Neuron, Vol. 35, 395–405, July 18, 2002, Copyright ©2002 by Cell Press

A Neural Basis for Social Cooperation

James K. Rilling,[1,2] David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, and Clinton D. Kilts
Department of Psychiatry and Behavioral Sciences
Emory University
Atlanta, Georgia 30322

Summary

Cooperation based on reciprocal altruism has evolved in only a small number of species, yet it constitutes the core behavioral principle of human social life. The iterated Prisoner's Dilemma Game has been used to model this form of cooperation. We used fMRI to scan 36 women as they played an iterated Prisoner's Dilemma Game with another woman to investigate the neurobiological basis of cooperative social behavior. Mutual cooperation was associated with consistent activation in brain areas that have been linked with reward processing: nucleus accumbens, the caudate nucleus, ventromedial frontal/orbitofrontal cortex, and rostral anterior cingulate cortex. We propose that activation of this neural network positively reinforces reciprocal altruism, thereby motivating subjects to resist the temptation to selfishly accept but not reciprocate favors.

Introduction

Evolutionary biologists have long theorized about how altruistic behavior can exist, given that natural selection is based primarily on the differential survival and reproductive success of individual organisms rather than groups of organisms. W.D. Hamilton's kin selection theory nicely accounts for altruism among relatives (Trivers, 1985). But cooperation among nonrelatives is pervasive in human society and must also be explained. Reciprocity, including both direct and indirect reciprocity, has been proposed to account for altruism toward nonrelatives. In direct reciprocity, individuals dispense favors, and these favors are likely to be returned by the recipient, in one form or another, in times of future need (Sahlins, 1972; Trivers, 1971). In indirect reciprocity, the favor is returned by a third party (Nowak and Sigmund, 1998). This study sought to define the neural basis of direct reciprocity.

The paradigmatic example of reciprocal altruism is food sharing, which human beings engage in far more deliberately and pervasively than any other species (Ridley, 1996), and which was almost certainly essential to the survival of our hominid ancestors in their African savannah niche (Lee and DeVore, 1968). This deeply ingrained tendency manifests itself in myriad ways in modern human social life, including the exchange of expertise, information, opportunities, and a host of material resources.

On the other hand, cooperation based on reciprocal altruism is rare in the rest of the animal kingdom and has only been convincingly demonstrated for a handful of species (Axelrod and Hamilton, 1981; Ridley, 1996; Trivers, 1971). In attempting to provide an explanation for its scarcity, evolutionary biologists have theorized that two or more preconditions must be satisfied for reciprocal altruism to evolve in a species: (1) individuals must interact repeatedly with social partners over the course of their lifetime, and (2) individuals must be able to recognize conspecifics and discriminate against those who do not reciprocate altruism (Axelrod and Hamilton, 1981; Trivers, 1971). As an additional precondition, there must also be a mechanism that enables individuals to inhibit the temptation to accept but not reciprocate altruism; a mechanism that weights long-term rewards and punishments over immediate and transient, short-term gains (Frank, 1988). Only with such a mechanism can the long-term benefits of sustained mutual cooperation be realized.

The iterated Prisoner's Dilemma Game has been used by investigators from a wide range of disciplines to model social relationships based on reciprocal altruism (Axelrod and Hamilton, 1981; Axelrod, 1984; Boyd, 1988; Nesse, 1990; Trivers, 1971). To elucidate the neural substrates of the emotional and cognitive processes that support cooperative, reciprocally altruistic relationships, we investigated game-related neural activations with fMRI as subjects played an iterated Prisoner's Dilemma Game with other subjects outside the scanner. In this game, two players independently 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 in that round. There are four possible outcomes of a round: player A and player B cooperate (CC), player A cooperates and player B defects (CD), player A defects and player B cooperates (DC), or 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 (Figure 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 two separate experiments, we scanned a total of 36 women with fMRI as they played the Prisoner's Dilemma Game. Experiment 1 was designed to isolate the neural correlates of cooperation and noncooperation in social and nonsocial contexts, and of monetary reinforcement of behavior. Nineteen subjects were scanned during each of four game sessions. The results of the first experiment revealed different patterns of neural activation depending on whether the playing partner was identified as a human or a computer. This motivated a second experiment in which 17 subjects were scanned during

[1]Correspondence: jrilling@princeton.edu
[2]Present address: Green Hall, Princeton University, Princeton, New Jersey 08544.

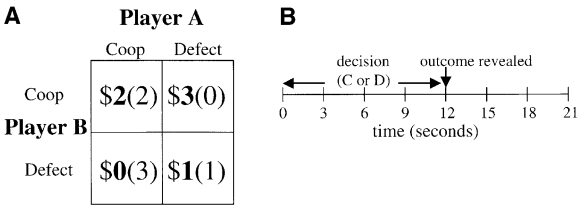


Figure 1. Study Design

(A) Payoff matrix used for the four outcomes in the Prisoner's Dilemma Game. Scanned subject's choices (C or D; player A) are listed atop columns and nonscanned subject's choices (C or D; player B) are listed aside rows. Dollar amounts in bold are awarded to player A. Amounts in parentheses are awarded to player B.

(B) Time course of a single round of the Prisoner's Dilemma Game.

each of three game sessions, focusing specifically on human versus computer interaction.

In both experiments, the players interacted via a networked computer that accepted the responses from the player inside the scanner (player A) and the nonscanned playing partner (player B). Each Prisoner's Dilemma game consisted of at least 20 rounds, with each round lasting 21 s (Figure 1B). During the first 12 s of each round, both players independently selected either to cooperate or defect. At 12 s, the square of the matrix where the two choices intersected was highlighted to reveal each player's choice and the resulting payoff for that round. Subjects were compensated in direct proportion to their accumulated earnings. The outcome was displayed for 9 s, and then the next round began (Figure 1B). Functional images were collected every 3 s. We analyzed both the BOLD response to the game outcome and the BOLD response during the decision-making period of each round. For the former, we examined the response for the epoch between 12 and 21 s. For the latter, we examined the 6 s epoch preceding the button press signaling a choice to cooperate or defect in each round.

Results

Prisoner's Dilemma Game Behavior

The number of occurrences of each outcome type was a function of the two players' choices and so was not specified in advance. Table 1 shows the average number of each outcome type per session for both experiments.

In Experiment 1, mutual cooperation was the most common outcome when the playing partner was a behaviorally unconstrained woman (see Experimental Procedures for details of experimental design). However, in the final rounds of the game, the frequency of mutual cooperation decreased, and mutual defection increased (Figure 2A). This pattern of behavior, switching to defection as the end of the game approaches, has been predicted on theoretical grounds (Axelrod, 1984) and observed empirically in previous studies (Andreoni and Miller, 1993). In games played with the provocative human confederate, the frequency of mutual cooperation was lower and mutual defection higher (Figure 2B). Because the "tit-for-tat" computer strategy initiated the game with defection, mutual cooperation was uncommon in early rounds but rebounded to levels observed with the unconstrained human partner as the game progressed before declining sharply on the very last round (Figure 2C).

