doi:10.1093/scan/nsq042 SCAN (2011) 6, 338–347

Neural basis of conditional cooperation

Shinsuke Suzuki,[1] Kazuhisa Niki,[2] Syoken Fujisaki,[3] and Eizo Akiyama[3]

[1]Laboratory for Integrated Theoretical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, [2]Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, and [3]Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8573, Japan

Cooperation among genetically unrelated individuals is a fundamental aspect of society, but it has been a longstanding puzzle in biological and social sciences. Recently, theoretical studies in biology and economics showed that conditional cooperation—cooperating only with those who have exhibited cooperative behavior—can spread over a society. Furthermore, experimental studies in psychology demonstrated that people are actually conditional cooperators. In this study, we used functional magnetic resonance imaging to investigate the neural system underlying conditional cooperation by scanning participants during interaction with cooperative, neutral and non-cooperative opponents in prisoner's dilemma games. The results showed that: (i) participants cooperated more frequently with both cooperative and neutral opponents than with non-cooperative opponents; and (ii) a brain area related to cognitive inhibition of pre-potent responses (right dorsolateral prefrontal cortex) showed greater activation, especially when participants confronted non-cooperative opponents. Consequently, we suggest that cognitive inhibition of the motivation to cooperate with non-cooperators drives the conditional behavior.

Keywords: neuroeconomics; functional MRI; prisoner's dilemma game; cooperation; reciprocity

INTRODUCTION

Cooperation is crucial for our survival in a society. In fact, we can find forms of cooperation, such as food-sharing and management of common-pool resources, all over the world (Kaplanand and Hill, 1985; Ostrom *et al.*, 1999).

On the other hand, cooperation among genetically unrelated individuals has been a longstanding puzzle in biology and economics. This is because cooperation is often costly to those who practice it, while others benefit. In recent decades, using the prisoner's dilemma (PD) game (Figure 1A), many theoretical studies showed that a conditional cooperation, cooperating with cooperators but not with non-cooperators, can spread over a society (Axelrod and Hamilton, 1985; Fudenberg and Maskin, 1986; Nowak and Sigmund, 1998). These works have been quite influential because no kinship or purely altruistic preference is needed to explain the evolution of cooperation. Moreover, many behavioral experiments confirmed that humans and some kinds of animal are actually conditional cooperators (Milinski, 1987; Wedekind and Milinski, 2000; Fischbacher *et al.*, 2001; Fehr and Fischbacher, 2003; Bshary and Grutter, 2006).

Especially, Axelrod and Hamilton (1981) investigated the evolution of cooperation in *fixed matching repeated PD,* where each participant plays the game with the same opponent repeatedly. They demonstrated that a type of conditional cooperators who practice 'tit-for-tat' behavior, which means that they cooperate on the first move and thereafter copy their opponents' moves, can evolve and form a cooperative society against defectors. This is because conditional cooperators cooperate with each other and never cooperate with defectors, hence conditional cooperators prevail in the population by natural selection. On the other hand, Nowak and Sigmund (1998) formalized *random matching repeated PD*, where each participant plays the game with a different opponent in every round. The opponent's past actions are shown to the participant as an *action history* (index of reputation). Those authors showed that conditional cooperators called 'Discriminators', who cooperate only with others who have a good action history, can establish a cooperative society. The former and latter forms of cooperation are respectively called *direct reciprocity* and *indirect reciprocity* because the cooperation is reciprocated by the recipient *directly* (by someone else *indirectly*) in fixed matching repeated PD (in random matching repeated PD). Subsequent to those works, various studies of reciprocal cooperation were conducted (May, 1987; Boyd and Richerson, 1988; Suziki and Akiyama, 2005, 2007, 2008). Furthermore, many experimental reports described that humans actually behave as conditional cooperators in broadly various situations (Wedekind and Milinski, 2000; Fischbacher *et al.*, 2001; Fehr and Fischbacher, 2003).

Despite the importance of conditional cooperation, the neural systems underlying it are not well understood. This is because it is difficult to investigate the issue using

Received 30 April 2009; Accepted 19 April 2010
Advance Access publication 25 May 2010

We thank M. Sakaki, H. Nakahara and the anonymous reviewers for their helpful comments. We also thank S.H. Oda and H. Kimura for their cooperation in conducting the experiment. Grants-in-Aid for Scientific Research, Nos 17103002, 18700220 and 20240026 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Correspondence should be addressed to Shinsuke Suzuki, Laboratory of Integrated Theoretical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.
E-mail: shinsuke@adds.co.jp.

