A functional imaging study of cooperation in two-person reciprocal exchange

Kevin McCabe*†‡, Daniel Houser*†§, Lee Ryan*¶, Vernon Smith*†, and Theodore Trouard*‖

*Cognition and Neuroimaging Laboratories, University of Arizona, Tucson, AZ 85721; †Interdisciplinary Center for Economic Science, George Mason University, 4400 University Drive, MSN 1B2, Fairfax, VA 22030; §Department of Economics, McClelland Hall 401, P.O. Box 210108, University of Arizona, Tucson, AZ 85721-0108; ¶Department of Psychology, Psychology 312, P.O. Box 210068, University of Arizona, Tucson, AZ 85721-0068; and ‖Biomedical Engineering Program, AHSC 5302, P.O. Box 245084, Tucson, AZ 85724

Contributed by Vernon Smith, August 7, 2001

Cooperation between individuals requires the ability to infer each other's mental states to form shared expectations over mutual gains and make cooperative choices that realize these gains. From evidence that the ability for mental state attribution involves the use of prefrontal cortex, we hypothesize that this area is involved in integrating theory-of-mind processing with cooperative actions. We report data from a functional MRI experiment designed to test this hypothesis. Subjects in a scanner played standard two-person ''trust and reciprocity'' games with both human and computer counterparts for cash rewards. Behavioral data shows that seven subjects consistently attempted cooperation with their human counterpart. Within this group prefrontal regions are more active when subjects are playing a human than when they are playing a computer following a fixed (and known) probabilistic strategy. Within the group of five noncooperators, there are no significant differences in prefrontal activation between computer and human conditions.

Reciprocal exchange (1, 2) is ubiquitous to the behavior of many species (3–5). To make an exchange, it is necessary to overcome the desire for immediate gratification in favor of greater but postponed gains from mutual cooperation. Increased specialization by humans in productive activities, together with the advantages this has produced, likely has been built on improved adaptations for social exchange. The social brain hypothesis (6) explains brain growth as largely an adaptation to more sophisticated forms of social interaction. Such an adaptation would support more sophisticated reciprocity strategies such as ''goodwill-accounting'' (7) and image-scoring strategies (8, 9).

The trust game, shown in Fig. 1, illustrates one of the joint decision tree tasks used in the experiment. In this task two subjects are paired with each other as decision makers 1 (DM1) and 2 (DM2). In behavioral experiments with similar decision trees (10, 11) 50% of the DM1 subjects make the trusting move right. In response 75% of the DM2 subjects reciprocate. In these cases DM1 and DM2 reach the cooperative outcome [180, 225]. If, however, DM1 moves left the game ends at the nonrisky outcome [45, 45]. By moving right DM1 takes the risk that DM2 will defect by moving right to the outcome [0, 405].

Based on imaging and lesion experiments (12–15) that study activations associated with understanding another person's mental states (16, 17), we hypothesize that cooperative behavior requires the binding of contingent information that allows subjects to evaluate the mental states of their counterpart and commit to a stimuli-conditioned reward-motivated choice. This commitment allows subjects to delay their desire for immediate gratification (18) and achieve a higher cooperative reward. We hypothesize that the medial prefrontal cortex serves as an important convergence zone in this decision problem, because it exhibits a pattern of connectivity (19, 20) that would enable the binding of game and counterpart entities to a mutual-gains event.

Materials and Methods

Subjects were recruited to participate in a paid functional MRI experiment lasting ≈1.5 h.

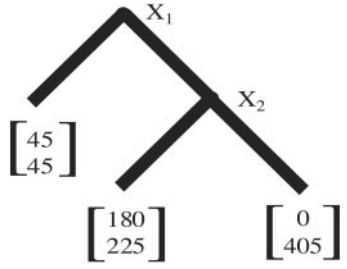

Fig. 1. Diagram of trust game used in the decision making. In the trust game, DM1 moves first (at node x_1) by either moving left, and ending the game, or moving right, giving DM2 a move. If DM1 moves right, DM2 gets the opportunity to move (at node x_2). Once DM2 moves, the game ends, DM1 is paid the top number as a payoff, and DM2 is paid the bottom number as a payoff. By moving right DM1 is trusting DM2 to reciprocate and not defect (move right). By substituting different payoff numbers, different incentives for cooperation can be studied.

