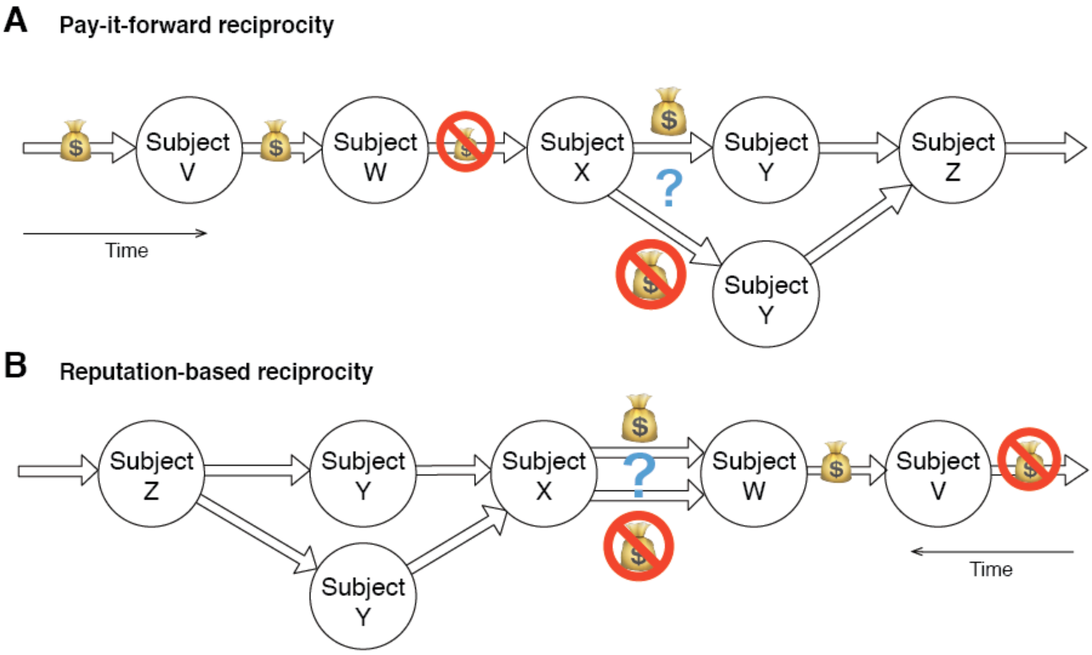
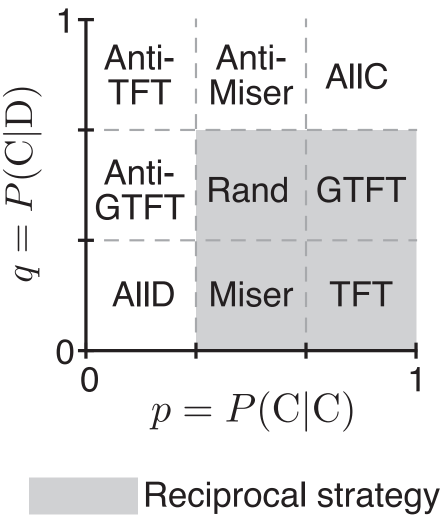
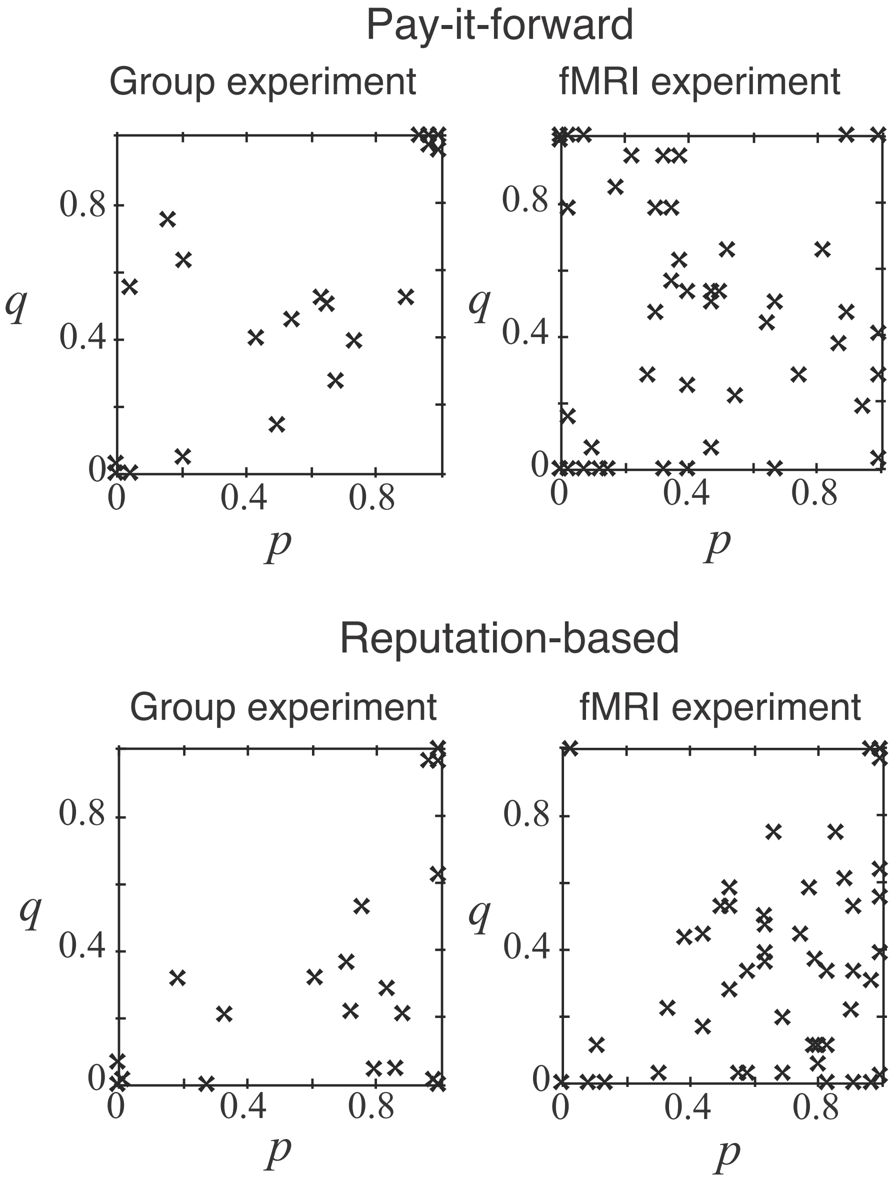
Supplementary Information
for "***Two distinct neural mechanisms underlying indirect reciprocity***"
by Watanabe *et al*.