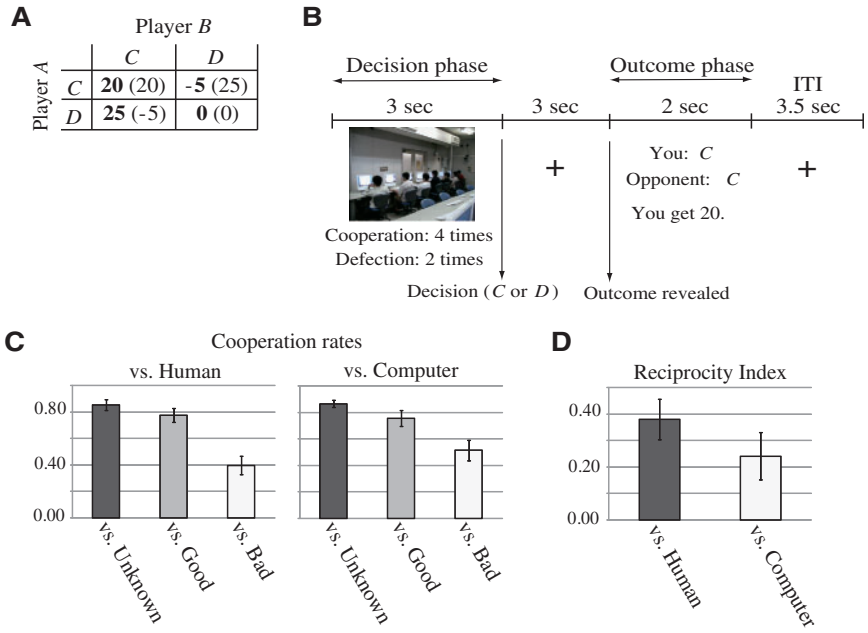

Fig. 1 (**A**) Payoff matrix for the prisoner's dilemma game (PD) used in this study. Side rows represent player A's choices (C or D), and top columns denote player B's choices. Numbers in bold are payoffs for player A, whereas those in parentheses are payoffs for player B. This game represents social interactions in which each player is better off by defecting irrespective of the opponent's choice. However, if both players cooperate, each can obtain a higher payoff than that obtained through mutual defection. (**B**) Timeline for a round of PD. Each round began with a 3 s *decision phase*. In this phase, participants saw photographs of the potential opponents and an action history of the actual opponent (no. of cooperation, no. of defection). Note that participants knew the action history of the actual opponent, but did not know who the actual opponent was among the potential opponents. Around the end of the decision phase, participants decided either to cooperate or to defect by pressing one of two buttons. After a 3-s fixation interval, they saw the outcome of the game, called the *outcome phase*. (**C**) Behavioral results: these are the cooperation rates when participants interacted with human and computer opponents whose action history was unknown, good, or bad, respectively (error bars denote s.e.m.). (**D**) Behavioral results: these are reciprocity indices in human and computer conditions. This index is defined as the difference between the rate of cooperation with good opponents and that with bad opponents.

fixed matching repeated PD. In fixed matching repeated PD, participants can make a decision about whether or not to cooperate in the subsequent game when they observe the outcome of the current game, because the opponent in the subsequent game is the same as in the current game. For example, if your opponent does not cooperate, you can immediately decide not to cooperate with her in the subsequent game in order to retaliate. Therefore, the neural activity used in decision making about the subsequent game cannot be dissociated from the activity in observing the outcome of the current game. To our knowledge, previous functional magnetic resonance imaging (fMRI) studies of repeated PD (Rilling *et al.*, 2002, 2007, 2008) had this problem, although they succeeded in reporting several interesting findings [e.g. mutual cooperation is associated with activation in reward-processing brain regions (Rilling *et al.*, 2002)].

To overcome this problem, in the present study we used random matching repeated PD with action histories reflecting participants' past actions, i.e. the indirect reciprocity paradigm (Nowak and Sigmund, 1998; Wedekind and Milinski, 2000), to investigate the neural and psychological processes underlying conditional cooperation. In random matching repeated PD, you cannot immediately decide on your behavior in the subsequent game based on the opponent's current action because the opponent in the subsequent game is different in most cases. So, we can dissociate the neural activity in decision making from the activity in observing the outcome of the game in random matching repeated PD.

In order to implement conditional cooperation, we need to reward cooperative opponents and punish non-cooperative ones. It was shown that activation of the right dorsolateral pre-frontal cortex (DLPFC) is associated with punishing norm violators in social situations (Knoch *et al.*, 2006; Buckholtz *et al.*, 2008). On the basis of the previous results, we hypothesized that, in our experimental task, the right DLPFC played important roles in participants' conditional cooperation, and the right DLPFC exhibited greater activation, especially when faced with non-cooperative opponents.

MATERIALS AND METHODS

Participants

In our experiment, 17 healthy volunteers participated (Japanese students at the University of Tsukuba; 12 men and 5 women; mean age = 21.35, s.d. = 1.46). Participants were excluded if they had medical, neurological or psychiatric illness. All participants gave their informed consent. The study was approved by the MRI Ethics Committee of

the National Institute of Advanced Industrial Science and Technology.

Experimental procedure

The participants engaged in six sessions, each consisting of 18 rounds of the PD game (Figure 1A) with action histories reflecting the participants' past actions, in the MRI scanner, after two practice sessions outside the scanner. The payoff matrix for PD used in this study was consistent with the low-cost condition of Bolton *et al.* (2005), in which there is relatively little motivation to defect. Although it is expected that participants' behaviors are sensitive to payoff structures such as the cost and benefit of cooperation, conditional cooperation was observed within a wide range of payoff parameters in random matching repeated PD with action histories (Wedekind and Milinski, 2000, Milinski *et al.*, 2001; Bolton *et al.*, 2005).

In all six sessions, they played random matching repeated PD with computer program opponents. However, they were instructed as follows. In three of the six sessions, called the *human condition*, they played PD with a human opponent chosen randomly from a pool of seven potential human opponents in every round. In the remainder, called the *computer condition*, they played the game with a randomly chosen computer opponent. The order of the six sessions they played was randomized across subjects.

To reinforce the belief that participants would be interacting with real human opponents in the human condition, before the practice sessions we showed the participants a movie of a room in which the seven potential human opponents were sitting in front of a computer screen. Although the participants were told that they were watching live movie, in fact they were watching a pre-recorded movie. Furthermore, in the practice and the scanned human sessions, the participants saw a photograph of the room (Figure 1B). After the experiment, we debriefed the participants and confirmed that they did not doubt the existence of the human opponents. Moreover, as we show later, behavioral and neuro-imaging data imply that participants can distinguish human opponents from computer opponents.

