



Universiteit
Leiden
The Netherlands

The neurocognitive development of social decision-making

Bos, W. van den

Citation

Bos, W. van den. (2011, April 12). *The neurocognitive development of social decision-making*. Retrieved from <https://hdl.handle.net/1887/16711>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/16711>

Note: To cite this publication please use the final published version (if applicable).

3. What motivates repayment?

Neural correlates of reciprocity in the Trust Game

Reciprocity of trust is important for social interaction and depends on individual differences in social value orientation (SVO). Here, we examined the neural correlates of reciprocity by manipulating two factors that influence reciprocal behavior: (1) the risk that the trustor took when trusting and (2) the benefit for the trustee when being trusted. fMRI results showed that anterior Medial Prefrontal Cortex (amPFC) was more active when participants defected relative to when participants reciprocated, but was not sensitive to manipulations of risk and benefit or individual differences in SVO. However, activation in the right temporal parietal-junction (rTPJ), bilateral anterior insula and anterior cingulate cortex (ACC) was modulated by individual differences in SVO. In addition, these regions were differentially sensitive to manipulations of risk for the trustor when reciprocating. In contrast, the ACC and the right dorsolateral prefrontal cortex were sensitive to the benefit for the trustee when reciprocating. Together, the results of this study provide more insight in how several brain regions work together when individuals reciprocate trust, by showing how these regions are differentially sensitive to reciprocity motives and perspective-taking.

3.1 Introduction

One of the key components of human social interaction is cooperation or the exchange of favor or goods between individuals for the attainment of mutual benefit. Cooperation depends to a large extent on trust and reciprocity. Trust is required because cooperative exchanges are often separated in time, whereas reciprocity, or the repayment of what others have provided us, is thought to be important for the maintenance of social relationships. That is, if favors are not returned relationships may be short-lived (Lahno, 1995).

Both the trustor and the trustee may obtain higher outcomes when trust is given relative to when no trust is given. However, trusting also involves a component of risk, because the trustor may attain higher personal benefit when not reciprocating. Consequently, trusting may result in a smaller outcome for

the trustor relative to when the trustor would not have trusted (Rousseau et al., 1998). Thus, the decision to trust another party involves risk for the trustor and the decision to reciprocate trust depend on the offset between maximizing personal outcomes relative to the appreciation of the trust that was given (i.e. repayment). This study will focus on different motives involved in reciprocal behavior.

Researchers have demonstrated that even for single anonymous transactions, individuals often reciprocate trust even when this leads to a smaller personal monetary outcome (Berg et al., 1995; McCabe et al., 2001). It has therefore been suggested that our motivation to reciprocate trust is not only guided by goals to maximize personal outcomes, but also by other-regarding preferences (Falk and Fischbacher, 2006; Fehr and Camerer, 2007; Fehr and Gintis, 2007; Van Lange, 1999). According to these studies, the decision to reciprocate is dependent on evaluating consequences for both self and others. Importantly, reciprocal behavior is dependent on individual differences in social value orientation (SVO), the general tendency of individuals to value the outcome of others (McClintock and Allison, 1989; De Dreu and Van Lange, 1995; Van Lange et al., 1997). Furthermore, decisions to reciprocate trust are not only motivated by outcome considerations but also involve considerations of the intentions of others, such as the risk that the trusting party took when trusting or the benefit for the trusted party when being trusted. Therefore, these decisions are thought to be dependent on our ability to take the perspectives of others.

Neuroimaging studies in combination with game theoretical paradigms have investigated the neural correlates of the cognitive processes involved in cooperation and reciprocal exchange (e.g. King-Casas et al., 2005; Krueger et al., 2007; McCabe et al., 2001; Rilling et al., 2002). Several of these neuroimaging studies have reported activation in the anterior medial prefrontal cortex (aMPFC) when participants are involved in interactions with another person relative to a computer (McCabe et al., 2001; Rilling et al., 2004), and when participants decide to trust relative to when they decide not to trust (McCabe et al., 2001; Delgado et al., 2005; King-Casas et al., 2005; Krueger et al., 2007; Baumgartner et al., 2008). Prior neuroimaging studies have considered the aMPFC together with the temporal-parietal-junction (TPJ) to be important for mentalizing and theory-of-mind. For example, neuroimaging studies have demonstrated that aMPFC and TPJ are active during theory-of-mind tasks, such as tasks that require participants to infer mental states of characters in stories (Fletcher et al., 1995) and cartoons (Gallagher et al., 2002) or while watching animations (Castelli et al., 2000). In addition, prior studies have suggested that in a social context the aMPFC is involved in evaluating the

mental content of others in relation to the self (Amodio and Frith, 2006), whereas the TPJ is thought to be important for redirecting or focusing attention on the other (Mitchell, 2008). However, the mentalizing requirements during these theory-of-mind tasks are complex, and therefore it is difficult to dissociate the putative roles of the aMPFC and TPJ in social interaction (Hampton et al., 2008). Therefore, it remains to be determined how activation in aMPFC and TPJ can be associated with the different processes, which may underlie reciprocal exchange.