**Supplementary Fig. S1. Illustration of indirect reciprocity games.**
**A** and **B**. Sequences of decision making in the two types of indirect reciprocity game. In the group experiment, subjects to be scanned in the subsequent fMRI experiment (i.e., subject X) were virtually embedded in the sequence. The subject that made a decision immediately after the virtual subject (i.e., subject Y) was asked to make two decisions depending on the possible actions of subject X. This was because the actual decision of the scanned subject was unknown until the fMRI experiment was performed. The two branches would merge into one chain in the subsequent step (i.e., subject Z). All subjects in the group experiment were informed of the existence of the error beforehand (see Supplementary Experimental Procedures for details and Group Experiment for the instructions that the subjects received).



**Supplementary Fig. S2. Classification of strategies.**

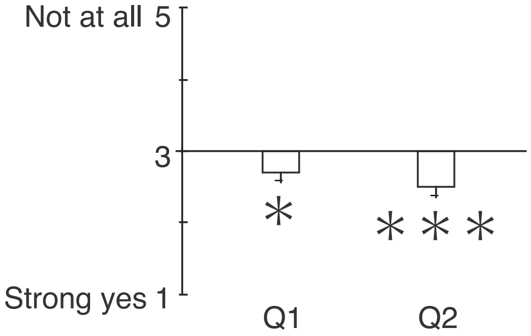


**Supplementary Fig. S3. Probability of conditional cooperation by different subjects.** A cross represents a subject. $p = P(\mathrm{C|C})$, $q = P(\mathrm{C|D})$**.** We slightly jittered the crosses originally located at ($p$, $q$) = (0, 0), (0, 1), (1, 0), and (1, 1) to indicate that multiple subjects possessed the same ($p$, $q$) values**.**

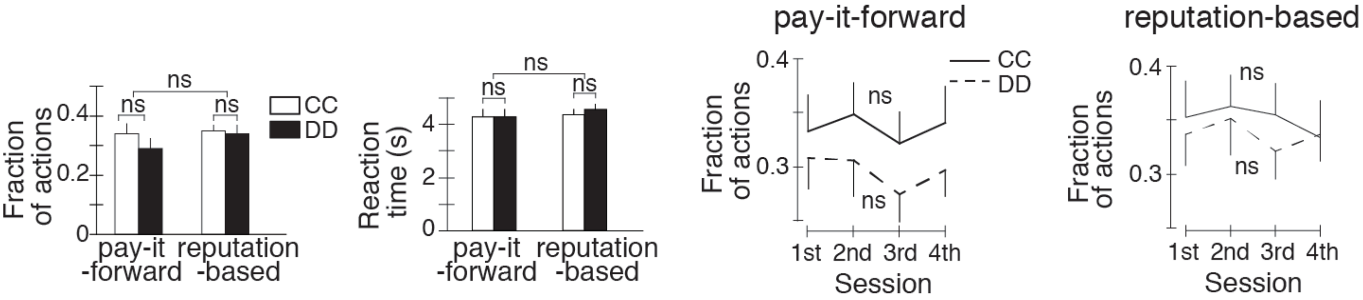
**Supplementary Results**

Questionnaires after the fMRI experiment

Immediately after the fMRI experiment, we asked the 48 subjects to answer a questionnaire to examine the extent to which they felt as if they had participated in actual economic games with other subjects. The answers to the first question (Q1): *Did you feel that you transacted real money?* (1, strong yes; 2, yes; 3, uncertain; 4, no; 5, not at all) indicated that the subjects had a sufficiently strong sense of handling real money during the fMRI experiment (score: 2.7 ± 0.14, mean ± s.e.m.; $t_{47} = 2.0$, $P = 0.05$, one-sample $t$ test; Fig. S4). The answers to another question (Q2): *Did you feel that your behavior would affect others' behavior?* (1, strong yes; 2, yes; 3, uncertain; 4, no; 5, not at all) indicated that the subjects undergoing the fMRI experiment had a sufficiently strong sense of belonging to a group of subjects (score: 2.54 ± 0.15; $t_{47} = 3.0$, $P = 0.005$, one-sample $t$ test; Fig. S4). These results suggest that the scanned subjects felt as if they had been participating in the group experiment.



**Supplementary Fig. S4. Results of the questionnaires conducted after the fMRI experiment.** *: $P = 0.05$, ***: $P = 0.005$ in one-sample $t$ test. Error bars: s.e.m. Q1: *Did you feel that you transacted real money?* Q2: *Did you feel that your behavior would affect others' behavior?*



**Supplementary Fig. S5. Behavioral results.**
Left two panels: fraction of reciprocal behaviors and reaction time of the scanned subjects. Error bars: s.e.m. Right two panels: time courses of fractions of CC and DD behavior. The averaged fractions of CC and DD for each of the four sessions are shown by the solid and dashed lines, respectively. Error bar: s.e.m. The error bars are only shown on one side of a line for clarity. ns: no significant difference.

Results of the group experiment

As in the fMRI experiment, we classified the subjects in the group experiment into nine strategies on the basis of the $p$ and $q$ values of each subject (Table S1). We then compared the distribution of the strategies for the group experiment ($N = 40$) and that for the fMRI experiment ($N = 48$). In both types of indirect reciprocity, Friedman tests of the fraction of the strategies (type of experiment [group and fMRI] × type of strategy [nine strategies]) did not detect a significant difference in the distributions between the group and fMRI experiments (pay-it-forward: $F_{1,8} = 0.11$, $P > 0.7$, reputation-based: $F_{1,8} = 0.14$, $P > 0.7$).

We also compared the distribution of strategies by conducting hierarchical log-linear analysis. We mapped the strategies on a 3 × 3 cross table based on the estimated $P$(C|C) and $P$(C|D) values (Fig. S2). On the basis of the tables, we separately conducted 2 × 3 × 3 hierarchical log-linear analysis (type of experiment [group and fMRI] × $P$(C|C) × $P$(C|D)) for the pay-it-forward and reputation-based indirect reciprocity games. None of the interaction effects including the type of experiment remained in the final models (final model for pay-it-forward: $P$(C|C) × $P$(C|D), type of experiment, $\chi^2_8 = 8.31$, $P = 0.40$; final model for reputation-based: $P$(C|C) × $P$(C|D), type of experiment, $\chi^2_8 = 6.02$, $P = 0.65$).

These results suggest that the subjects in the fMRI experiments behaved in a manner similar to that of subjects in the group experiment.