Behavioral Protocol. Subjects responded to cash-payoff salient features of a visually presented two-person binary game tree by pressing response buttons with their right (move right) or left hand (move left). The subjects played the role of either first decision maker or second decision maker in each game. Second decision makers saw the first decision makers' choice before making their decision. Subjects were matched with either a human or computer counterpart and were visually informed of their counterpart's type before seeing the game tree. When the subject in the scanner played the computer they were told that it would play a fixed probabilistic strategy of 75% left and 25% right as DM2 and that the computer plays 100% right as DM1. We provided this information to the subjects to reduce the chances that the subjects would try to predict the experimenters' intentions. Similarly, when the computer moved, it did so immediately to make it less likely that subjects would anthropomorphize the computer responses. The task was administered in six scanning runs. Each run consisted of 12 randomly presented games with different payoffs with counterbalanced roles and counterparts. Behavioral and functional MRI data were recorded simultaneously from 12 right-handed subjects who were trained before entering the scanner. The subjects provided written informed consent.

Experimental Design. Subjects, in pairs, played three types of games: a trust game, a punish game, and a mutual advantage game. In each experiment one subject played the game through an interactive goggle/button system in the scanner while the other subject made decisions from a computer in the control room that was connected

Abbreviations: TR, repetition time; DM1 and DM2, decision makers 1 and 2, respectively.

‡To whom reprint requests should be addressed. E-mail: kmccabe@gmu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.211415698

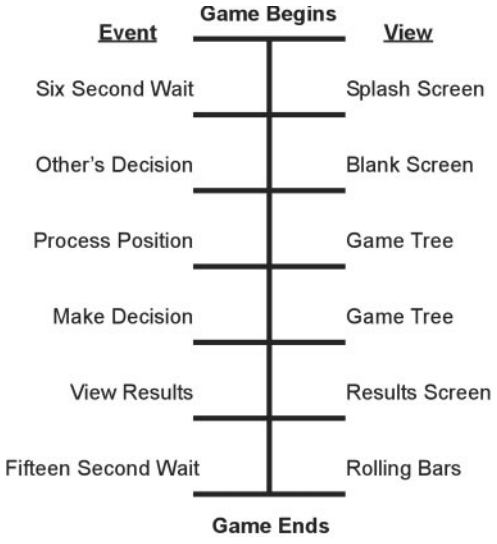

Fig. 2. Timeline of decisions and information for one game played by DM2 inside the scanner.

to the scanner system. The subject in the scanner played 72 games, in six blocks of 12, presented in a random order. In 36 of these games it was common knowledge that the subjects played each other (Human–Human), whereas in the other 36 games the subject who was in the scanner played a computer following a fixed (and known) probabilistic strategy (Human–Computer). All earnings were paid in cash at the end of each session.

The timeline for stimulus and response for a single play of a game is shown in Fig. 2. Each game begins with a splash screen, which is a 6-s introductory screen that informs subjects of which role they will be playing (DM1 or DM2) and whether their counterpart is a human or computer. If they are playing against the computer they are also told the probability that the computer will play left and the probability that the computer will play right. The remaining sequence of events depends on the subject's role. If the subject is playing as DM1 the subject sees the game tree with the appropriate payoffs and makes a decision to go either left or right; the subject then waits for DM2 to make a decision, sees the resulting play of the game, and indicates that he or she is ready to continue. If the subject is playing as DM2 the subject sees a blank screen until DM1 has made a decision; the subject then sees the game tree with DM1's decision and makes a decision, sees the resulting play of the game, and indicates that he or she is ready to continue. After looking at results the subject then watched a screen with rolling bars for 15 s .

Data Acquisition. Functional images were acquired on an 1.5-T whole body MRI scanner (Signa Echospeed, General Electric Medical Systems, Milwaukee, WI) equipped with a standard quadrature head coil. A single-shot gradient echo spiral acquisition was used (21) with repetition time (TR) = 2,000 ms, echo time = 40 ms, field of view = 220×220 mm^2, and matrix = 64×64. Fifteen contiguous 6-mm slices, oriented parallel to the AC-PC line, were imaged each TR. Six functional scans lasting an average of 7 min were taken as subjects played 12 decision problems presented in random order. High-resolution T1-weighted images were obtained over the same volume for registration of the functional results. After the functional examination, high resolution three-dimensional gradient echo images were obtained in the sagittal plane for registration of the functional data sets with the following parameters: 1.5-mm slice, TR = 22 ms, echo time = 5 ms, field of view = 250×250 mm^2, matrix = 256×256, and flip angle = 30°.