Besides the aMPFC and TPJ, neuroimaging studies of social decision-making have also suggested that brain regions that are associated with reward processing and arousal can mark social interactions as positive or aversive. For example, one neuroimaging study demonstrated that activation in the ventral striatum correlates positively with cooperation choices in a Prisoners Dilemma Game (Rilling et al., 2004). Two other neuroimaging studies showed that unfair treatment by a partner in the Ultimatum Game results in increased activation in the insula (Sanfey et al., 2003; Tabibnia et al., 2008), and this region has also been engaged during unreciprocated trust (Rilling et al., 2008). A recent study, which examined iterated two-person trust exchanges, demonstrated that the insula is more active for low relative to high levels of reciprocity. This finding was explained by suggesting a role of the insula in signaling personal norm violations (King-Casas et al., 2008). Thus, the ventral striatum and the insula seem to be involved in the pleasant and unpleasant aspects of social interactions, which may explain how lower level affective processes can result in encouragement or discouragement of social behavior (Sanfey, 2007). However, even though this pattern of activity is consistent over a wide range of social interactions paradigms, it has not been shown how these regions are associated with the choice and motivation to reciprocate.

Finally, the anterior cingulate cortex (ACC) and the right dorsolateral prefrontal cortex (rDLPFC) are typically engaged when individuals make decisions in which there is conflict between social norms and personal interest (Sanfey et al., 2003; Spitzer et al., 2007) or when individuals make decisions that may be counter to their own response tendencies (Rilling et al., 2002, 2007). In addition, transcranial magnetic stimulation of the right DLPFC lead to an increase of accepting unfair offers in the Ultimatum Game (Knoch et al., 2006). These control-related structures may therefore be involved in overriding self-oriented impulses.

Neuroimaging methods may allow us to examine the possible dissociations between different processes that underlie an individual's decision to reciprocate. Indeed, the review of prior neuroimaging studies suggests that the brain regions, which have been reported in social interaction studies, may indeed contribute in

different ways to different motives for reciprocity. However, to date, most neuroimaging studies of social interaction have examined the neural correlates of different types of choices (e.g. reciprocate vs defect) but have not attempted to dissociate between processes that may underlie the decision to reciprocate or defect, such as the risk that the trusting party took or the benefit the trusted party gained by being trusted. Therefore, the question remains how the brain regions, which have previously been associated with lower-level cognitive and affective processes and have been suggested to be involved in social interaction, are differentially involved in reciprocal behavior. This question can be addressed by investigating how these brain regions are differentially sensitive to the putative motives for reciprocity, which have been outlined above. In this study, we will manipulate the risk for the trustor and the benefit for the trustee, and we will examine the effects of these manipulations on the neural correlates of reciprocal behavior under these conditions. Thus, the goal of the current study was to determine whether the appreciation of different motives for reciprocity can be dissociated on a neural level by manipulating the risk that the trustor took when trusting and the benefit for the trustee when being trusted.

Participants played several one-shot rounds of the Trust Game, in which they had to make the decision whether or not to reciprocate trust given by another individual (Berg et al., 1995). In the Trust Game, two anonymous players are involved in dividing a certain amount of money. The first player (trustor) has two options. One option is to divide the money according to a predetermined scheme (e.g. eight for first player and seven for second player; see Figure 3.1 A), the other option is to trust the second player (trustee) and to give him/her the choice to divide the money. The latter option potentially leads to a higher pay-off for both players. If trusted, the second player has two options: (1) reciprocate the trust given by the first player (e.g. 11 for first player and 10 for second player) or (2) defect and maximize personal gains (e.g. 5 for first player and 17 for second player). All participants were assigned to the role of the second player and always had two fixed choices. This design allowed us to (a) concentrate on the decision to reciprocate or not and (b) systematically vary the main variables of interest: the risk for the trustor and the benefit for the trustee.

We predicted that the extent to which second players are motivated to reciprocate depends on the risk that the first player has taken (i.e. the amount of money the first player can lose by trusting) and the benefit that the second player receives when being trusted (i.e. the amount of money that the second mover receives when trusted relative to not being trusted) (Pillutla et al., 2003; Malhotra, 2004; van den Bos et al., 2010). More specifically, we expected that

participants were more motivated to reciprocate when either the risk or the benefit was high rather than low.