Table 1. Average Number of Outcome Types per Session, for Sessions with Presumed Human Playing Partners

Experiment	Partner	CC	CD	DC	DD	Total
1	unconstrained	11.2	2.3	3.2	3.2	20
1	confederate	6.4	4.6	4.2	4.7	20
2	open ended	11.9	3.8	3.6	4	23
2	closed	9.9	2.8	2.5	5	20

As for Experiment 1, mutual cooperation was the most common outcome in games played with presumed human partners for Experiment 2 (Figures 2D and 2E). In these games, the observed reduction in mutual cooperation in rounds 18–20 was forced by the computer strategy, which defected automatically in these rounds. The rebound of mutual cooperation in rounds 20–23 (Figure 2D) was induced by programmed cooperation by the computer in these three rounds. When subjects in Experiment 2 were instructed that they were playing the game with a computer rather than another person, mutual cooperation was less common throughout the game (Figure 2F; paired $t = 4.90$, 19 df, $p < 0.001$), even though subjects were actually playing against exactly the same computer strategy.

In both experiments, there was a tendency for subject pairs who arrived at a CC outcome to persist with mutual cooperation so that a CC outcome in the current round was most likely to be followed by a CC outcome in the next (Table 2).

fMRI Data

Neural Activations Related to the Reaction to the Game Outcome (Seconds 12–21 of Each Round)

The BOLD response to a given outcome type (i.e., CC, CD, DC, or DD) could be attributable to an effect of either the partner's choice, the player's choice, or to an interaction between the two that exceeded the sum of their respective main effects. The statistical interaction is of special interest because it relates specifically to the social interaction rather than to independent effects of player and partner decisions. Therefore, we began by testing for main effects of player and partner choices and for an interaction between the two. More specifically, for the games in which player A assumed she was playing with a human partner, we examined the main effect of player A's decision (irrespective of player B's decision) on neural activity in player A during the reaction epoch (seconds 12–21 for each round), the main effect of player B's decision (irrespective of player A's decision) on neural activity in player A, and the interaction effect of player A and player B's decisions on neural activity in player A.

Experiments 1 and 2 were analyzed separately. The following procedure was used to identify brain regions that were activated in both experiments. For a given contrast, we masked the thresholded ($p < 0.01$) t statistic map for that contrast from Experiment 1 and limited our analysis of Experiment 2 to voxels within the mask. We then calculated the same contrast for Experiment

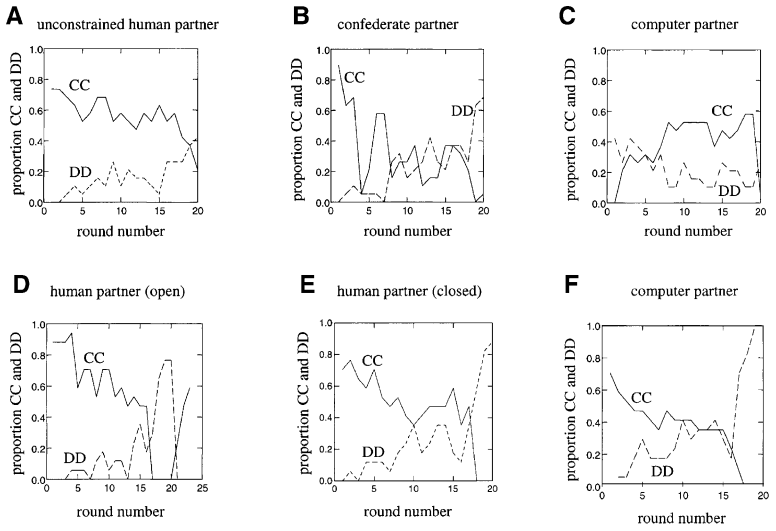


Figure 2. Round by Round Depiction of the Proportion of All Subjects Pairs Who Mutually Cooperated in Three Sessions from Each of Two Experiments

Results for Experiment 1 are shown for sessions with (A) unconstrained human playing partner, (B) provocative confederate playing partner, (C) computer partner playing "tit-for-tat". Results for Experiment 2 are shown for sessions with (D) an assumed human partner with the number of rounds unspecified in advance, and an assumed (E) human and (F) computer partner with the number of rounds specified in advance.

2. Voxels within the mask that survived at $p < 0.01$ were then reported as replicated activations.

Main Effects

No main effects replicated across both experiments.

Interaction Effects

Consistent interaction effects were observed across the two experiments (Table 3). These effects were restricted to one side of the interaction, namely ([CC + DD]) − [CD + DC]) and not the opposite ([CD + DC] − [CC + DD]). That is, for the regions listed in Table 3, the neural response to game outcomes CC and DD combined was greater than activation following outcomes CD and DC combined. This is of interest because CD and DC outcomes are typically aversive to at least one of the two subjects and are consequently unlikely to be repeated. On the other hand, CC and DD are more stable in the sense that subjects often persist with these outcomes. Hence, CC and DD outcomes might be considered behaviorally reinforcing. In terms of spatial extent, the largest activation for this interaction involving symmetric social behavior is in the anteroventral striatum and subgenual anterior cingulate cortex (BA 25). The striatal activation includes the caudate nucleus and nucleus accumbens (Nac), both of which receive midbrain dopamine projections known to be involved with processing reward (Schultz, 1998). The ventromedial/orbitofrontal cortex (OFC), another brain area involved in reward processing (Rolls, 1999), was also activated for the interaction (Figure 3).

Table 2. Transition Probabilities Following CC Outcomes in Experiments 1 and 2

Experiment	Partner	CC	CD	DC	DD
1	unconstrained human	0.79	0.06	0.11	0.04
1	confederate human	0.47	0.34	0.13	0.06
2	assumed human open ended	0.82	0.11	0.05	0.01
2	assumed human close ended	0.77	0.14	0.06	0.03

The probability of each outcome, given a CC outcome in the previous round, is listed as a function of experiment and partner type.

Event-Related Plots

Event-related plots were constructed to determine which outcome or outcomes were responsible for the interaction effect. Plots were made for the peak voxels in the anteroventral striatum and OFC ROIs for every subject. Across both experiments, for each session with a human partner in which all four outcomes occurred ($n = 61$ sessions), we examined which of the four outcomes had the largest amplitude-fitted response in the general linear model and whether that response was positive or negative. The plot for one subject is shown in Figures 4A and 4B. For the peak voxel of the anteroventral striatal ROI, CC had the largest fitted response for 30 of the 61 sessions, whereas DD had the largest response for 19 sessions, and both CD and DC had the largest response for only 6 sessions each (Figure 4C). This distribution differed significantly from chance (chi-square = 26.8, $p < 0.001$). In 25 of the 30 sessions where CC had the largest fitted response, that response was positive. In other words, CC was associated with increased activation relative to the other conditions, rather than less deactivation. Though not as pronounced, there was also evidence of deactivation for the CD and DC outcomes at this location (Figure 4C). In 19 of the 23 cases where CD had the smallest fitted response, that response was negative, and the DC response was negative in 21 of the 23 cases where DC had the smallest fitted response.