**Supplementary Table S1.**

Distribution of the strategies of the subjects

| | Anti-TFT | Anti-GTFT | AllD | Anti-Miser | Rand | Miser | AllC | GTFT | TFT |
|---|---|---|---|---|---|---|---|---|---|
| **Pay-it-forward** | | | | | | | | | |
| Group experiment | 0.10 | 0.05 | 0.20 | 0.00 | 0.20 | 0.05 | 0.25 | 0.05 | 0.10 |
| fMRI experiment | 0.21 | 0.02 | 0.21 | 0.04 | 0.17 | 0.13 | 0.06 | 0.10 | 0.06 |
| **Reputation-based** | | | | | | | | | |
| Group experiment | 0.00 | 0.00 | 0.25 | 0.00 | 0.05 | 0.05 | 0.15 | 0.10 | 0.40 |
| fMRI experiment | 0.02 | 0.00 | 0.10 | 0.00 | 0.15 | 0.13 | 0.10 | 0.19 | 0.31 |

Brain activations related to reciprocal behavior

We searched for brain regions activated by reciprocal behavior (i.e., CC and DD) as compared with non-reciprocal behavior (i.e., CD and DC) separately in each type of indirect reciprocity. With the exception of the dorsal precuneus, the regions activated by reciprocal behavior were similar between the two types of indirect reciprocity (Fig. 1G, Table S2). In fact, we did not find significant differences in brain activity during reciprocal behavior between the two types of game (i.e., no significant activation in $(CC+DD–CD–DC)_{\text{pay-it-forward}}$ versus $(CC+DD–CD–DC)_{\text{reputation-based}}$; $P > 0.005$, uncorrected).

**Supplementary Table S2**

Brain regions related to reciprocal behavior

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **Pay-it-forward** | | | | |
| **(CC+DD) – (CD + DC)** | | | | |
| Lt inferior frontal gyrus | -56 | 6 | 38 | 4.5 |
| posterior dorsal mPFC | -2 | 16 | 50 | 4.3 |
| Rt inferior parietal lobule | 30 | -50 | 50 | 4.2 |
| Lt inferior parietal lobule | -48 | -38 | 56 | 6.1 |
| **Reputation-based** | | | | |
| **(CC+DD) – (CD + DC)** | | | | |
| Lt inferior frontal gyrus | -56 | 8 | 32 | 4.5 |
| Rt inferior parietal lobule | 24 | -60 | 46 | 4.2 |
| dorsal precuneus | -20 | -64 | 46 | 4.3 |
| posterior dorsal mPFC | 0 | 16 | 48 | 4.9 |
| Lt inferior parietal lobule | -46 | -36 | 54 | 4.5 |

$P_{\text{FWE}} < 0.05$. Rt, right; dmPFC, dorso-medial prefrontal cortex; Lt, left. MNI, Montreal Neurological Institute.

## Supplementary Table S3

Brain regions differentially or commonly related to two types of indirect reciprocity

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **(CC−DD) pay-it-forward > (CC−DD) reputation-based** | | | | |
| Rt. AI | 34 | 30 | 0 | 5.3 |
| post dmPFC | -4 | 28 | 50 | 4.6 |
| **(CC−DD) reputation-based > (CC−DD) pay-it-forward** | | | | |
| ventral precuneus | -4 | -62 | 24 | 4.9 |
| dorsal precuneus | 0 | -50 | 66 | 5.1 |
| **(CC−DD) pay-it-forward & (CC−DD) reputation-based** | | | | |
| Lt. caudate | -10 | 0 | 16 | $P < 10^{-8}$ |
| ant dmPFC | -6 | 50 | 18 | |

Activations specific to either type of indirect reciprocity: $P_{\text{FWE}} < 0.05$. Activations found in the conjunction analysis: $P < 10^{-8}$ (i.e., $10^{-4} \times 10^{-4}$). AI, anterior insula; post, posterior; ant, anterior. See Table S2 for the other abbreviations.

DD-specific brain activity

Table S3 shows the coordinates of the brain regions whose CC-specific activity (i.e., CC−DD) differed significantly between the two types of indirect reciprocity ($P_{\text{FWE}} < 0.05$). The table also shows the coordinates for the brain regions that were commonly activated by both types of indirect reciprocity ($P < 10^{-8}$).

We also examined the difference in DD-specific activity (i.e., DD−CC) between the two types of indirect reciprocity by using the same ANOVA as that used in the case of CC-specific activity. We identified significant activations in the dorsal and ventral precuneus in a contrast between (DD−CC)$_{\text{pay-it-forward}}$ > (DD−CC)$_{\text{reputation-based}}$. The right AI and post dmPFC showed significant activations in a contrast between (DD−CC)$_{\text{reputation-based}}$ > (DD−CC)$_{\text{pay-it-forward}}$.

However, these activations do not reflect DD-specific activity. These brain regions are not significantly more activated during DD than during CC in each type of indirect reciprocity (Fig. 2D and 2E). For example, the dorsal precuneus was significantly active in a contrast between (DD−CC)$_{\text{pay-it-forward}}$ > (DD−CC)$_{\text{reputation-based}}$, which would lead us to speculate that DD-specific activity in the dorsal precuneus may be higher in pay-it-forward reciprocity than in reputation-based reciprocity. However, there was no significant difference in the brain activity between the CC and DD conditions within the pay-it-forward indirect reciprocity game (Fig. 2E). Therefore, we concluded that the significant effect detected by the ANOVA was derived mainly from an increase in the activity of the dorsal precuneus during CC in the reputation-based reciprocity game, rather than from a DD-specific activity in the pay-it-forward reciprocity game.

Furthermore, exploratory whole-brain analysis of brain activity within each type of game supports the same results. For example, the dorsal precuneus was significantly activated in CC relative to DD (i.e., CC−DD) in the reputation-based reciprocity game, but not in DD relative to CC (i.e., DD−CC) in the pay-it-forward reciprocity game (Table S4). The other four brain regions detected by the ANOVA (i.e., the AI, post dmPFC, dorsal precuneus, and ventral precuneus) showed significantly greater activity in CC than DD in the corresponding type of game (Fig. 2D and 2E, Table S4). In conclusion, the four regions are likely to be involved in CC in the corresponding indirect reciprocity.

It should be noted that DD-specific activity in the bilateral inferior parietal lobules was observed in both types of game ($P < 10^{-8}$ in a conjunction analysis).

**Supplementary Table S4**

Brain regions related to each type of indirect reciprocity

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **Pay-it-forward** | | | | |
| **CC – DD** | | | | |
| Rt amygdala | 30 | -6 | -18 | 4.5 |
| Rt anterior insula | 28 | 30 | 6 | 5.3 |
| Lt caudate | -8 | 6 | 12 | 4.7 |
| anterior dorsal mPFC | -2 | 50 | 12 | 4.6 |
| Lt middle frontal gyrus | -38 | 16 | 40 | 4.5 |
| posterior dorsal mPFC | 2 | 34 | 44 | 5.5 |
| | | | | |
| **DD – CC** | | | | |
| Rt inferior temporal gyrus | 50 | -58 | -6 | 3.7 |
| Lt superior temporal sulcus | -66 | -26 | 15 | 5.7 |
| Lt inferior parietal lobule | -62 | -20 | 40 | 4.5 |
| | | | | |
| **Reputation-based** | | | | |
| **CC – DD** | | | | |
| Lt caudate | -14 | 6 | 14 | 4.5 |
| anterior dorsal mPFC | -2 | 50 | 18 | 4.6 |
| ventral precuneus | -4 | -64 | 28 | 5.3 |
| Rt superior temporal lobule | 34 | -40 | 60 | 4.6 |
| dorasl precuneus | 0 | -46 | 66 | 4.7 |
| | | | | |
| **DD – CC** | | | | |
| Lt inferior parietal lobule | -56 | -34 | 42 | 4.7 |
| Rt middle frontal gyrus | 26 | 34 | 50 | 4.6 |

$P_{\mathrm{FWE}} < 0.05$. See Table S2 and S3 for abbreviations.