Analysis. The functional MRI data were analyzed with SPM99 (www.fil.ion.ucl.ac.uk/spm). Each subject's images were realigned, normalized to the Montreal Neurological Institute template, spatially smoothed (Gaussian kernel, full width at half maximum = 12 mm), and temporally smoothed using standard SPM99 procedures. For each subject, the areas of activation were assessed by creating statistical parametric maps based on voxel-wise linear multiple regression with conditions (convolved with SPM99's ideal hemodynamic response function) and orthogonal temporal basis functions (SPM99's default high pass filter to account for nuisance effects including temporal drift and physiological artifacts on the signal) as regressors. Data from games 3, 6, 9, and 12 are not expected to invoke theory-of-mind reasoning and were not included in the analysis. Also, because there are only four human and four computer decisions within each session, each subject's data were pooled across all of their sessions. Hence, we restricted session effects to those captured by the default high pass filter. For each individual, a cluster was considered to be significantly more active in the human than computer condition if it contained at least 12 contiguous voxels that evidenced greater activity in that condition ($P < 0.001$ uncorrected, according to SPM's standard statistical procedures). An activation was considered to be in the medial prefrontal cortex if any part the cluster fell within that region. To determine the location of activations we relied on the atlas of Tailarach and Tournoux and a visual inspection of the Montreal Neurological Institute normalized structural scans.

A conjunction analysis as implemented in SPM99 was used to determine areas of activation common to all cooperators. A second conjunction analysis was used to determine the areas of activation common to all noncooperators. Our study includes seven cooperators and five noncooperators. The null hypothesis of interest is that at any given voxel, there is not differential activation between the human and computer conditions for every relevant subject. Hence, a voxel is considered "active" across subjects if and only if, for each subject, the t statistic at that voxel exceeds a given threshold. To obtain an overall $P < 0.001$ (uncorrected), we used T thresholds of 0.32 ($P = 0.37$ uncorrected) for our seven cooperators and 0.67 ($P = 0.25$ uncorrected) for our five noncooperators. To determine the location of common activations we again relied on the atlas of Tailarach and Tournoux and a visual inspection of the Montreal Neurological Institute normalized structural scans.

Results

This study examines the bold response one TR (1.5 s) before the results screen, because decision making for cooperation is likely to be salient at this TR independent of the subject's position in the game. If the subject is DM1 or DM2 with a move, which occurs in 92% of the games, then this is the final TR of their decision period, whereas if they are DM2 without a move this is the final TR of the "wait" period. DM2s are likely to ask themselves during the wait condition, "What is my counterpart doing?" and begin to form beliefs about what a delay means about their counterpart's desires. We expect the human and computer treatments to generate differential activations associated with predicting and understanding the cooperative intentions of another human. Our analysis treats the rolling-bars condition as the baseline.

We compute the number of cooperative moves in the trust and punish games. The maximum possible score is 18. At the individual level we looked for significant activation differences when playing a human vs. the computer. Fig. 3 shows the pattern of activation for a cooperator (subject 19). The six subjects (24, 9, 25, 19, 6, and 2) with the highest cooperation scores show significant increases in activation in medial prefrontal regions during human–human interactions when compared with human–computer interactions. However, the location of these activations vary by subject. The six subjects who received the