For the peak voxel in the OFC ROI (see Figure 3), CC had the largest fitted response for 32 of the 61 sessions (versus 15 for DD, 7 for CD, and 7 for DC), and 24 of these were positive in amplitude. This distribution also differed significantly from chance (chi-square = 27.9, $p < 0.001$). Thus, for both ROIs, the interaction effect was dominated by the positive response to CC.

CC versus the Other Outcomes

Given that the BOLD response to CC was largely responsible for the interaction effect, we decided to focus more specifically on this outcome by contrasting the BOLD response to the CC outcome with the average response of the other three outcomes combined. Masking the results of the Experiment 2 with Experiment 1 revealed larger and more significant activations in ventromedial

Table 3. Reaction Epoch: Location of Brain Activations in Player A Related to the Interaction of Player B's and Player A's Decision to Cooperate or Defect

Brain Region	Coordinates	Peak t Statistic	Number Voxels
Player A × player B interaction (CC + DD) − (CD + DC)			
R caudate	6 18 0	4.94	8
L post-central gyrus (BA 1/3)	−27 −39 60	3.86	5
R central sulcus (BA 4)	18 −30 72	3.35	7
R medial frontal gyrus (BA 11)	6 51 −18	3.26	7

Activations are for Experiment 2 ($p < 0.01$, $n = 17$ subjects) after limiting the search volume to voxels that survived a statistical threshold $p < 0.01$ in Experiment 1 ($n = 19$ subjects). Activations consisting of fewer than five contiguous voxels are not reported. L, left hemisphere; R, right hemisphere.

frontal cortex and anteroventral striatum than were found for the interaction analysis (Table 4; Figure 5A). In contrast to the interaction t map, the activation in the ventromedial frontal cortex extended dorsally into the rostral anterior cingulate cortex (BA32). In Figures 5C and 5D, the statistical parametric map for this contrast is displayed on a spatially normalized EPI image to demonstrate that the observed ventromedial frontal/orbitofrontal activation is not within an area of high magnetic susceptibility artifact. Figure 6 is an event-related plot for one subject for the peak voxel of the OFC ROI.

CC Compared with Monetary Reinforcement

To investigate the possibility that this pattern of activation was simply a consequence of monetary reinforcement ($2 for a CC outcome), we tested for a condition (human partner versus control) by monetary outcome ($2 versus others) interaction in Experiment 1. That is, we asked whether earning $2 when playing with a human partner produced more activation than earning $2 in the nonsocial control condition. The test for interaction revealed activation in the anteroventral striatum and OFC (Figure 5E). Thus, the anteroventral striatum, rACC, and OFC were activated more by reciprocated social cooperation than by a $2 reward in a nonsocial context.

CC with Computer versus Human Playing Partners

Finally, sessions with computer playing partners were included in both experiments to determine whether activations detected with human partners were specific to human social interaction. In both experiments, mutual cooperation with a computer playing partner activated regions of the ventromedial/orbital frontal cortex (BA 11) that were also activated with human playing partners (Table 5), although for Experiment 1, the overlap was only observed if the t statistic threshold was decreased to $p < 0.05$. In neither of the two experiments did mutual cooperation with a computer activate the rostral anterior cingulate or the anteroventral striatum observed for human playing partners.

Neural Activation Related to Social Decision Making (6 s Epoch Preceding the C or D Choice)

Given that subjects make their choices early (mean = 3.4 s) within the 12 s decision-making period of each round, it seems likely that the 9 s period during which the game outcome was displayed (and over which neural activity was sampled for the reaction epoch) involves not only the reaction to the outcome of the current round but also decision making related to the choice for the next round. In our experiments, CC outcomes tended to occur in consecutive strings so that a CC outcome in one round was most likely to be followed by a CC outcome in the next (Table 2). Thus, the intervals following CC outcomes typically involved a decision to continue cooperating, rather than defect. To more systematically investigate neural activity related to opting for social cooperation, a model was specified that compared the BOLD signal in the 6 s interval immediately preceding the choice to cooperate or defect (as marked by a button press), and analyzed as a function of the partner's decision in the previous round. The four conditions were XC,CX (i.e., choosing to cooperate after the partner had cooperated in the previous round), XC,DX (i.e., choosing to defect after the partner had cooperated in the previous round), XD,CX (i.e., choosing to cooperate after the partner had defected in the previous round, and XD,DX (i.e., choosing to defect after the partner had defected in the previous round).

The decision to cooperate following a cooperative choice by one's partner in the previous round activated the left anterior caudate and the right post-central gyrus (Table 6; Figure 7). The decision to reciprocate cooperation was also associated with activation in two regions that were activated following mutual cooperation in the reaction epoch: the rostral anterior cingulate cortex and the anteroventral striatum (Table 6; Figure 7).

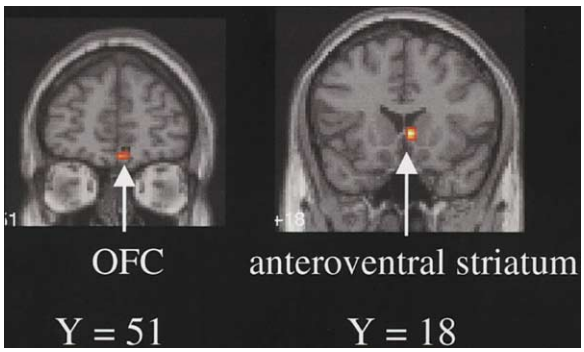
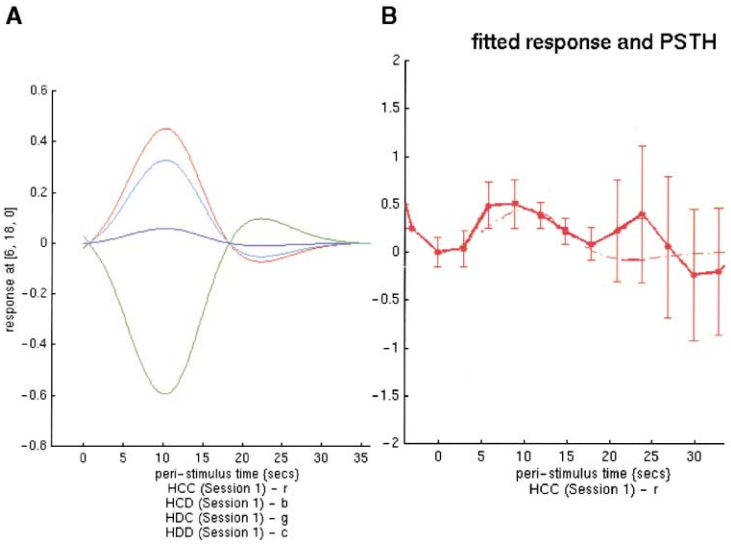


Figure 3. Reaction Epoch

Activation in player A when playing with an assumed human partner. Voxels activated for the interaction of player A and player B's choices (CC − CD) − (DC − DD) in Experiment 2 ($p < 0.01$), after masking the results with voxels activated for the same contrast in Experiment 1 ($p < 0.01$). OFC = ventromedial frontal/orbitofrontal cortex.