Robustness of neuroimaging results against variation in subject grouping

We repeated the neuroimaging analysis after excluding Rand strategy (Fig. S2) from the definition of reciprocal subject. Then, the number of reciprocal subjects decreased from 21 to 14. With the data collected from the 14 subjects, we calculated the difference in the CC-specific activity between the two types of indirect reciprocity (Supplementary Table S5). The results were qualitatively the same as those for the 21 reciprocal subjects including the Rand strategists. Although the statistical significance for the present results was weaker than that for the original results, all significant activations observed in the original analysis (Supplementary Table S3) were reproduced. Therefore, our findings bear robustness with respect to the definition of reciprocal subject.

**Supplementary Table S5**

Brain regions differentially or commonly related to two types of indirect reciprocity: without Rand subjects

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **(CC–DD) pay-it-forward > (CC–DD) reputation-based** | | | | |
| Rt. AI | 36 | 30 | 2 | 4.9 |
| post dmPFC | 0 | 26 | 50 | 4.3 |
| **(CC–DD) reputation-based > (CC–DD) pay-it-forward** | | | | |
| ventral precuneus | -2 | -60 | 24 | 4.4 |
| dorsal precuneus | 2 | -46 | 64 | 4.9 |
| **(CC–DD) pay-it-forward & (CC–DD) reputation-based** | | | | |
| Lt. caudate | -8 | 0 | 14 | $P < 10^{-8}$ |
| ant dmPFC | 0 | 48 | 16 | |

See Tables S2 and S3 for abbreviations.

Analysis with the tendency of reciprocal behavior as a covariate.

To accentuate the results shown in Table S3, we performed the GLM analysis to search for CC–DD activity by incorporating the tendency of reciprocal behavior as a covariate. The tendency was defined as $p + (1–q)$, where the distributions of $p$ and $q$ are shown in Fig. S3. We obtained the qualitatively same results as the original ones obtained without the new covariate (Supplementary Table S6).

**Supplementary Table S6**

Brain regions differentially or commonly related to two types of indirect reciprocity: with the tendency of reciprocal behavior as a covariate

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **(CC–DD) pay-it-forward > (CC–DD) reputation-based** | | | | |
| Rt. AI | 34 | 32 | 2 | 5.5 |
| post dmPFC | -2 | 26 | 48 | 4.5 |
| **(CC–DD) reputation-based > (CC–DD) pay-it-forward** | | | | |
| ventral precuneus | -4 | -58 | 26 | 4.6 |
| dorsal precuneus | 0 | -52 | 64 | 5.3 |
| **(CC–DD) pay-it-forward & (CC–DD) reputation-based** | | | | |
| Lt. caudate | -10 | 0 | 14 | $P < 10^{-8}$ |
| ant dmPFC | -2 | 48 | 16 | |

$P_{\text{FWE}} < 0.05$. See Tables S2 and S3 for abbreviations.

**Supplementary Table S7**

Results of PPI analysis (CC – DD)

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **PPI from AI in pay-it-forward reciprocity** | | | | |
| Rt caudate | 14 | 4 | 6 | 4.8 |
| Lt caudate | -8 | 4 | 12 | 4.7 |
| Middle cingulate gyrus | 4 | -6 | 42 | 4.5 |
| **PPI from dorsal precuneus in reputation-based reciprocity** | | | | |
| Lt caudate | -10 | 2 | 16 | 4.9 |
| Rt caudate | 16 | 4 | 8 | 4.7 |
| Middle cingulate gyrus | -2 | 2 | 30 | 4.5 |

$P_{\mathrm{FWE}} < 0.05$. See Tables S2 and S3 for abbreviations.

**Supplementary Table S8**

Results of VBM

| | MNI coordinate | | | |
|---|---|---|---|---|
| Anatomical label | x | y | z | t value |
| **Correlation with *P*(C\|C) in pay-it-forward reciprocity** | | | | |
| Rt anterior insula | 40 | 12 | -4 | 4.9 |
| Lt inferior frontal gyrus | -56 | 22 | 10 | 4.5 |
| Rt inferior parietal lobule | 34 | -50 | 58 | 4.5 |
| **Correlation with *P*(C\|C) in reputation-based reciprocity** | | | | |
| Lt inferior parietal lobule | -46 | -46 | 38 | 4.8 |
| dorsal precuneus | -10 | -56 | 60 | 4.9 |
| Rt superior parietal lobule | 32 | -40 | 60 | 4.6 |

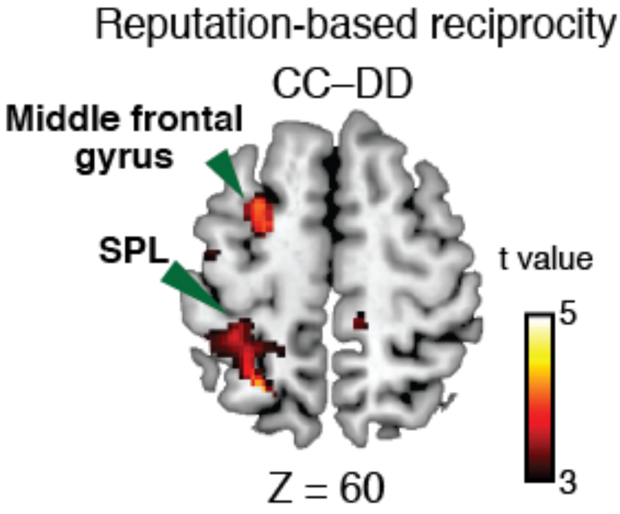
$P_{\mathrm{FWE}} < 0.05$. See Tables S2 and S3 for abbreviations.

Brain activity specific to positive reciprocity in non-reciprocal subjects

To compare with the brain activity of the reciprocal subjects, we examined the brain activity patterns of the subjects whose data were excluded in our original analysis.

First, we investigated whether or not the observed CC–DD activity in reputation-based reciprocity was specific to the reciprocal subjects. To this end, we estimated the CC–DD activity in reputation-based reciprocity in the non-reciprocal 12 subjects, but activations in the dorsal precuneus or dmPFC were insignificant even with a moderate statistical threshold ($P < 0.005$, uncorrected). Activations in the right AI during CC–DD condition in pay-it-forward reciprocity were also insignificant ($P < 0.005$, uncorrected). Instead, for these non-reciprocal subjects, we found mild activations in the middle frontal gyrus and left superior parietal lobule (SPL) in CC–DD in reputation-based reciprocity ($P < 0.001$, uncorrected; Fig. S6).

