith, U., Frith, C., 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. NeuroImage 12, 314–325.

Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: evidence on reciprocation. Econ. J. 111, 51–68.

Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S.J., Frith, C.D., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57, 109–128.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J., Turner, R., 1995. Characterizing dynamic brain responses with fMRI: a multivariate approach. NeuroImage 2, 166–172.

Gallagher, H.L., Frith, C.D., 2003. Functional Imaging of "theory of mind". Trends Cogn. Sci. 7, 77–83.

Gallagher, H., Happe, F., Brunswick, N., Fletcher, P., Frith, U., Frith, C., 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. Neuropsychologia 38, 11–21.

Gallagher, H., Jack, A., Roepstorff, A., Frith, C.D., 2002. Imaging the intentional stance in a competitive game. NeuroImage 16, 814–821.

Gorno-Tempini, M.L., Price, C.J., Josephs, O., Vandenberghe, R., Cappa, S.F., Kapur, N., Frackowiak, R.S., Tempini, M.L., 1998. The neural systems sustaining face and proper-name processing. Brain 121, 2103–2118.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face processing. Trends Cogn. Sci. 4, 223–233.

Leveroni, C., Seidenberg, M., Mayer, A., Mead, L., Binder, J., Rao, S., 2000. Neural systems underlying the recognition of familiar and newly learned faces. J. Neurosci. 20, 878–886.

Maddock, R.J., 1999. The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain (comment). Trends Neurosci. 22, 310–316.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T., 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Natl. Acad. Sci. 98, 11832–11835.

Miller, E., Cohen, J., 2001. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 24, 167–202.

Mitchell, J.P., Heatherton, T.F., Macrae, C.N., 2002. Distinct neural systems subserve person and object knowledge. Proc. Natl. Acad. Sci. 99, 15238–15243.

Nash, J.F., 1950. Equilibrium points in n-person games. Proc. Natl. Acad. Sci. U. S. A. 36, 48–49.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1, 515–526.

Rapoport, A., Chammah, A.M., 1965. Prisoner's Dilemma; A Study in Conflict and Cooperation. University of Michigan Press, Ann Arbor.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D., 2002. A neural basis for social cooperation. Neuron 35, 395–405.

Sally, D.F., 1995. Conversation and cooperation in social dilemmas. Ration. Soc. 7, 58–92.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 300, 1755–1758.

Squire, L.R., Zola, S.M., 1996. Structure and function of declarative and nondeclarative memory systems. Proc. Natl. Acad. Sci. U. S. A. 93, 13515–13522.

Talairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publishers, New York.

Vogeley, K., Bussfeld, P., Newen, A., Hermann, S., Happe, F., Falkai, P., Maier, W., Shah, N., Fink, G., Zilles, K., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. NeuroImage 14, 170–181.