C

	largest response	smallest response
CC	30	5
CD	6	23
DC	6	23
DD	19	10
total	61	61

Figure 4. Reaction Epoch: Event-Related Plot for the Peak Voxel in the Anteroventral Striatum

(A) Fitted response for all four outcomes in a single subject, CC = red, CD = blue, DC = green, and DD = cyan.

(B) Raw data for CC outcome for a single subject. The outcome is revealed at t = 0 s and displayed for 9 s.

(C) Distribution of outcome types having the largest and smallest amplitude fitted response in the GLM, across all 61 sessions.

Discussion

Reaction Epoch

Postscan subject interviews revealed that mutual cooperation was typically considered the most personally satisfying outcome. The more profitable DC outcome was typically described as less desirable than CC outcomes either because it provoked guilt over having profited at the partner's expense, or because subjects realized that the outcome would likely provoke defection by the partner, thereby destabilizing the relationship and leading to lower cumulative earnings. Combined with the neuroimaging and electrophysiological evidence linking the orbitofrontal cortex (Francis et al., 1999; O'Doherty et al., 2001; Rolls, 1999; Schultz et al., 2000; Thut et al., 1997) and ventral striatum (Berns et al., 2001; Breiter et al., 2001; Koepp et al., 1998; Pagnoni et. al., 2002; Schultz, 1998) to reward processing, this suggests that the orbitofrontal and anteroventral striatal activations associated with the CC outcome in our experiment may relate to the rewarding effects of arranging and/or experiencing a mutually cooperative social interaction.

Recent evidence indicates that reward-related neural activity is greater for unpredicted than predicted rewards (Schulz et. al., 1997). Our results are consistent with this observation insofar as subjects exerted no control over their partners' decisions so that the game outcome always had an element of unpredictability. A subject could never know for certain if her cooperative choice would be reciprocated. However, when subjects choose to cooperate, they are guessing that their partner will do the same; and when their cooperation is met with defection, an anticipated reward is omitted. Schulz et al. (1997) have demonstrated that the omission of expected rewards deactivates midbrain dopamine neurons (decreases spike production), an observation that leads to the prediction that the CD outcome should be associated with deactivation of the midbrain and perhaps the striatal neurons to which it projects. Indeed, CD was often associated with deactivation of the anteroventral

Table 4. Reaction Epoch: Location of Significant Brain Activations for the Contrast Comparing the CC Outcome with the Average of the Other Three Outcomes when Playing with a Human Partner

Region	Coordinates	Peak t Statistic	Number Voxels
CC versus all other choices			
L paracentral lobule (BA 7)	−18 −39 54	6.45 *	22
R caudate	3 18 0	5.35 *	14
L postcentral gyrus (BA 1)	−39 −30 60	4.3	7
R medial frontal gyrus (BA 11)	3 48 −12	4.03	28
rostral anterior cingulate gyrus (BA 32)	−3 51 6	3.65	6
L superior temporal gyrus (BA 22/42)	−51 −30 12	3.57	7
R paracentral lobule (BA 5/7)	18 −45 60	2.99	5

Activations are for Experiment 2 ($p < 0.01$, n = 17 subjects) after limiting the search volume to voxels that survived ($p < 0.01$) in Experiment 1 (n = 19 subjects). Voxels surviving a corrected p value <0.05 after small volume correction with the mask from Experiment 1 are marked with an asterisk. Activations consisting of fewer than five contiguous voxels are not reported in the table. L, left hemisphere; R, right hemisphere.

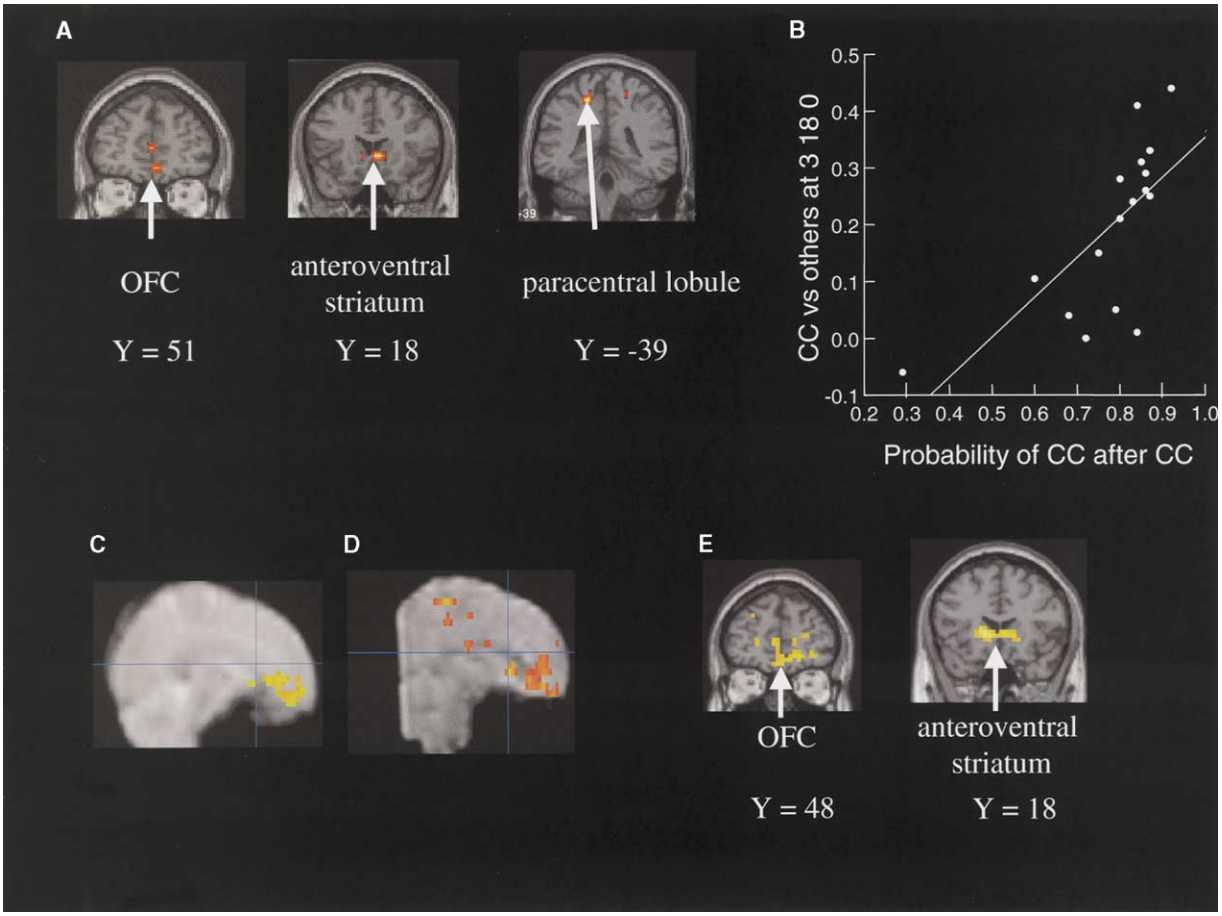


Figure 5. Reaction Epoch: Activation in Player A in Response to CC Outcomes

(A) Voxels activated more by mutual cooperation (CC) than the mean of the other three outcomes in Experiment 2 ($p < 0.01$) after masking the results with voxels activated for the same contrast in Experiment 1 ($p < 0.01$).