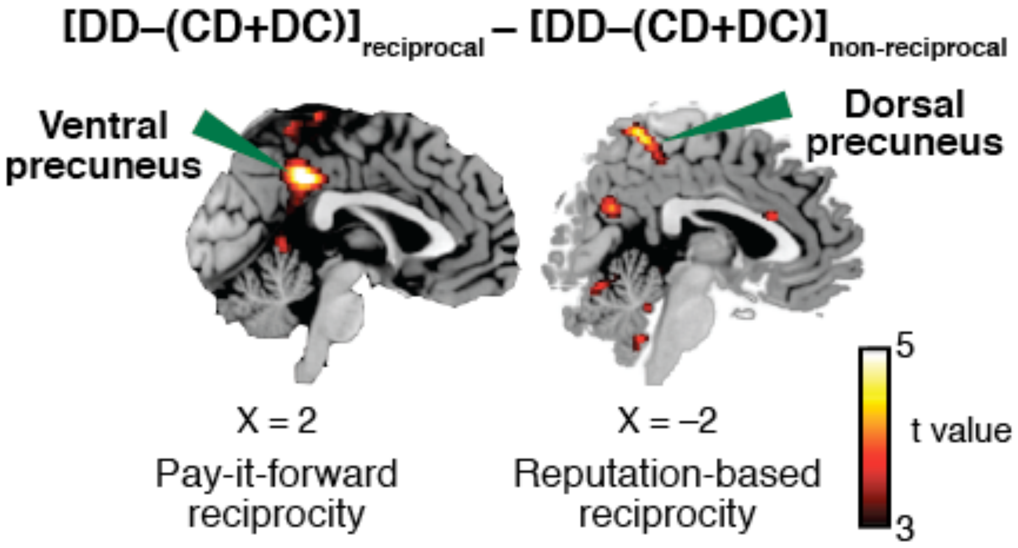
These results support that our main findings that the activations in the right AI and dorsal precuneus are specific to positive pay-it-forward reciprocity and positive reputation-based reciprocity, respectively.



**Supplementary Fig. S6. Brain activity in the non-reciprocal subjects during positive reputation-based reciprocity, defined as CC–DD.**

Difference in DD brain activity between reciprocal and non-reciprocal subjects

In the essentially same manner as the analysis of the CC brain activity (Fig. 3K), we compared the brain activity specific to DD (DD–(CD+DC)) between the 21 reciprocal and 12 non-reciprocal subjects. We found that, compared with the non-reciprocal subjects, the reciprocal subjects significantly recruited the ventral and dorsal precuneus during the pay-it-forward and reputation reciprocity games, respectively ($P_{\text{FWE}} < 0.05$; Fig. S7). The consistent activations observed in the precuneus area implies that a similar neural mechanism involving the precuneus is shared by positive (i.e., CC) and negative (i.e., DD) reputation-based reciprocity and negative pay-it-forward reciprocity, but not by positive pay-it-forward reciprocity.


[DD–(CD+DC)]reciprocal – [DD–(CD+DC)]non-reciprocal
Ventral precuneus
Dorsal precuneus
5
t value
3
X = 2
Pay-it-forward reciprocity
X = –2
Reputation-based reciprocity

**Supplementary Fig. S7. Difference in DD brain activity between the reciprocal and non-reciprocal subjects.**

## Supplementary Experimental Procedures

### Overall experimental design

The study consisted of two experiments: a group experiment and an fMRI experiment (Fig. 1A and 1B). Before the fMRI experiment, we collected a series of behavioral patterns in the group experiment. Based on the behavioral data, we prepared a series of stimuli used in the fMRI experiment. Written informed consent was obtained from all subjects in the group and fMRI experiments. All experiments complied with the requirements of the Declaration of Helsinki.

### Pay-it-forward reciprocity game

Figure S1A shows a part of a chain of subjects participating in the pay-it-forward reciprocity game. For example, subject X is first given 75 JPY (ca. £0.5 or $0.75) and presented with the previous decision of the upstream neighbor in the chain, i.e., subject W. Then, subject X is urged to make a decision toward the downstream neighbor, i.e., subject Y. Subject X selects either donation (i.e., cooperation) or non-donation (i.e., defection). If subject X decides to donate, subject X loses the 75 JPY given before the game, and the reward for the recipient of the action, i.e., subject Y, increases by 150 JPY. If subject X decides not to donate, the reward does not change for any player such that subject X keeps the 75 JPY (Fig. 1E). Theoretically speaking, in the Nash equilibrium of the single game, subject X is tempted to defect to keep 75 JPY unless an incentive to cooperate is presented.

After subject X has submitted a decision, subject Y is informed of the decision of subject X and makes a decision, i.e., to donate or not to donate to the downstream neighbor, i.e., subject Z. Then, subject Z makes a decision, and so forth. A sequence of decisions flows along the arrows in the chain network shown in Fig. S1A. Note that this experimental design does not allow the subject to specifically reciprocate or retaliate against other subjects that have made decisions before. For example, subject X knows beforehand that his or her decision will be revealed to the downstream neighbor subject Y, who directly receives the subject X's action. However, subject Y cannot affect subject X because the subsequent decision of subject Y is directed toward someone else (i.e., subject Z). The first subject is required to submit two decisions corresponding to the hypothetical situations in which the subject just observed C and D, because there is no upstream neighbor for this subject.

### Reputation-based reciprocity game

Figure S1B shows a part of a chain of subjects participating in the reputation-based reciprocity game. Subject X, for example, in the middle of the chain is first presented with the decision of the downstream neighbor, i.e., subject W, toward the downstream neighbor of subject W, i.e., subject V. The possible actions and their consequences in terms of the reward are the same as those for the pay-it-forward reciprocity game (Fig. 1E). Subject X may be tempted to donate to subject W if and only if subject W donated to subject V in the previous game, indicating that subject W has a good reputation. Such a mechanism for sustaining cooperation is called the image scoring [1, 2]. Cooperation under image scoring theoretically requires some intricate conditions to be satisfied, and more complex reputation-assignment rules facilitate reputation-based cooperation far more easily [3, 4]. However, in different experiments, humans have shown reputation-based cooperation that is consistent with image scoring [1, 2, 5-7]. After subject X has submitted a decision, the decision is revealed to the upstream neighbor, i.e., subject Y. This information serves as the reputation of subject X. Then, subject Y makes a decision toward subject X, subject Z makes a decision toward subject Y, and so forth. The first subject is required to make two decisions corresponding to a good (i.e., C) and bad (i.e., D) reputation of the hypothetical downstream neighbor. It should be noted that, before making a decision of whether to donate, each subject was informed of the decision of the previous subject, to whom the decision is directed.

Group experiment

Forty healthy Japanese university students (9 males, 31 females) were allocated to four groups of 10 subjects. Each group separately played four sessions of just one of the two types of reciprocity game: pay-it-forward and reputation-based reciprocity games (Fig. 1A and 1B, Fig. S1).