Each round in a session began with a 3 s *decision phase*. In this phase, participants saw an *action history* of the actual opponent and a photograph of the potential opponents in the human condition set (in the computer condition set, the action history of the computer opponent and a photograph of a computer were presented to participants). Around the end of the phase, the participants decided either to cooperate or to defect based on the action history by pressing one of two buttons (average reaction time was 2.78 s from the beginning of a decision phase). During the decision phase, participants knew the action history of the actual opponent, but did not know who among the potential opponents was the actual opponent. Furthermore, participants were instructed that their opponents also made decisions in this phase (i.e. simultaneous PD). Next, after a 3 s fixation interval, they were informed of the result of the game for 2 s during the *outcome phase*. The intertrial interval (ITI) was 3.5 s. Regarding the length of ITI, there is a trade-off. In terms of gathering a clear BOLD signal related to the task, it is ideal to employ a longer ITI (>12 s). On the other hand, it is possible to use an experimental design with a short ITI in order to ensure a sufficient number of trials. In this study, we adopted a short ITI (3.5 s) experimental design. The timeline for a PD round is portrayed in Figure 1B.

An action history was described as the set of the numbers of instances of cooperation and defection in past rounds of the session. However, in 6 of the 18 rounds, an action history was hidden. In this case, the history was displayed as '$*$'.

For analyses, we defined an opponent's history '$*$' as 'unknown', a history where the number of instances of cooperation is greater than that of defection as 'good', and the reverse history as 'bad'. Each of the three types of histories were presented six times per session (18 rounds) in random order.

Note that each participant was told that, in each round, others not chosen as the opponent were playing PD with each other. Furthermore, each participant was informed that his or her action history was shown to the opponent; and that his or her own and all the others' action histories were updated based on their own actions.

Moreover, in both human and computer conditions, decision-making by computer opponents was programmed to be more likely to cooperate with a participant who had cooperated many times in prior rounds of the session. It was shown that real human subjects have this tendency (Wedekind and Milinski, 2000). That is, an opponent's action depends on the participant's own action history, but not on the opponent's own reputation. Specifically, if a participant had cooperated c times and defected d times in the past rounds of a session, the probability that an opponent of the participant cooperated was

$$P = 0.1 + 0.8 \times \frac{c}{c+d}. \quad (1)$$

In the first round, an opponent cooperated with a $P = 0.5$.

Furthermore, after the experiment, participants' payoffs were converted into cash (Japanese yen). They received 0–2500 yen in addition to a fixed reward of 5500 yen ($\$1 \approx 100$ yen).

Rational behavior in this experiment and rationality index

Here, we show the rational behavior in this task in terms of monetary reward maximization. Of course, participants could not know the rational behavior exactly because they did not know their opponents' decision-making algorithm described in Equation (1). However, they might be able to learn the rational behavior through trial and error.

In this task, each of the six sessions was mutually independent because the action histories of the participants were reset at the beginning of a session.

Each session consisted of 18 rounds. Therefore, the actions of a participant are represented as an 18-D vector $\mathbf{a} = (a_1, \ldots, a_{18})$, where $a_t \in \{0, 1\}$. The t-th element of the vector represents an action of the participant at round t: 1 denotes cooperation; 0 denotes defection. Given the opponent's decision-making algorithm described in Equation (1), the expected payoff of a participant at round t is represented as

$$\pi(t) = \begin{cases} a_1(0.5 \times 20 - 0.5 \times 5) + (1 - a_1)(0.5 \times 25) & \text{if } t = 1, \\ a_t\left(\frac{\sum_{j=1}^{t-1} a_j}{t-1} \times 20 - \left(1 - \frac{\sum_{j=1}^{t-1} a_j}{t-1}\right) \times 5\right) & \\ +(1 - a_t)\left(\frac{\sum_{j=1}^{t-1} a_j}{t-1} \times 25\right) & \text{if } t > 1, \end{cases} \quad (2)$$

(see the payoff matrix of PD in Figure 1A). Therefore, the expected payoff throughout the session is

$$\sum_{t=1}^{18} \pi(t) = 12.5 - 5 \sum_{t=1}^{18} a_t + 25 \sum_{t=2}^{18} \frac{\sum_{j=1}^{t-1} a_j}{t-1}. \quad (3)$$

The second term represents the cost of cooperation. The third one denotes the benefit from cooperation. The equation above can be rewritten as

$$\begin{aligned} \sum_{t=1}^{18} \pi(t) = 12.5 &+ a_1\left[-5 + 25\left(1 + \frac{1}{2} + \cdots + \frac{1}{17}\right)\right] \\ &+ a_2\left[-5 + 25\left(\frac{1}{2} + \cdots + \frac{1}{17}\right)\right] \\ &+ \cdots \\ &+ a_{16}\left[-5 + 25\left(\frac{1}{16} + \cdots + \frac{1}{17}\right)\right] \\ &+ a_{17}\left[-5 + \frac{25}{17}\right] + a_{18}(-5). \end{aligned} \quad (4)$$

In the expressions shown above, the coefficient of a_t is positive for $t \leq 14$ and negative for $t > 15$. Therefore, the rational behavior that maximizes the expected payoff of a participant is cooperation during the first 14 rounds and defection in the remaining four rounds. Intuitively speaking, participants should cooperate in the first 14 rounds to obtain a better action history in order to induce opponents' cooperation; but, in the last four rounds, they should defect to exploit their opponents' cooperation.

We define a participant's rationality index as the ratio of rational actions to all of his or her own actions in the human condition.