(B) Plot of contrast value for CC versus others in the peak voxel of the anteroventral striatal ROI against the probability of CC repeating in consecutive rounds, for the 17 subjects in Experiment 2.

(C and D) Statistical parametric map for the contrast in (A) displayed on a normalized EPI image for (C) Experiment 1 and (D) Experiment 2. Statistical t images are thresholded at $p < 0.01$ (uncorrected).

(E) Voxels showing a significant condition (human versus control) by monetary outcome ($2 versus others) interaction ($p < 0.01$) in which the response to $2 is greater for the social than the control condition. Data are for Experiment 1 only because the control condition was not included in Experiment 2. OFC = ventromedial frontal/orbitofrontal cortex.

striatum in our experiment (Figure 4C). DC was also often associated with striatal deactivation, an observation that could be reconciled with predictions if subjects find DC more aversive than DD (they defect to protect themselves from potential exploitation by a defecting partner but experience guilt upon realizing a DC outcome).

Cooperating is always risky given the unpredictability of the intentions of another person in a social dyad. So, it is possible that the observed pattern of activation relates more generally to a realization of success following a risky decision and not specifically to a reciprocated act of altruism. Alternatively, it may be the case that the observed activation is associated with positive feelings toward one's partner; that activation of anteroventral striatum and OFC can result in feelings of trust and comradery that reinforce the cooperative act, superseding any conscious recognition that material gains will flow from mutual cooperation. Indeed, some theorists have proposed that many of the social emotions have evolved in the service of preserving social relationships based on reciprocity (Trivers, 1971; Frank, 1988). This agrees with the everyday observation that we often behave altruistically toward others simply because we like them, not because we consciously calculate that they are likely to reciprocate in the future.

Subjects who find the CC outcome rewarding would be expected to persist with CC outcomes more than other subjects. We were therefore interested in whether the magnitude of the activation in the anteroventral striatum and OFC was related to subjects' tendencies to persist with CC outcomes. Indeed, subject's who were more likely to experience consecutive CC outcomes had greater activation in the peak voxel of the anteroventral striatum ROI ($r = 0.70$; $p = 0.002$, Figure 5B). There was no such behavioral correlation for the peak voxel of the OFC ROI.

Comparisons between human and computer activation patterns show that the orbitofrontal activation asso-

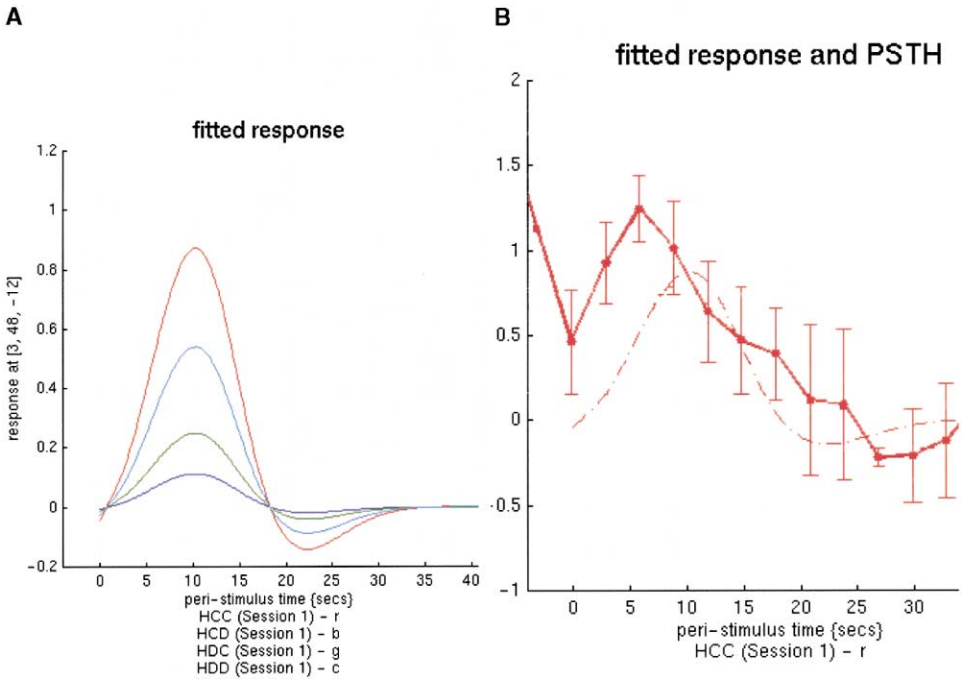


Figure 6. Reaction Epoch: Event-Related Plot for the Peak Voxel in the OFC for the Contrast CC versus Other Outcomes

(A) Fitted response for all four outcomes in a single subject, CC = red, CD = blue, DC = green, and DD = cyan.

(B) Raw data for CC outcome for a single subject. The outcome is revealed at t = 0 s and displayed for 9 s.

ciated with CC outcomes is not specific to rewarding *human* social interaction but can also be elicited by interactive computer programs, at least when the latter are programmed to be responsive to their partner's behavior. On the other hand, cooperation with a human partner may be a more effective stimulus for striatal mechanisms related to reward since we did not observe striatal activation in association with CC for computer partners.

Finally, we note that the most significant activation in association with the CC outcome was in neither the OFC nor striatum, but in somatosensory association cortex in the medial posterior parietal lobe (BA 7; see Table 4 and Figure 5A). A prominent theory of emotion processing proposes that a neural representation of an organism's somatic state is an important referent of emotional experience (Bechara et. al., 2000), and that somatosensory association cortex is largely where these representations are formed. Thus, one possible interpretation of this activation is that it is related to a representation of a somatic state of an emotional experience that follows mutual cooperation.

Decision Making

The decision to cooperate following cooperation by one's partner in the previous round was associated with activation in the right post-central gyrus. The post-central gyrus activation is in primary somatosensory cortex and could be a neural representation of a somatic response to an imagined decision to reciprocate cooperation (Damasio, 1994; Aziz et. al., 2000).

The anteroventral striatum was also activated for this contrast (i.e., XC,CX versus others). Our social decision-making epoch (for round n + 1) trails but overlaps with the reaction epoch (to round n), raising the possibility that the anteroventral striatal activation represents pro-

Table 5. Reaction Epoch: Location of Significant Brain Activations for Contrast Comparing the CC Outcome with the Average of the Other Three Outcomes when Playing with a Computer Partner

Region	Coordinates	Peak t Statistic	Number Voxels
Experiment 1 (n = 19 subjects)			
No activations			
Experiment 2 (n = 17 subjects)			
L insula	−39 −3 18	4.7	18
L OFC (BA 11)	−3 36 −12	4.39	12
L anterior insula	−27 9 6	4.15	6
L frontal pole (BA 10)	−6 66 6	3.86	6
R OFC (BA 11)	6 48 −18	3.35	18

Activations for computer partners ($p < 0.01$) were masked with the results of the same contrast for human partners ($p < 0.01$) to show areas of overlap. Activations consisting of fewer than five contiguous voxels are not reported in the table. L, left hemisphere; R, right hemisphere.