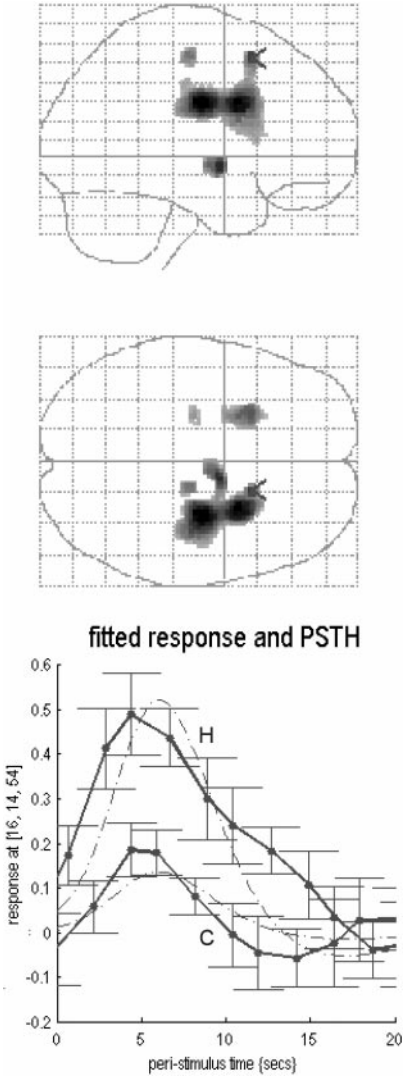

Fig. 3. Bold response of a cooperator for the contrast human (H) > computer (C). The blobs on the glass brain are clusters of at least 12 contiguous voxels that show significantly more activation in the human than computer condition (SPM *t* map, $P < 0.001$ uncorrected). The cursor on the glass brain is located at the voxel with the greatest *t* statistic within the medial prefrontal clusters. The graph immediately below the glass brains displays the peristimulus time histogram at the voxel indicated by the cursor. This is the mean of the adjusted (for time and physiological effects) response to the computer and human conditions over all the trials. The bar extends one standard error above and below the mean.

lowest cooperation scores (22, 10, 18, 21, 11, and 3) did not show significant activation differences in medial prefrontal cortex between the human and computer conditions.

An aggregate analysis suggests a common cortical network for cooperation. Consistent with previous behavioral studies (22, 23) subjects who made cooperative moves at least one third of the time, i.e., scores of six or higher, were precategorized as cooperative. This categorization results in seven cooperative subjects (24, 9, 25, 19, 6, 2, and 22), and five noncooperative subjects (10, 18, 21, 11, and 3). Fig. 4 shows the results of a conjunction analysis for human vs. computer differences for the seven cooperators ($P = 0.001$ uncorrected). A conjunction analysis ($P = 0.001$ uncorrected) with noncooperators shows no significant activation differences between the human and computer conditions.

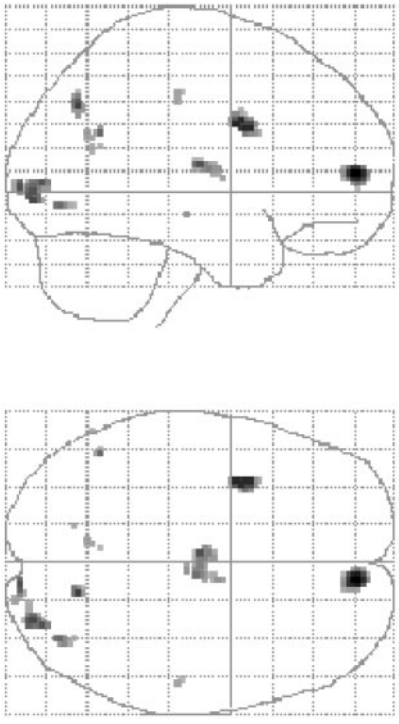

Fig. 4. Aggregate conjunction analysis of the contrast human > homputer for the seven cooperators showing a cooperation score of 6 or better. ($P = 0.001$ uncorrected.)

Discussion

Fig. 4 suggests that cooperators have a common pattern of "bold" activation differences. This suggests that cooperation requires an active convergence zone (24), possibly in prefrontal cortex, that binds joint attention to mutual gains with the inhibition of immediate reward gratification to allow cooperative decisions. Systematic activation differences are observed in (*i*) the occipital lobe (Brodmann area 17, 18), in which we hypothesize greater visual demands are placed on subjects who are trying to understand both their own payoff/incentives and the payoff/incentives of their counterparts. Common activation differences are also observed in (*ii*) the parietal lobe (Brodmann area 7), which is part of the "where" pathway for primate vision (25) and (*iii*) the thalamus. Consistent with our hypothesis that cooperation requires prefrontal control (26) activation, differences are observed in (*iv*) the middle frontal gyrus and (*v*) the frontal pole (Brodmann area 10).