In each session, a subject decided 10 times whether or not to donate. Nine out of the 10 decisions were directed to any of the other nine subjects in the group experiment. The remaining decision was directed to a subject in the fMRI experiment. The subjects sequentially made decisions in a pseudo-random order. To assure anonymity, the names of the subjects were displayed on the computer screen as randomly chosen pairs of letters. Therefore, the subjects could not identify the person to whom they donated, who donated to whom, and so on. They were also told that the same name never appeared on the computer screen so that the entire experiment mimicked a series of one-shot interactions in a larger group. The subjects made decisions on computers and were also told that the experimenters facing the subjects (i.e., some of the authors) were informed of neither their decisions nor the amount of reward that the subjects obtained.

The subjects in the group experiment were instructed that there was an 11th subject who was not participating simultaneously. In fact, the 11th subject participated in the fMRI experiment. We used the decisions made just before the 11th subject's turns as the stimuli presented in the following fMRI experiment. Because the 11th subject was also displayed as two-letter name on the computer screen, the subjects could not know whether the adjacent subject in a chain network was the 11th subject. Because the decisions of the subjects in the fMRI experiment were unknown before we conducted the fMRI experiment, the subjects in the group experiment that had to submit decisions right after the 11th subject were told to make two decisions. For example, in the pay-it-forward reciprocity game, the subjects answered both questions: "If XQ gives money to you, will you give money to JW?" and "If XQ does not give money to you, will you give money to JW?" The subjects were told that, when they made disparate decisions for the two cases, the next subject also needed to answer for both cases and that the same parallel questions would be repeated until a subject selected the same action for both cases. The subjects were also told that there was a small probability that their decisions would be conveyed to the next subjects incorrectly, and that this error could also terminate the sequence of parallel questions. In practice, the error was implemented only when disparate decisions were made so that the parallel questions were terminated. The exact value of the error probability was not informed to the subjects.

Participants in the fMRI experiment

Fifty healthy right-handed Japanese male subjects (19−30 years) participated in the fMRI experiment. None of the subjects participated in the group experiment. None of the subjects had a history of neuropsychiatric disorder. The ethics committee of the University of Tokyo Hospital approved the study. Written informed consent was obtained from all the subjects before the experiment. Behavioral responses of two subjects were not recorded owing to a technical problem. Therefore, we analyzed behavioral data obtained from the remaining 48 subjects.

MRI acquisition

We used a 3T magnetic resonance imaging scanner (Discovery MR750w, GE, USA) with a 32-channel head coil in the University of Tokyo Hospital. High-resolution T1-weighted images (IR-prepared fast spoiled-gradient-recalled acquisition; TR = 6.8ms; voxel size, 1×1×1 mm) were acquired for each subject. We also acquired T2*-weighted functional images in a gradient-echo echo-planar sequence (TR = 3s; TE = 35ms; flip angle = 80°; voxel size, 4×4×4 mm; 42 slices; interleaved scan). We discarded the first five T2*-weighed images in each session to remove the transient before an equilibrium of longitudinal magnetization was reached.

### Tasks and stimuli used in the fMRI experiment

Each subject participated in eight scanning sessions (i.e., runs), each of which required approximately 4 min. The subject played the pay-it-forward reciprocity game in four sessions and the reputation-based reciprocity game in the other four sessions. The subject played either type of game 18 times with different partners in a session to undergo a total of 72 games of each type. The order of the two types of sessions was counter-balanced across subjects. When the type of game was switched, the subjects were given a brief instruction about the rules of the next game before the new session began.

In a single game of either type, a subject was first presented with a cue image showing the relationship among the subject and their neighbors in a chain network (Cue in Fig. 1C and 1D). The same names of neighbors did not appear in different single games. The cue image continued for 1 s. Next, the subject was presented with a condition image for 3 s showing the previous action of a neighbor, i.e., either donation (cooperation) or non-donation (defection). Donation and non-donation were indicated by a red circle and cross, respectively (Condition in Fig. 1C and 1D). Third, the subject was presented with a decision image (Decision in Fig. 1C and 1D) and urged to select the action (i.e., donation or non-donation) towards the specified recipient. The image also contained a sentence saying that the next decision maker, whose names and positions in the chain network were shown in the image, would find out the subject's decision. The position of the recipient and the next decision maker differed between the pay-it-forward and reputation-based reciprocity games (Fig. 1A and 1B; also see Fig. S1). The subject was instructed to make a decision within 6 s by pressing one of two buttons. Fourth, a fixation image with a white cross in the centre of the screen was presented (Fixation in Fig. 1C and 1D), during which the subject was instructed to look at the white cross. The length of the fixation period was uniformly distributed between 2 and 3 s.

Before the fMRI experiment, an experimenter who was not responsible for paying the reward to the subject provided a thorough explanation to the subject. (see Supplementary Materials).

### Identification of subjects carrying out indirect reciprocity

We focused on the brain activity of subjects engaged in indirect reciprocity in chains of donation games. Therefore, we analyzed the MRI data obtained from subjects showing at least some reciprocal tendencies. If a subject cooperated and defected right after observing a neighbor's cooperation and defection, we judged that the subject had implemented reciprocal behavior. Otherwise, the subject was considered not to have behaved reciprocally. Not all of the 48 subjects showed reciprocal behavior; hence, we classified them as follows.

We model the subject as a so-called reactive strategist that determines the action (i.e., C or D) on the basis of the two conditional probabilities. We denote by $p$ and $q$ the conditional probabilities to select C and D after observing C in the last game, respectively.

The log likelihood for the reactive strategy is given by

$$\log L = n_{\mathrm{CC}} \log p + n_{\mathrm{CD}} \log(1-p) + n_{DC} \log q + n_{\mathrm{DD}} \log(1-q),$$

where $n_{\mathrm{CC}}$, $n_{\mathrm{CD}}$, $n_{\mathrm{DC}}$, and $n_{\mathrm{DD}}$ are the numbers of conditional actions that the subject selected in the experiment (Fig. 1F). The first subscript (i.e., C or D) represents the action that the subject observes in the last game. In the case of the pay-it-forward reciprocity game, this is the action of the upstream neighbor (subject W in Fig. S1A) towards the subject in the last game. In the case of the reputation-based reciprocity game, it is the action of the downstream neighbor (subject W in Fig. S1B) towards his/her downstream neighbor (subject V in Fig. S1B) in the last game. The second subscript represents the action selected by the focal subject. Note that, for each of the pay-it-forward and reputation-based reciprocity games, $n_{\mathrm{CC}} + n_{\mathrm{CD}} + n_{\mathrm{DC}} + n_{\mathrm{DD}}$ was equal to 36 for each subject in the group experiment and 72 in the fMRI experiment.