Table 6. Decision-Making Epoch: Location of Significant Brain Activations for Contrast Comparing Cooperation Following a Cooperative Choice by One's Partner in the Previous Round (XC,CX) with the Average of the Other Three Outcomes, (XD,CX), (XC,DX), (XD,DX), when Playing with a Human Partner

Region	Coordinates	Peak t Statistic	Number Voxels
(XC,CX) versus all other conditions			
L anterior caudate	−12 24 12	5.23*	10
R post-central gyrus	36 −27 54	4.76*	5
R anterior cingulate gyrus (BA32)	3 36 −6	4.06	5
R collateral sulcus	39 −45 −6	3.87	5
R caudate	6 21 6	3.79	5

Activations are for Experiment 2 ($p < 0.01$, $n = 17$ subjects) after limiting the search volume voxels that survived in Experiment 1 ($p < 0.01$, $n = 19$ subjects). Voxels surviving a corrected p value <0.05 after small volume correction with the mask from Experiment 1 are marked with an asterisk. Activations consisting of fewer than five contiguous voxels are not reported in the table. L, left hemisphere; R, right hemisphere.

longed responses to the CC outcome that extend into our decision-making epoch. However, it is also possible that some of the activations in Table 6 relate specifically to social decision making. For example, the anterior cingulate cortex is involved in the detection of cognitive conflict (Cohen et. al., 2000). The decision to persist with cooperation may involve conflict given the ever present temptation to defect and earn an extra dollar. However, processing of cognitive conflict has been linked with the caudal anterior cingulate, known as its cognitive division, whereas the cingulate activation we report here is in rostral anterior cingulate cortex (Bush et. al., 2000). Nevertheless, conflict based on emotional interference reportedly activates the rostral ACC (Whalen et. al., 1998), and it has been hypothesized that this region may be generally involved with processing conflict related to emotions (Davidson, 2000). Its involvement with emotion is also supported by multiple neuroimaging studies (Drevets and Raichle, 1998). Thus, the observed rostral anterior cingulate activation may reflect the emotional tone of social decision making.

The decision to continue cooperating following a CC outcome in the previous round also requires overcoming a putative bias that humans and other animals have to weight the attractiveness of a reward in inverse proportion to its delay (Chun and Herrnstein, 1967), a bias that would encourage our subjects to value the immediate reward of defection and its $3 payoff more than the delayed reward from sustained mutual cooperation. In other words, persisting with mutual cooperation requires restraining the impulse to defect and achieve immediate gratification. Accumulating evidence impli-

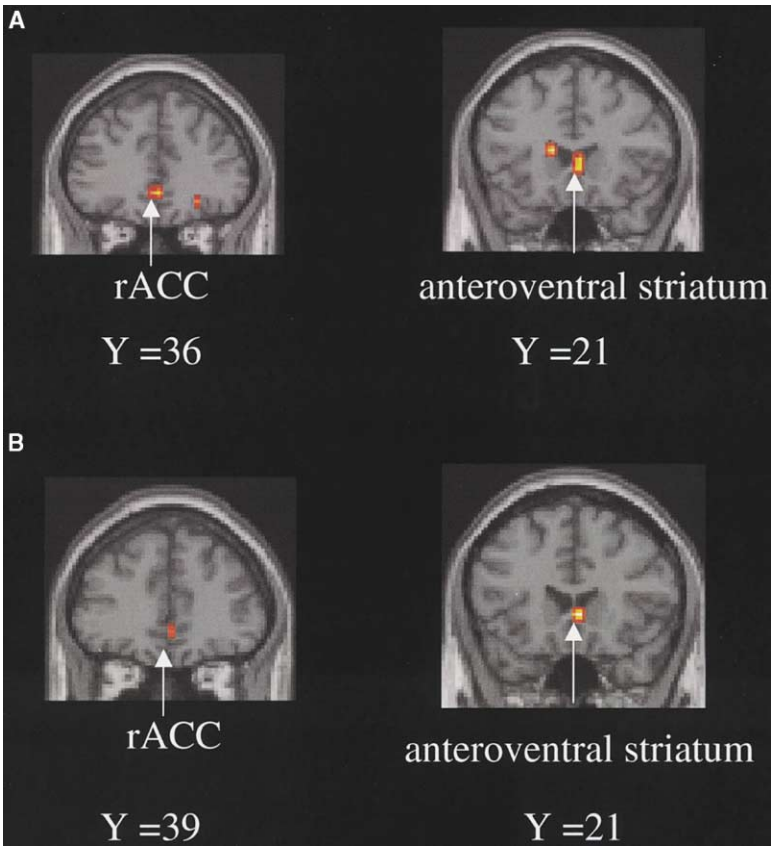


Figure 7. Decision-Making Epoch: Activation Related to the Decision-Making Epoch

(A) Voxels activated more when player A chose cooperation following a cooperative choice by her partner in the previous round (XC,CX) than for the average of the other three conditions: cooperation following partner defection (XD,CX), defection following partner cooperation (XC,DX), and defection following partner defection (XD,DX). Results are for Experiment 2 ($p < 0.01$) after masking with voxels activated for the same contrast in Experiment 1 ($p < 0.01$).

(B) Results from (A) further masked by voxels that were activated in both experiments for the contrast CC versus others during the reaction epoch (i.e., Table 4; Figure 5A) to show areas activated during both reaction and decision-making epochs. rACC = rostral anterior cingulate cortex.

cates the ventromedial frontal/orbitofrontal cortex in this role (Grafman et. al., 1996). Although ventromedial/orbitofrontal activation was not detected in our combined analysis of the decision-making epoch, it was activated ($p < 0.001$) in Experiment 2. Patients with damage to the ventromedial frontal lobe are characterized by impaired personal and social decision making (Damasio, 1994; Bechara et. al., 2000) and have been described as lacking the ability to delay gratification. Analogously, subjects who defect out of mutually cooperative social interactions in the Prisoner's Dilemma Game opt for immediate gratification (attaining the maximum payoff for that round) and may overlook or fail to consider the future consequences of defection (partner retaliation and lower cumulative earnings). The corollary is that subjects who resist the temptation to defect for short-term gain and instead persist in mutual cooperation may be better guided by the future consequences of their decisions. Thus, our findings are consistent with the notion that the ventromedial frontal cortex is involved with increasing sensitivity to distant rewards and punishments (Rogers et. al., 1999).

Summary

In summary, mutually cooperative social interactions in the Prisoner's Dilemma Game were associated with activations in anteroventral striatum, rostral ACC, and OFC that were not observed in response to monetary reinforcement in a nonsocial control condition. OFC, but not rostral ACC or anteroventral striatum, activation was also observed for mutual cooperation with a computer partner, suggesting that the ACC and striatal activations may relate specifically to cooperative social interactions with human partners.

Cooperative social interactions with nonkin are pervasive in all human societies and generally emerge from relationships based on reciprocal altruism. Such relationships arguably lay the foundation for the interdependence upon which societal division of labor is based. We have identified a pattern of neural activation that may be involved in sustaining cooperative social relationships, perhaps by labeling cooperative social interactions as rewarding, and/or by inhibiting the selfish impulse to accept but not reciprocate an act of altruism.