In conclusion, our behavioral data shows that half the subjects in our experiment consistently attempted cooperation with their human counterpart. Within this group, and within subjects comparison, we find that regions of prefrontal cortex are more active when subjects are playing a human than when they are playing a computer following a fixed (and known) probabilistic strategy. Within the group of noncooperators we find no significant differences in prefrontal cortex between the computer and human conditions. One possible explanation for our results is that within this class of games, subjects learn to adopt game form-dependent rules of thumb when playing the computer or when playing noncooperatively with a human counterpart. In comparison, cooperation requires an active convergence zone that binds joint attention to mutual gains with sufficient inhibition of immediate reward gratification to allow cooperative decisions.

We acknowledge Georgio Coricelli and Mary Rigdon for their help in collecting the data and the help of Ming Hsu and Adam Talenfeld in assisting with the analysis of the data.

1. Trivers, R. L. (1971) *Q. Rev. Biol.* **46,** 35–57.
2. Axelrod, R. & Hamilton, W. D. (1981) *Science* **211,** 1390–1396.
3. Milinski, M. (2987) *Nature (London)* **25,** 433–435.
4. DeWaal, F. B. M. (1997) *Evol. Hum. Behav.* **18,** 375–386.
5. DeWaal, F. B. M. & Berger, M. L. (2000) *Nature (London)* **404,** 563.
6. Dunbar, R. I. M. (1996) in *Evolution of Social Behavior Patterns in Primates and Man*, eds. Runciman, W.G., Smith, J. M. & Dunbar, R. I. M. (Oxford Univ. Press, Oxford, UK).
7. McCabe, K. & Smith, V. (2000) in *Bounded Rationality: The Adaptive Toolbox*, eds. Gigerenzer, G. & Selten, R. (MIT Press, Cambridge, MA).
8. Nowak, M. A. & Sigmund, K. (1998) *Nature (London)* **393,** 573–577.
9. Wedekind, C. & Milinski, M. (2000) *Science* **288,** 850–852.
10. McCabe, K. & Smith, V. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3777–3781. (First Published March 21, 2000; 10.1073/pnas.040577397)
11. McCabe, K., Rassenti, S. & Smith, V. (1996) *Proc. Natl. Acad. Sci. USA* 13421–13428.
12. Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J. & Frith, C. D. (1995) *Cognition* **57,** 109–128.
13. Castelli, F., Happe, F., Frith, U. & Frith, C. (2000) *Neuroimage* **12,** 314–325.
14. Sabbagh, M. A. & Taylor, M. (2000) *Psychol. Sci.* **11,** 46–50.
15. Leslie, A. (2000) in *The New Cognitive Neurosciences*, ed. Gazzaniga, M. (MIT Press, Cambridge, MA), pp. 1235–1248.
16. Baron-Cohen, S. (1995) in *Mindblindness An Essay on Autism and Theory of Mind* (MIT Press, Cambridge, MA).
17. Galagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C. & Frith, C. D. (2000) *Neuropsychologia* **38,** 11–21.
18. Metcalfe, J. & Mischel, W. (1999) *Psychol. Rev.* **106,** 3–19.
19. Petrides, M. & Pandya, D. N. (1994) in *Handbook of Neuropsychology*, eds. Boller, F. & Grafman, J. (Elsevier Sciences, Amsterdam), Vol. 9.
20. Pandya, D. N. & Yeterian, E. H. (1998) in *The Prefrontal Cortex Executive and Cognitive Functions*, eds. Roberts, A. C., Robbins, T. W. & Weiskrantz, L. (Oxford Univ. Press, New York).
21. Lee, R., Glover, G. & Meyer, G. (1995) *Magn. Res. Med.* **33,** 745–754.
22. Isaac, M. & Walker, J. (1988) *Q. J. Econ.* **103,** 179–200.
23. Davis, D. & Holt, C. (1993) *Experimental Economics* (Princeton Univ. Press, Princeton).
24. Damasio, A. R. (1989) *Neural Comput.* **1,** 123–132.
25. Rolls, E. & Treves, A. (1997) *Neural Networks and Brain Function* (Oxford Univ. Press, Oxford).
26. Miller, E. (2000) *Nat. Rev. Neurosci.* **1,** 59–65.