The maximum likelihood estimators are simply given by $\hat{p} = n_{\mathrm{CC}}/(n_{\mathrm{CC}} + n_{\mathrm{CD}})$ and $\hat{q} = n_{\mathrm{DC}}/(n_{\mathrm{DC}} + n_{\mathrm{DD}})$ (Fig. S3). We classified the subjects into the following nine categories: unconditional cooperator (AllC), unconditional defector (AllD), random player (Rand), tit-for-tat (TFT), generous tit-for-tat (GTFT), Miser, Anti-TFT, Anti-GTFT, and Anti-Miser (Fig. S2).

AllC, AllD, and Rand are unconditional on the result of the last game that the subject observes (i.e., $p = q$). They are defined by $p = q = 1 - \epsilon$ , $p = q = \epsilon$ , and $p = q = 0.5$, respectively, where $\epsilon$ is the probability of error in selecting the action. TFT is defined by $p = 1 - \epsilon$ and $q = \epsilon$ . Although TFT is usually used in the context of direct reciprocity (i.e., repeated games between the same pair of players) [8, 9], here we use it to refer to the corresponding strategy in the two indirect reciprocity games without ambiguity. In reputation-based reciprocity, the TFT is usually referred to as discriminator [10, 11].

In direct reciprocity, GTFT, a more generous variant of TFT, is a strong competitor in evolutionary dynamics and is defined by $p = 1$ and $q = \frac{c}{b}$ [11-13]. Because $b = 150$ JPY and $c = 75$ JPY in our experiments, we refer to the reactive strategy $(p,q) = (1-\epsilon, 0.5)$ as GTFT. Similarly, the Miser strategy, originally defined by $(p,q) = (0.5,\epsilon)$ [14], is defined in our case by $(p,q) = (0.5,\epsilon)$ .

We refer to $(p,q) = (\epsilon, 0.5)$ , $(\epsilon, 1-\epsilon)$, and $(0.5, 1-\epsilon)$ as the Anti-Miser, Anti-TFT, and Anti-GTFT strategies, respectively.

We classified each subject into one of the nine strategies by selecting the strategy that minimizes the minimum description length (MDL) for the optimal $\epsilon$ value. The MDL is given by $-\log L + \frac{k}{2}\log(n_{\mathrm{CC}} + n_{\mathrm{CD}} + n_{\mathrm{DC}} + n_{\mathrm{DD}})$ , where $k = 0$ for Rand, which is parameter-free, and $k = 1$ for the other eight strategies, which have $\epsilon$ as the single free parameter. For each of the eight strategies, we allowed $\epsilon$ to vary to minimize the MDL.

We analyzed the subjects classified as Rand, TFT, GTFT, or Miser, which we call the reciprocating subjects. These four strategies are the closest to TFT among the nine strategies and are considered to comply with indirect reciprocity, at least to some extent. We included Rand and Miser in reciprocal behavior, because subjects showing intermediate behavior like our Rand and Miser have often been grouped together with more cooperative strategists [15].

In the group experiment, there were 8 and 12 reciprocating subjects out of 40 subjects in the pay-it-forward and reputation-based reciprocity games, respectively. In the fMRI experiment, there were 22 and 37 reciprocating subjects out of 48 subjects in the pay-it-forward and reputation-based reciprocity games, respectively. All of the reciprocating subjects in the pay-it-forward reciprocity game were also reciprocating subjects in the reputation-based reciprocity game in the fMRI experiment. Because the MRI data recorded from one of the 22 reciprocating subjects in the pay-it-forward reciprocity game were lost owing to technical problems in transferring the data, we subjected the data recorded from the remaining 21 reciprocating subjects identified in the pay-it-forward reciprocity game to the imaging analysis.

Some subjects yielded low likelihood values because they were not close to any of the nine strategies. However, the aim of our classification was to identify subjects that showed indirect reciprocity at least to some extent, but not to accurately model the behavior of the subjects. In particular, players with relatively large $p$ and small $q$ were classified as reciprocating subjects regardless of the bare likelihood values. In fact, such subjects behaved fairly consistently with indirect reciprocity; they tended to cooperate and defect after observing cooperation and defection in the last game, respectively.

Number of reciprocal behaviors and reaction time

To compare the fractions of CC and DD in the different types of games, we performed a repeated-measures two-way ANOVA of the fraction of the reciprocal behavior (type of action [CC and DD] × type of game [pay-it-forward and reputation-based]). We also compared the reaction times for CC and DD by using another ANOVA with the same structure.

fMRI analysis: task-related images

We preprocessed fMRI images recorded during the pay-it-forward and reputation-based reciprocity games by using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). The fMRI images were realigned, corrected for slice timing, normalized against the standard template image (ICBM 152), temporally filtered (high-pass filter: 128 s), and spatially smoothed (full width at half maximum, FWHM = 8 mm).

For each reciprocal subject, we analyzed the preprocessed images by using a general linear model in a standard event-related design. The general linear model consists of three regressors for each type of reciprocity: one for CC, another for DD, and the other for the mergence of CD and DC. Because the numbers of the CD and DC were considerably smaller than those of CC and DD in some subjects, we combined CD and DC. The onsets for these regressors were set at the time when each conditioning image appeared. The response time to each stimulus for each subject was also used as the duration time in each regressor. In addition to these regressors of interest, we added run-effect regressors and six motion-parameter regressors.

At a group level, we used the random effects model to analyze the fMRI images that were subjected to analysis at the single-subject level. We first conducted a repeated-measures two-way ANOVA of the fMRI signals (type of action [CC and DD] × type of game [pay-it-forward and reputation-based]). We then conducted post-hoc paired *t* tests of CC-specific activity (i.e., CC − DD) between the two types of indirect reciprocity across subjects. The ANOVA adopted FWE-corrected $P < 0.05$ as a statistical threshold, whereas the post-hoc *t* tests adopted $P < 0.05$ that was Bonferroni-corrected among the number of brain regions detected by the ANOVA.

To find the brain regions involved in both types of indirect reciprocity, we also conducted a conjunction analysis [16-18]. On the basis of the standard procedure of conjunction analysis with a conservative null hypothesis, we first estimated a whole-brain statistical map for CC-specific activity in each type of reciprocity (i.e., map showing CC − DD during the pay-it-forward reciprocity game and that during the reputation-based reciprocity game). We then binarized the two maps by thresholding the fMRI signals at one of three *P* values ($10^{-3}$, $5 \times 10^{-4}$, and $10^{-4}$, uncorrected). By multiplying the two binary maps obtained with the same threshold, we derived a whole-brain map indicating a conjunction of the two conditions (i.e., pay-if-forward and reputation-based) with three different *P* values (i.e., $10^{-6}$, $2.5 \times 10^{-7}$, and $10^{-8}$, uncorrected). This statistical threshold ($P < 10^{-6}$) is at least as conservative as that used in previous studies [17, 18].

ROI analysis and PPI analysis

We defined the six brain regions that we identified as CC-specific regions in either type of indirect reciprocity in the fMRI analysis (Fig. 2A, 2B and 2C, Table S2) as ROIs. We then extracted the CC-specific brain activity of each ROI by averaging the fMRI signals (so-called parameter estimates or beta values) in a 4-mm-radius sphere. The coordinates of the centre of the sphere are shown in Table S2.