Experimental Procedures

Subjects

The mean age of the 19 female participants in Experiment 1 was 28.8 years (range 20–60 years). The mean age of the 17 female participants in Experiment 2 was 23.8 years (range 20–30). The subject pool was restricted to women because of published reports that men and women play the game differently, particularly in the presence of a male experimenter (Hottes and Kahn, 1974; Rapoport and Chammah, 1965; Skotko et al., 1974).

Prior to scanning, all participants completed a 10 min computer tutorial, complete with examples, intended to familiarize them with the Prisoner's Dilemma game and with appropriate strategies for maximizing earnings. Specifically, it was pointed out that two players would both earn $40 if they both cooperated each round, but only $20 if they both defected each round. They were also told that one would earn $60 and the other $0 in the unlikely event that one player cooperated each round and the other defected each round. Subsequently, all players completed a two question multiple-choice quiz designed to assess their comprehension of the game. For subjects who answered one or both questions incorrectly, efforts were made to clarify the game. Only after the investigators concluded that subjects understood the task were subjects positioned in the scanner. Players were instructed to adopt a strategy that would maximize their earnings (with the exception of the constraints imposed on the confederate) and were compensated in direct proportion to their accumulated total.

Experimental Design

The game matrix was projected onto a screen that player A viewed through a mirror mounted on the head coil and player B viewed on a computer screen in an adjacent room. Player A indicated her decision to cooperate or defect by pressing one of two buttons on a fiber optic button box. Player B chose to cooperate or defect using two keys on the computer keyboard. When either player pressed a button or key, their choice was indicated by a color change of the corresponding selection above the column (Figure 1). Their partner's choice would not be revealed until 12 s after the round started, when the game outcome for that round was displayed. The outcome of each round was recorded and saved to a computer file that was used to specify the general linear model design matrices for each subject.

Experiment 1

For Experiment 1, subjects were informed that each game would consist of 20 rounds. In one game, the subject played 20 rounds with an unconstrained human player. In another game, the playing partner was a provocative human confederate who was constrained in her choices by having to cooperate on round 1 and defect if both players mutually cooperated on three previous rounds. Scanned subjects were unaware of these constraints. In a third session, subjects played the game with a preprogrammed computer strategy. The computer defected on round 1 of the game and subsequently played a "tit-for-tat" strategy in which it mimicked the human subject's selection from the previous round. The remaining session was a control task to determine brain activation related to monetary reward in a nonsocial context. For the control task, subjects pressed one of four buttons to select one square of an empty payoff matrix, during the first 12 s of each of 20 rounds. Each round, the computer randomly assigned $0, $1, $2, or $3 to each square of the matrix. At 12 s, the random payoff for the selected square was revealed and displayed for 9 s.

Prior to each run, subjects were reminded whom they would be playing with (the partner's name, a "preprogrammed computer strategy," or the "control task"). We hypothesized that the confederate run would be more provocative if it followed a run with a typically less provocative (i.e., more cooperative) human partner. Therefore, we used a fixed order for runs. In attempting to control for the potential confounds related to task novelty (e.g., anxiety associated with the very first run of the experiment), the control scan was placed first rather than last for 9 of the 16 subjects.

Experiment 2

In each of three sessions, subjects played against the same preprogrammed computer strategy that made cooperate or defect choices according to probabilities derived from the behavior of the unconstrained human subjects from the first experiment. That is, behavioral data from the unconstrained human subjects who played outside the scanner in Experiment 1 were used to calculate the probability that a person would cooperate, as a function of the outcome of the previous two rounds of the game. Thus, a different probability was calculated for each of the 16 possible contingencies (e.g., CC,CC; CC,CD; … DD,DD). In all three games, the computer was programmed to defect automatically in rounds 18–20 in order to ensure sufficient non-CC outcomes for statistical analysis. To protect against the possibility of subjects recognizing a predictable strategy that always defected on the last three rounds, game one included an additional three rounds (21–23) in which the computer always cooperated. In two of the three sessions, subjects were told that their playing partner was one of two women whom they had just previously met. In a third session, they were told the playing partner would be a computer. The first game was open ended in the sense that subjects were not told how many rounds the game would consist of. We included an open-ended game to control for

brain activations related to anticipating the game's end. For the two remaining games, subjects were told in advance that each would consist of 20 rounds, with one game played with a human playing partner and the other with a computer partner. The order of the two sessions was counterbalanced. The identity of the playing partner was announced before each game.

For both experiments, subjects were introduced to two human partners prior to scanning in order to reinforce the belief that they would be playing the game with real people. In both experiments, for games against the "computer," subjects were told they would play the game with a "preprogrammed computer strategy that does not play a fixed sequence of choices. Instead, it responds to your choices from earlier rounds with specified probabilities," but they were not told what strategy the computer would play.

Image Acquisition and Analysis

A 1.5 Tesla Philips NT scanner was used to acquire T1-weighted structural images and gradient echo, echoplanar T2*-weighted images with blood oxygen level-dependent (BOLD) contrast (Ogawa et al., 1992). For Experiment 1, we acquired 28 axial slices (5 mm thick) in a plane parallel to the anterior-posterior commissural line that included the entire brain volume (TR = 3000 ms, TE = 40 ms, flip angle = 90°, 64 × 64 matrix). For Experiment 2, we acquired scans coronally with a reduced TE in an attempt to minimize magnetic susceptibility artifacts in orbitofrontal and medial temporal lobe regions (O'Doherty et al., 2001). 27 slices (6 mm thick) were collected perpendicular to the anterior-posterior commissural line (TR = 3000 ms, TE = 28 ms, flip angle = 90 degrees, 64 × 64 matrix). The most caudal aspect of the occipital lobe was excluded in those cases where we could not cover the entire brain volume. For Experiment 1, functional images were acquired in four runs of 145 volumes. For Experiment 2, functional images were collected in a single run of 480 volumes in which the three games were presented in succession, with intervening 1 min rest periods. Head movement was minimized by padding and restraint.

Data were analyzed using statistical parametric mapping SPM 99 (Wellcome Department of Cognitive Neurology, London, UK). Motion correction of images to the first functional scan was performed within subject using a 6 parameter rigid-body transformation (Friston et al., 1995a). Images were then spatially normalized to the Montreal Neurological Institute (MNI) template by applying a 12 parameter affine transformation followed by nonlinear warping using basis functions (Ashburner and Friston, 1999). Images were subsequently smoothed with a Gaussian kernel of 8 mm FWHM.