Data acquisition

All scanning was performed using an MRI scanner (3.0-T MRI, Signa GE 3T; GE Healthcare) equipped with EPI capability using the standard head coil for radiofrequency transmission and signal reception. In all, 27 axial slices (4 mm thick with a 0.2 mm gap, interleaved) were prescribed to cover the whole brain. T2*-weighted gradient echo EPI was used. The imaging parameters were TR = 2 s, TE = 30 ms, FA = 75, FOV = 20 × 20 cm (64 × 64 mesh). To avoid head movement, participants wore a neck brace and were asked not to talk or move during scanning.

Data analysis

The image data were analyzed using SPM5. After slice timing correction, the images were realigned to the first volume to correct for participants' motion, spatially normalized and spatially smoothed using an 8 mm FWHM Gaussian kernel. High-pass temporal filtering (using a filter width of 128 s) was also applied to the data.

A separate general linear model (GLM) was defined for each subject. The model included six regressors that modeled the BOLD response to interacting with opponents of six types in the decision phase: unknown-human, good-human, bad-human, unknown-computer, good-computer and bad-computer; and eight regressors that modeled the BOLD response to combinations of the PD outcomes of four types and opponents of two types (eight combinations in all) in the outcome phase. Each regressor was convolved with a standardized model of the hemodynamic response. For each subject, contrasts of parameter estimates were computed at every voxel of the brain. One-sample t-tests were used to identify voxels, where the average contrasts for the whole group differed significantly from zero (i.e. a random effect analysis).

We included no data on four participants in the analyses of the outcome phase because sufficient mutual defection (DD) or unilateral defection (DC) outcomes were not ensured in their data. On the other hand, for analyses of the decision phase, we included data on all 17 participants.

RESULTS

Behavioral results

Behavioral results showed that participants were conditional cooperators (Figure 1C); that is, more likely to cooperate with opponents whose action history was good or unknown than with opponents whose action history was bad (Tukey's multiple-comparison test: $n = 17$; good-unknown, $P > 0.05$; good–bad, $P < 0.01$; unknown-bad, $P < 0.01$). (From here on, we call opponents with unknown action histories 'unknown opponents'). This result suggests that participants strongly distinguished bad opponents from the others, and so they were more likely to defect when interacting with bad opponents, whereas they were more likely to cooperate with both unknown and good opponents. In addition, the tendency was more robust when interacting with human opponents than when interacting with computer opponents (Figure 1D). Precisely, the difference between the rate of cooperation with good opponents and that with bad opponents, i.e. the *reciprocity index*, was

greater when participants interacted with humans than when they interacted with the computer ($n = 17$, $t_{16} = 1.78$, $P < 0.05$, one-tailed).

Neuro-imaging results

Decision phase

Here, we show the results of neuro-imaging in the decision phase, during which the participants made decisions based on their opponents' action histories. (In the remainder of this article, we present results in the human condition unless we specifically mention otherwise.)

Comparing participants' neural activities when faced with bad opponents to their activities when faced with good opponents, we found significant activity in the right DLPFC (Figure 2 and Table 1). In addition, bilateral posterior superior temporal sulcus and temporo–parietal junction (pSTS/TPJ) showed greater activation associated with interaction with bad opponents (Table 1).

We also analyzed the opposite contrast, (good–bad). However, no region was detected under the same statistical threshold. Under a more liberal threshold ($P < 0.01$, uncorrected), we found activation in the bilateral temporal pole (Supplementary Table S1), which is thought to be a part of the social brain (Frith and Frith, 2003; Gallagher and Frith, 2003).

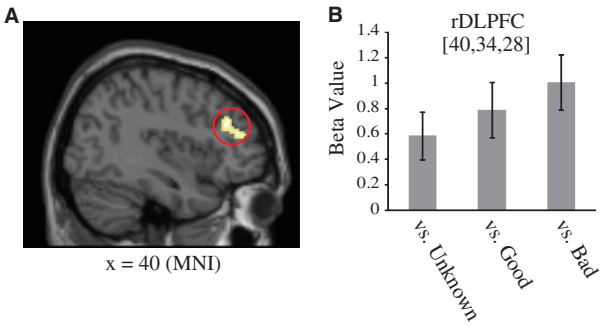

Fig. 2 fMRI results in the decision phase in the human condition: activation related to interaction with bad opponents. (**A**) Maps of the *t*-statistic for the contrast (bad–good) showing activation ($P < 0.001$, uncorrected, cluster size $k > 10$) in the right DLPFC peaked at (40, 34, 28) (MNI). (**B**) Right DLPFC activation for interacting with unknown, good and bad opponents (error bars indicate s.e.m.).

Table 1 fMRI results in the decision phase in the human condition: areas showing activation during interaction with bad opponents compared with good opponents, the contrast bad–good; ($n = 17$, $P < 0.001$, uncorrected, cluster size $k > 10$)

Region	MNI coordinates (*x y z*)	Spatial extent (voxels)	*t*-static
R middle frontal gyrus (BA9)	40 34 28	118	4.94
R superior temporal gyrus (BA22)	56 −58 16	53	4.78
L middle temporal gyrus (BA39)	−54 −64 32	32	4.64
R superior frontal sulcus (BA6)	24 0 46	19	4.30
R superior frontal sulcus (BA6)	28 2 60	18	4.25
L inferior parietal lobule (BA40)	−58 −36 28	11	3.89

Outcome phase

Next, we show neuro-imaging results in the outcome phase, during which participants observed the result of a round of PD.