For the 21 scanned subjects, we then calculated Pearson's correlation coefficient between brain activity and the probability of cooperation after the subject had observed cooperation (i.e., $P(C|C)$). For the four of the six ROIs specific to the type of indirect reciprocity (i.e., AI, posterior dmPFC, dorsal and ventral precuneus), we calculated $P(C|C)$ for each subject by using the behavioral results for the corresponding type of game. For the two ROIs found in the conjunction analysis (i.e., caudate and anterior dmPFC), $P(C|C)$ refers to the value calculated from the behavior in all the games.

This analysis left three out of six ROIs (i.e., AI, dorsal precuneus, and caudate) with high positive correlations. We then estimated task-related functional connectivity among these ROIs by calculating the PPIs by using the standard procedure implemented in SPM8 [19]. At the single-subject level, we first estimated CC-specific PPIs with each of the three ROIs as a seed for each type of game. The PPI analysis involved the following three regressors: the time course of the fMRI signals in the ROI, a regressor representing the psychological variable of interest (i.e., CC − DD), and a regressor representing the cross product of the previous two regressors. The regression coefficients for the third regressor are defined as the PPIs [19]. By conducting a one-sample *t* test of the results of the single-subject analysis, we calculated the PPI at a group level. Similarly, we estimated the PPIs for CC (i.e., CC − Fixation), that for DD (i.e., DD − fixation), and that for CD and DC (i.e., CD + DC − Fixation) for each type of game. For CC − DD, we also conducted exploratory whole-brain analysis of the PPI on the basis of a random effects model.

Resting-state functional connectivity (rsFC)

We obtained resting-state fMRI signals for approximately 10 min (5min/session × 2 sessions) after the subjects completed all the games. For each session and subject, the fMRI images were realigned, corrected for slice timing, and normalized against the standard template image (ICBM 152) by SPM8 [20]. For each session, we applied temporal band-pass filtering (0.01-0.1Hz) implemented by in-house MATLAB scripts and then spatially smoothed the images (FWHM = 8 mm) by SPM8. After combining the smoothed data across two sessions, we applied a general linear model to correct the images for the subjects' head motion, whole-brain signals, ventricular signals, white matter signals, and run effects.

By calculating Pearson's correlation coefficient for the preprocessed data, we determined the rsFC between the AI and caudate and between the dorsal precuneus and caudate. The activity of each ROI was defined as an averaged value of the beta value in a 4-mm-radius sphere, the centre of which is shown in Table S2. Finally, we transformed the obtained Pearson's correlation coefficient to a *Z* value by applying Fisher's transformation.

VBM analysis

To perform VBM analysis, we first preprocessed the high-resolution T1-weighed images in SPM8 as follows [21]: For each subject, the images were segmented into gray matter, white matter, and cerebrospinal fluids in the native space with the New Segment Toolbox [22]. The segmented gray and white matter images underwent alignment and warp to a template space. The images were then resampled down to 1.5-mm isotropic voxels. We registered the gray and white matter images to a subject-specific template by using the DARTEL Toolbox [23]. We normalized individual gray matter images to Montreal Neurological Institute (MNI) spaces by using the DARTEL Toolbox and smoothed the images with a Gaussian kernel (FWHM = 8mm).

We then analyzed the preprocessed gray matter images with a multiple regression model in SPM8. We searched for brain regions whose gray matter volume was correlated with *P*(C|C) in each type of game. To this end, we first conducted ROI-based VBM analysis. In this analysis, we regressed out the effect of the total brain volume of each individual and measured the regional gray matter volume of the AI and dorsal precuneus. For each subject, the gray matter volume was defined as the averaged gray matter volume in a 4-mm-radius sphere whose centre is shown in Table S2. Subsequently, we calculated Pearson's correlation coefficient between the gray matter volume of either ROI and *P*(C|C) in each type of game. We also performed an exploratory whole-brain VBM analysis.

**Supplementary References**

1. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960):785–791.
2. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437(7063):1291–1298.
3. Milinski M, Semmann D, Krambeck HJ (2002) Reputation helps solve the 'tragedy of the commons'. *Nature* 415(6870):424–426.
4. Bolton GE, Katok E, Ockenfels A (2005) Cooperation among strangers with limited information about reputation. *J Public Econ* 89(8):1457–1468.
5. Yamagishi T, Cook KS (1993) Generalized exchange and social dilemmas. *Soc Psychol Quaterly* 56(4): 235–248.
6. Dufwenberg M, Gneezy U, Guth W (2001) Direct versus indirect reciprocity: an experiment. *Homo Oeconomicus* 18:19–30.
7. Engelmann D, Fischbacher U (2009) Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ Behav* 67(2):399–407.
8. Boyd R, Richerson PJ (1989) The evolution of indirect reciprocity. *Soc Netw* 11(3):213–236.
9. Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393(6685): 573–577.
10. Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. *Proc Biol Sci* 268(1468):745–753.
11. Nowak, M. A. (2006). Evolutionary Dynamics (Harvard University Press, Boston, MA).
12. Axelrod, R. (1984). The Evolution of Cooperation (Basic Books, New York, NY).
13. Nowak MA, Sigmund K (1992) Tit for tat in heterogeneous populations. *Nature* 355(6357):250–253.
14. Frean MR. (1994) The prisoner's dilemma without synchrony. *Proc Biol Sci* 257(1348):75–79.
15. McCabe K, Houser D, Ryan L, Smith V, Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci USA* 98(20):11832–11835.
16. Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. *NeuroImage* 25(3):653–660.
17. Klasen M, Kenworthy CA, Mathiak KA, Kircher TTJ, Mathiak K (2011) Supramodal Representation of Emotions. *J Neurosci* 31(38):13635–13643.
18. Watanabe T, Yahata N, Kawakubo Y, et al. (2013) Network structure underlying resolution of conflicting non-verbal and verbal social information. *Soc Cogn Affect Neurosci* doi 10.1093/scan/nst046
19. Friston KJ, Buechel C, Fink GR, Morris J, Rolls ET, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* 6(3):218–229.
20. Fox MD, Snyder AZ, Vincent JL, Corbetta M, van Essen DC, Raichle ME (2005) The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci USA* 102(27):9673–9678.
21. Ashburner J, Friston KJ. Voxel-based morphometry--the methods. (2001) *NeuroImage* 11(6 Pt 1):805–821.
22. Ashburner J, Friston KJ Unified segmentation. (2005) *NeuroImage* 26(3):839–851.
23. Ashburner J A fast diffeomorphic image registration algorithm. (2007) *NeuroImage* 38(1):95–113.

**Supplementary Materials**

The following pages contain slides used for instructing the subjects on the procedure used in the fMRI experiment. The original slides were written in Japanese, and the authors translated them into English.