A random-effects, event-related, statistical analysis was performed with SPM (Friston et al., 1995b). The experiment was analyzed as a 2 × 2 factorial design. First, a separate general linear model (GLM) was specified for each subject with four conditions representing the four possible choice pairs of a trial: CC, CD, DC, and DD. This was done for each of the game sessions (e.g., unconstrained human, confederate, computer, and control for Experiment 1), yielding a 16 column design matrix if all four outcomes occurred at least once in all four runs. Global differences were controlled by proportional scaling (Friston et al., 1995a), high-frequency noise was removed by temporally filtering the data with a hrf low pass filter, and linear trends were removed by entering scan number as a covariate in the design matrix. For each of the game sessions, we calculated three two-sided contrast images that corresponded to the main effects of the subject's decision (contrast vector [1 1 −1 −1]), partner's decision (contrast vector [1 −1 1 −1]), and the interaction term (contrast vector [1 −1 −1 1]). Contrasts were also performed for all possible pair-wise comparisons of the four outcomes, and for each outcome relative to the mean of the other three outcomes. These individual contrast images were entered into a second-level analysis, using a separate one-sample t test. Each experiment was analyzed separately. The resulting statistical parametric map from Experiment 1 was thresholded at $p < 0.01$ (uncorrected) to generate an initial brain map of activations related to social cooperation. A mask was made of voxels surviving this threshold, and this mask was used to constrain the anatomical search space in Experiment 2. Voxels within the mask that also survived a threshold of $p < 0.01$ (uncorrected) in Experiment 2 were reported as activations. We present analyses for computer partners and assumed human playing partners. The latter includes data for both the session with the unconstrained and confederate human partners from Experiment 1 and both the open-ended and closed sessions from Experiment 2.

Acknowledgments

We thank Drs. Hui Mao and Stephan Hamann for assistance with various aspects of this study. This research was supported by a Markey Center for Neurological Sciences Fellowship (to J.K.R.), NIDA (DA00367 to G.S.B.), NIMH (MH61010 to G.S.B), and NARSAD (to G.S.B.).

Received: October 5, 2001
Revised: May 3, 2002

References

Andreoni, J., and Miller, J.H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. Econ. J. *103*, 570–585.

Ashburner, J., and Friston, K.J. (1999). Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. *7*, 254–266.

Axelrod, R., and Hamilton, W.D. (1981). The evolution of cooperation. Science *211*, 1390–1396.

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

Aziz, Q., Thompson, D.G., Ng, V.W.K., Hamdy, S., Sarkar, S., Brammer, M.J., Bullmore, E.T., Hobson, A., Tracey, I., Gregory, L., et al. (2000). Cortical processing of human somatic and visceral sensation. J. Neurosci. *20*, 2657–2663.

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

Berns, G.S., McClure, S.M., Pagnoni, G., and Montague, P.R. (2001). Predictability modulates human brain response to reward. J. Neurosci. *21*, 2793–2798.

Boyd, R. (1988). Is the repeated Prisoner's Dilemma a good model of reciprocal altruism? Ethol. Sociobiol. *9*, 211–222.

Breiter, H.C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron *30*, 619–639.

Bush, G., Luu, P., and Posner, M.I. (2000). Cognitive and emotional influences in anterior cingulate cortex. Trends Cogn. Sci. *4*, 215–222.

Chun, S.-H., and Herrnstein, R. (1967). Choice and delay of reinforcement. J. Exp. Anal. Behav. *10*, 67–74.

Cohen, J.D., Botvinick, M., and Carter, C.S. (2000). Anterior cingulate and prefrontal cortex: who's in control? Nat. Neurosci. *3*, 421–423.

Damasio, A.R. (1994). Descartes' Error: Emotion, Reason, and the Human Brain (New York: G.P. Putnam).

Davidson, R.J. (2000). Affective style, psychopathology, and resilience: brain mechanisms and plasticity. Am. Psychol. *55*, 1196–1214.

Drevets, W.C., and Raichle, M.E. (1998). Reciprocal suppression of regional cerebral blood flow during emotional versus higher cognitive processes: implications for interactions between emotion and cognition. Cognition Emotion *12*, 353–385.

Francis, S., Rolls, E.T., Bowtell, R., McGlone, F., O'Doherty, J., Browning, A., Clare, S., and Smith, E. (1999). The representation of pleasant touch in the brain and its relationship with taste and olfactory areas. Neuroreport *10*, 453–459.

Frank, R.H. (1988). Passions within Reason: The Strategic Role of the Emotions, First Edition (New York: Norton).

Friston, K., Ashburner, J., Frith, C., Poline, J.-B., Heather, J., and Frakowiak, R. (1995a). Spatial registration and normalization of images. Hum. Brain Mapp. *2*, 1–25.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J., and Turner, R. (1995b). Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage *2*, 166–172.

Grafman, J., Schwab, K., Warden, D., Pridgen, A., Brown, H.R., and Salazar, A.M. (1996). Frontal lobe injuries, violence, and aggression: a report of the Vietnam head injury study. Neurology *46*, 1231–1238.

Hottes, J.H., and Kahn, A. (1974). Sex differences in a mixed-motive conflict situation. J. Pers. *42*, 260–275.

Koepp, M.J., Gunn, R.N., Lawrence, A.D., Cunningham, V.J., Dagher, A., Jones, T., Brooks, D.J., Bench, C.J., and Grasby, P.M. (1998). Evidence for striatal dopamine release during a video game. Nature *393*, 266–268.

Lee, R.B., and DeVore, I. eds. (1968). Man the Hunter (Chicago: Aldine).

Nesse, R.M. (1990). Evolutionary explanations of emotions. Hum. Nat. *1*, 261–289.

Nowak, M.A., and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. Nature *393*, 573–577.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nat. Neurosci. *4*, 95–102.

Ogawa, S., Tank, D.W., Menon, R., Ellermann, J.M., Kim, S.G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc. Natl. Acad. Sci. USA *89*, 5951–5955.

Pagnoni, G., Zink, C.F., Mantague, P.R., and Berns, G.S. (2002). Activity in human ventral striatum locked to errors of reward prediction. Nat. Neurosci. *5*, 97–98.

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

Ridley, M. (1996). The Origins of Virtue (London: Viking).

Rogers, R.D., Owen, A.M., Middleton, H.C., Williams, E.J., Pickard, J.D., Sahakian, B.J., and Robbins, T.W. (1999). Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. J. Neurosci. *19*, 9029–9038.

Rolls, E.T. (1999). The Brain and Emotion (Oxford, UK: Oxford).

Sahlins, M.D. (1972). Stone Age Economics (Chicago: Aldine-Atherton).

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J. Neurophysiol. *80*, 1–27.

Schulz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schultz, W., Tremblay, L., and Hollerman, J.R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb. Cortex *10*, 272–283.

Skotko, V., Langmeyer, D., and Lundgren, D. (1974). Sex differences as artifact in the Prisoner's Dilemma game. J. Confl. Resolut. *18*, 707–713.

Thut, G., Schultz, W., Roelcke, U., Nienhusmeier, M., Missimer, J., Maguire, R.P., and Leenders, K.L. (1997). Activation of the human brain by monetary reward. Neuroreport *8*, 1225–1228.

Trivers, R.L. (1971). The evolution of reciprocal altruism. Q. Rev. Biol. *46*, 35–57.

Trivers, R. (1985). Social Evolution (Menlo Park, CA: Cummings).

Whalen, P.J., Bush, G., McNally, R.J., Wilhelm, S., McInerney, S.C., Jenike, N.A., and Rauch, S.L. (1998). The emotional counting stroop paradigm: a functional magnetic resonance imaging probe of the anterior cingulate affective division. Biol. Psych. *44*, 1219–1228.