Several studies showed that brain regions associated with reward-feedback-based learning play important roles in social decision-making (Barraclough *et al.*, 2004; Lee *et al.*, 2004; Delgado *et al.*, 2005; King-Casas *et al.*, 2006; Sanfey, 2007) as well as in non-social decision-making (O'Doherty, 2004; O'Doherty *et al.*, 2004; Doya, 2008; Rushworth and Behrens, 2008).

To detect brain regions involved in reward processing in our task, we compared brain activity when observing mutual cooperation (CC) or unilateral defection (DC) with unilateral cooperation (CD) or mutual defection (DD). (Recall that monetary payoffs for CC and DC were greater than that for CD and DD as described in Figure 1A.) For the contrast [CC + DC − CD − DD], we found activation in the striatum and left orbital medial pre-frontal cortex (OMPFC; Figure 3A–D and Table 2).

Furthermore, we found a significant correlation between each participant's *rationality index* and the activity of OMPFC associated with the contrast (Figure 3E; $r = 0.69$, $P < 0.01$), although no significant correlation was observed in the activity of the striatum. The rationality index of a participant is defined as the ratio of rational actions to all of his or her own actions in the human condition (we exactly show the rational behavior and the definition of the rationality index in 'Materials and Methods' section).

Furthermore, recent neuro-imaging studies suggested that human subjects make decisions based not only on the prospects of monetary reward but also on those of social rewards (Rilling *et al.*, 2002; Moll *et al.*, 2006; Fehr and Camerer, 2007; Harbaugh *et al.*, 2007).

Harbaugh *et al.* (2007) and Moll *et al.* (2006) suggested that donation or cooperation itself is rewarding for humans. So, we analyzed a contrast (CC + CD − DC − DD) to investigate neural activity in observing outcomes where the subject had cooperated with others. For the contrast, we found posterior cingulate cortex (PCC) activity (Figure 4A and B, Table 3). Moreover, each participant's activity was significantly correlated with his or her cooperation rate (Figure 4C; $r = 0.56$, $P < 0.01$).

On the other hand, Rilling *et al.* (2002) suggested that mutual cooperation (CC) is associated with activation in reward-processing brain regions. So, we tried to find brain regions specifically involved in observing CC by a contrast (3*CC − CD − DC − DD). Although some reward-related brain regions were detected under the contrast, we could not show that any of them were specifically engaged in CC (see Supplementary data).

Human condition–computer condition

Finally, we compared brain activity during interaction with human opponents to that during interaction with computer

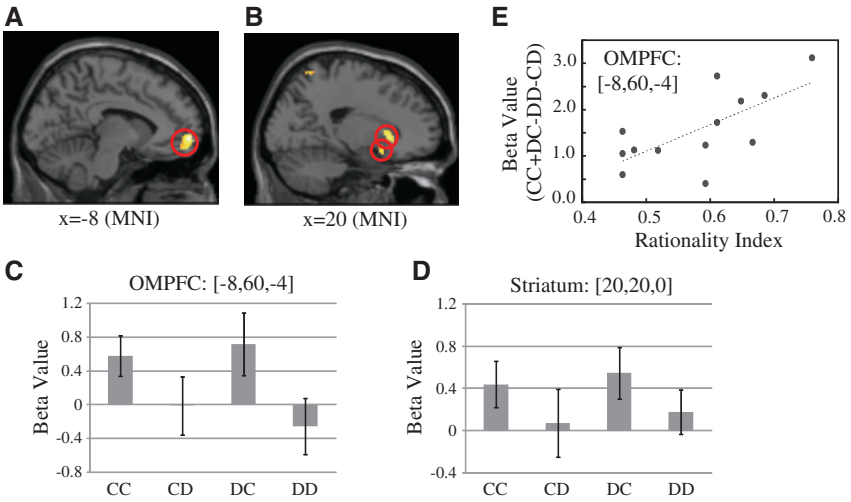

Fig. 3 fMRI results in the outcome phase in the human condition: activation related to a monetary reward. (**A** and **B**) Maps of the *t*-statistic for the contrast (CC + DC − CD − DD) showing activation ($P < 0.001$, uncorrected, cluster size $k > 10$) in (**A**) the OMPFC peaked at (−8, 60, −4) (MNI), and (**B**) the striatum peaked at (20, 20, 0) (MNI). (**C** and **D**) OMPFC and the striatum activation for each of the four outcomes (error bars indicate s.e.m.). (**E**) Plot of the correlation between the β-value of the peak voxel in the OMPFC and the rationality index for each participant ($r = 0.69$, $P < 0.01$). The rationality index of a participant is defined as the ratio of rational actions to all of their own actions in the human condition (we show exactly the rational behavior and the definition of the rationality index in 'Materials and Methods' section).

Table 2 fMRI results in the outcome phase in the human condition: areas showing activation for the contrast [CC+DC−CD−DD] ($n = 13$, $P < 0.001$, uncorrected, cluster size $k > 10$)

Region	MNI coordinates (x y z)			Spatial extent (voxels)	*t*-static
L medial frontal gyrus	−8	60	−4	338	7.22
L superior frontal gyrus (BA6)	−24	14	64	16	6.87
R sublobar/extranuclear	32	−20	−10	19	6.49
R caudate	20	20	0	96	6.03
L frontal lobe/subgyral	−26	36	−2	20	5.66
L sublobar/extranuclear	−30	2	−14	25	5.32
R superior parietal lobule (BA7)	20	−60	68	60	4.85
L fusiform gyrus (BA37)	−56	−52	−20	17	4.80
R putamen	20	12	−16	39	4.78
R superior temporal gyrus (BA22)	60	−4	−8	72	4.73
L superior parietal lobule (BA7)	−20	−54	70	13	4.60
R superior temporal gyrus (BA6)	60	4	2	22	4.45

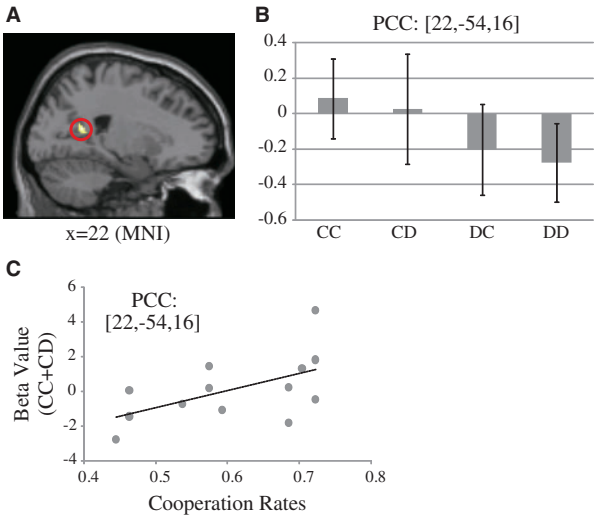

Fig. 4 fMRI results in the outcome phase in the human condition: activation related to social reward derived by cooperating. (**A**) Maps of the *t*-statistic for the contrast (CC + CD − DC − DD) showing activation ($P < 0.001$, uncorrected, cluster size $k > 10$) in the PCC peaked at (22, −54, 16) (MNI). (**B**) PCC activation for each of the four outcomes (error bars indicate s.e.m.). (**C**) Plot of the correlation between the β-value of the peak voxel in PCC and the cooperation rate for each participant ($r = 0.56$, $P < 0.01$).

opponents. Consistent with previous studies (Gallagher *et al.*, 2002; Gallagher and Frith, 2003; Amodio and Frith, 2006, Frith and Frith, 2006), in the decision phase, this contrast revealed activation in anterior cingulate cortex (ACC) thought to be associated with mentalizing and theory of mind (Table 4). Furthermore, the temporal pole and insula, which are also related to mentalizing, theory of mind and empathy (Frith and Frith, 2003; Gallagher and Frith, 2003; Singer *et al.*, 2004; Lissek *et al.*, 2008), were activated in the outcome phase, although with only a few voxels (Table 4). This implies that participants distinguished between human opponents and computer opponents.

DISCUSSION

Cooperation is a costly behavior that benefits others. Why is self-sacrificial behavior observed in human and some kinds of animal societies? How does cooperation evolve? This issue has been of considerable concern in various fields such as biology (Nowak, 2006), economics (Fudenberg and Maskin, 1986) and psychology (Dawes, 1980). In recent

Table 3 fMRI results in the outcome phase in the human condition: areas showing activation for the contrast [CC + DC − CD − DD] ($n = 13$, $P < 0.001$, uncorrected, cluster size $k > 10$)

Region	MNI coordinates (*x y z*)	Spatial extent (voxels)	*t*-static
R posterior cingulate	22 −54 16	33	5.49
R precentral gyrus (BA4)	40 −24 70	41	5.46

Table 4 fMRI results: areas showing differences in activation during interaction with humans compared with computer opponents ($n = 17$, $P < 0.001$, uncorrected)

Region	MNI coordinates (*x y z*)	Spatial extent (voxels)	*t*-static
Decision phase			
Pons	−18 −34 −28	14	6.03
L inferior frontal gyrus	−48 6 16	92	5.48
Thalamus	−8 −8 −4	13	4.43
Midbrain	−6 −38 −18	10	4.39
L inferior frontal gyrus	−44 36 0	12	4.27
Anterior cingulate (BA32)	8 38 20	12	4.25
Outcome phase			
R insula (BA13)	34 −30 18	2	4.74
R middle temporal gyrus (BA21)	48 −2 −40	1	4.24

decades, many theoretical studies showed that a conditional cooperation, cooperating with cooperators but not with non-cooperators, can spread over a society (Axelrod and Hamilton, 1981; Fudenberg and Maskin, 1986; Nowak and Sigmund, 1998). Moreover, behavioral experiments confirmed that humans and some kinds of animals are actually conditional cooperators (Milinski, 1987; Wedekind and Milinski, 2000; Fischbacher *et al*., 2001; Bshary and Grutter, 2006). However, the neural and psychological processes underlying conditional cooperation are still unknown.

In this study, we investigated the neural and psychological bases of conditional cooperation using fMRI. Specifically, we scanned participants while they played random matching repeated PD, i.e. the indirect reciprocity paradigm (Nowak and Sigmund, 1998; Wedekind and Milinski, 2000), with opponents whose action history was good, unknown or bad.

The behavioral results showed that participants were conditional cooperators, which is consistent with the results of previous studies (Wedekind and Milinski, 2000; Fischbacher *et al*., 2001; Fehr and Fischbacher, 2003). In addition, we found that participants strongly distinguished bad opponents and were therefore more likely to defect when faced with bad opponents, whereas they tended to cooperate with both unknown and good opponents.

In the rest of this section, we first discuss neural activity when participants observed a result of the game, and then discuss the activity when they made decisions based on the opponents' action histories.

Neural activity when observing a result of the game

We found that OMPFC and striatum were activated when participants received a large monetary reward. The OMPFC and the striatum are thought to be parts of a neural circuit that guides behavior based on reward feedback (Bechara *et al*., 1999, 2000; O'Doherty, 2004; O'Doherty *et al*., 2004; Doya, 2008; Hare *et al*., 2008; Rushworth and Behrens, 2008), which is consistent with our results.

Furthermore, several studies suggested that OMPFC is more directly involved in action–selection than the striatum (Bechara *et al*., 1997, 1999, 2000; Hampton *et al*., 2008; Hare *et al*., 2008). Specifically, lesion studies showed that patients with OMPFC damage were impaired in gambling tasks (they continued to choose disadvantageously; Bechara *et al*., 1997, 1999, 2000). Moreover, Hare *et al*. (2008) showed that OMPFC encodes the goal value, whereas the striatum encodes the value update signal, the so-called reward prediction error. In addition, Hampton *et al*. (2008) suggested that, in the context of a strategic game, expected values of actions that guide behaviors are maintained in OMPFC and are updated by a prediction error encoded in the striatum. Indeed, our experiment revealed a significant correlation between each participant's *index of rationality*, in terms of monetary reward maximization, and the OMPFC activity associated with receiving a large monetary reward (Figure 2C), although no significant correlation was observed in the activity of the striatum. In other words, participants who acted rationally exhibited greater activation in OMPFC associated with monetary reward. A similar correlation was observed in a non-social gambling task (De Martino *et al*., 2006).

Taking these factors into account, we suggest that, in our experiment, OMPFC was involved in guiding behavior towards the rational based on reward feedback. Specifically, we speculate that the OMPFC activation was associated with the learning signal (i.e. reward feedback) to acquire the rational behavior.

In addition, our analysis revealed that PCC activity was associated with observing outcomes where the participants had cooperated with others. Moreover, participants who showed greater PCC activation were more likely to cooperate with others. PCC is thought to be involved in reward processing (McCoy *et al*., 2003; Fliessbach *et al*., 2007; Ballard and Knutson, 2009) and in the theory of mind or mentalizing (Seger *et al*., 2004; Chiu *et al*., 2007). We therefore suggest that one possible function of PCC activity is to process social reward elicited by cooperating with others.

Another possibility is that PCC activity reflected a reputation building or an expectation of future reward. This is because, in our random matching repeated PD task, a participant's cooperation (i.e. the outcome of mutual

cooperation or unilateral cooperation) caused his or her own action history to improve, meaning that opponents would cooperate with him or her in the future. Consistent with that, in the fixed matching repeated PD task, Rilling *et al.* (2002) showed that mutual cooperation was most likely to be followed by mutual cooperation in the subsequent round and that observing mutual cooperation was associated with neural activity in reward-processing brain regions.

Neural activity during decision making

We found greater neural activity in the right DLPFC when interacting with bad opponents than with good opponents.

The right DLPFC is generally thought to be involved in cognitive top-down inhibition of pre-potent responses (Miller and Cohen, 2001; Mansouri *et al.*, 2007, 2009). Then, in our task, what pre-potent responses did the right DLPFC inhibit? Recall that, in our task, it is speculated that participants made decisions for the sake of monetary and social rewards, as suggested in the discussion about neural activity when observing results of the game. To gain monetary rewards, cooperation is rational behavior in most rounds (see 'Materials and Methods' section). This is because opponents in our experiment were more likely to cooperate with a participant who had a good action history. So, participants should cooperate to obtain better action histories in the long run. Likewise, in terms of social rewards, participants obviously should attempt to cooperate. Moreover, the right DLPFC exhibited greater activation when the participant engaged in interaction with bad opponents. These factors suggest that a pre-potent response in our experiment was cooperating with others, and that the DLPFC activity inhibited the response. In other words, only when faced with bad opponents, participants inhibited the pre-potent motivation of cooperation and so were more likely to defect (i.e. reciprocally fair behavior) in association with the brain activity in the right DLPFC. In fact, we confirmed that the right DLPFC activity was greater when participants defected than cooperated (see details in Supplementary data).

Consistent with that inference, a previous study showed that disruption of the right DLPFC by repetitive transcranial magnetic stimulation (rTMS) diminished rejection of unfair offers in ultimatum games, although the disruption did not affect judgment of the fairness of the offers (Knoch *et al.*, 2006). The authors of the study (Knoch *et al.*, 2006) suggested that the right DLPFC played a key role in overriding self-interested motivation (to accept unfair offers in ultimatum games) and thus enabled participants to implement their fairness goals. Moreover, Rilling *et al.* (2007) reported the activity in DLPFC during defection in fixed matching repeated PD and suggested that pre-potent motivation to cooperate was overridden by cognitive control associated with DLPFC.

Several studies have reported a similar conflict-induced neural activity in ACC as well as DLPFC in social situations (Sanfey *et al.*, 2003; Rilling *et al.*, 2007; Baumgartner *et al.*, 2008). In contrast, in our experiment, ACC did not show greater activation, especially when the participant was interacting with bad opponents. Although it is not easy to explain why, one possibility is that the ACC is less directly related to resolution of conflict between incompatible motor plans than the DLPFC, as was recently suggested (Rushworth *et al.*, 2004; Mansouri *et al.*, 2007, 2009).

In addition to the right DLPFC, the bilateral pSTS/TPJ showed greater activation during interactions with bad opponents. The pSTS/TPJ is known to be associated with mentalizing or the theory of mind (Frith and Frith 2003; Gallagher and Frith, 2003; Overwalle, 2009). The activation might be consistent with our statement that cooperating was a pre-potent response in our experiment. That is, participants needed to deliberately think about their opponents only when the opponents were bad, which elicited the activation of pSTS/TPJ.

Taken together, these findings indicate that, in this task, the brain had two parallel systems. One is a pre-potent system that taught participants to cooperate for the sake of social and long-term monetary rewards. The system was established through trial-and-error learning associated with the activity of reward-processing brain regions such as OMPFC and PCC. The other is a system that inhibited the motivation of cooperate and motivated participants to defect when faced with bad opponents. The right DLPFC, a brain area involved in cognitive inhibition of pre-potent responses, was implicated in the system. This idea is consistent with our behavioral results that participants were more likely to defect on bad opponents and cooperate with both unknown and good opponents.

SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

Conflict of Interest

None declared.

REFERENCES

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*, 268–77.

Axelrod, R., Hamilton, W.D. (1981). The evolution of cooperation. *Science*, *211*, 1390–96.

Ballard, B., Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. *NeuroImage*, *45*, 143–50.

Barraclough, D.J., Conroy, M.L., Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*, 404–10.

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*, 639–50.

Bechara, A., Damasio, H., Tranel, D., Damasio, A.R. (1997). Deciding Advantageously before knowing the advantageous strategy. *Science*, *275*, 1293–5.

Bechara, A., Damasio, H., Damasio, A.R., Lee, G.P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *The Journal of Neuroscience*, *19*, 5473–81.

Bechara, A., Damasio, H., Damasio, A.R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, *10*, 295–307.

Bolton, G.E., Katok, E., Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics*, *89*, 1457–68.

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

Bshary, R., Grutter, A.S. (2006). Image scoring and cooperation in a cleaner fish mutualism. *Nature*, *441*, 975–8.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., et al. (2008). The neural correlates of third-party punishment. *Neuron*, *60*, 930–40.

Chiu, P.H., Amin Kayali, M.A., Kishida, K.T., et al. (2007). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, *57*, 463–73.

Dawes, R.M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–93.

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

De Martino, B.D., Kumaran, D., Seymour, B., Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*, 684–7.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*, 410–6.

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

Fehr, E., Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785–91.

Fischbacher, U., Gachter, S., Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*, 397–404.

Fliessbach, K., Weber, B., Trautner, P., et al. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, *318*, 1305–8.

Frith, C.D., Frith, U. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions Royal Society of London, Series B*, *358*, 459–73.

Frith, C.D., Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*, 531–4.

Fudenberg, D., Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*, 533–54.

Gallagher, H.L., Jack, A.I., Roepstorff, A., Frith, C.D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *16*, 814–21.

Gallagher, H.L., Frith, C.D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, *7*, 77–83.

Harbaugh, W.T., Mayr, U., Burghart, D.R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*, 1622–5.

Hampton, A.N., Bossaerts, P., John, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6741–6.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, *28*, 5623–9.

Kaplanand, H., Hill, K. (1985). Food sharing among Ache foragers. tests of explanatory hypotheses. *Current Anthropology*, *26*, 223–46.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R. (2006). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*, 78–83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*, 829–32.

Lee, D., Conroy, M.L., McGreevy, B.P., Barraclough, D.J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research*, *22*, 45–58.

Lissek, S., Peters, S., Fuchs, N., et al. (2008). Cooperation and deception recruit different subsets of the theory-of-mind network. *PLoS ONE*, *3*, 1–11.

McCoy, A.N, Crowley, J.C., Haghighian, G., Dean, H.L., Platt, M.L. (2003). Saccade reward signals in posterior cingulate cortex. *Neuron*, *40*, 1031–40.

Mansouri, F.A., Buckley, M.J., Tanaka, K. (2007). Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment. *Science*, *318*, 987–90.

Mansouri, F.A., Tanaka, K., Buckley, M.J. (2009). Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. *Nature Reviews Neuroscience*, *10*, 141–52.

May, R.M. (1987). More evolution of cooperation. *Nature*, *327*, 15–7.

Miller, E.K., Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167.

Milinski, M. (1987). TIT FOR TAT in sticklebacks and the evolution of cooperation. *Nature*, *325*, 433–5.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceeding of Royal Society of London*, *268*, 2495–501.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 15623–28.

Nowak, M.A., Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–7.

Nowak, M.A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560–3.

O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, *14*, 769–76.

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*, 452–4.

Ostrom, E., Burger, J., Field, C.B., Norgaard, R.B., Policansky, D. (1999). Revisiting the commons: local lessons, global challenges. *Science*, *284*, 278–82.

Overwalle, F.V. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, *30*, 829–58.

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.

Rilling, J.K., Glenne, A.L., Jairama, M.R., et al. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry*, *61*, 1260–71.

Rilling, J.K., Goldsmitha, D.R., Glenna, A.L., et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia*, *46*, 1256–66.

Rushworth, M.F.S., Walton, M.E., Kennerley, S.W., Bannerman, D.M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*, 410–7.

Rushworth, M.F.S., Behrens, T.E.J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*, 389–97.

Sanfey, A.G., Rilling, J.K., Aronson, J.A. Nystrom, Cohen, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–8.

Seger, C.A., Stoneb, M., Keenan, J.P. (2004). Cortical activations during judgments about the self and another person. *Neuropsychologia*, *42*, 1168–77.

Suzuki, S., Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. *Proceeding of Royal Society of London series B*, *272*, 1373–7.

Suzuki, S., Akiyama, E. (2007). Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. *Journal of Theoretical Biology*, *245*, 539–52.

Suzuki, S., Akiyama, E. (2008). Evolutionary stability of first-order-information indirect reciprocity in sizable groups. *Theoretical Population Biology*, *73*, 426–36.

Sanfey, A.G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, *318*, 598–602.

Singer, T., Seymour, B., O'Doherty, J.P., Kaube, H., Dolan, R.J., Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*, 1157–62.

Wedekind, C., Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–2.