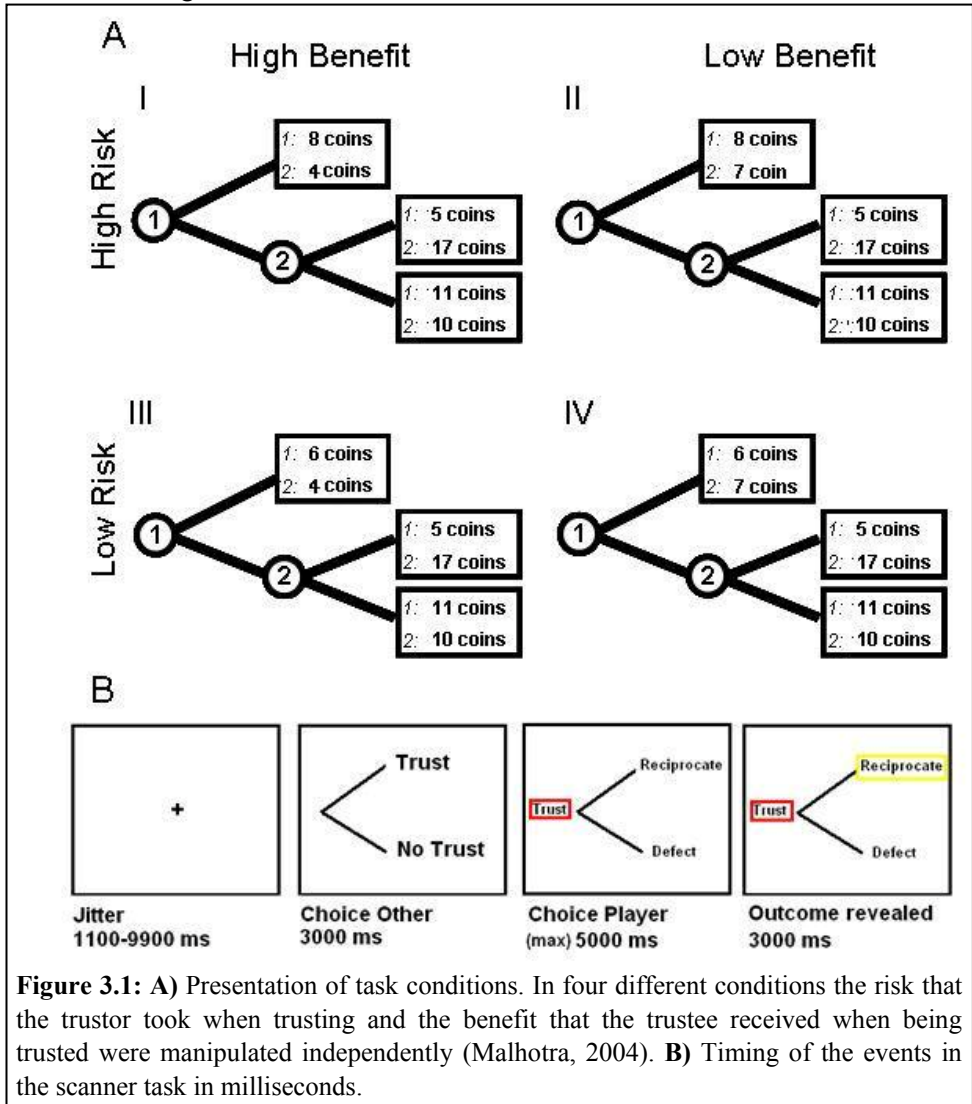


Figure 3.1: A) Presentation of task conditions. In four different conditions the risk that the trustor took when trusting and the benefit that the trustee received when being trusted were manipulated independently (Malhotra, 2004). B) Timing of the events in the scanner task in milliseconds.

We hypothesized that regions that are involved in mentalizing would be modulated by both risk and benefit manipulations. However, we expected that the type of perspective-taking would be associated with distinct neural correlates. In particular, we posited that regions that are important for taking the perspective of the other would be especially sensitive to the risk manipulation because the risk manipulation requires participants to take into account the outcomes of the other (first) player. Thus, the risk manipulation focused on neural correlates of mentalizing about how the different outcomes affect the

first player. In contrast, we posited that regions, which are associated with self-referential thought, would be sensitive to the benefit manipulation, because the benefit manipulation involves taking into account the second player's own increased outcome in case of trust. Thus, the benefit manipulation focused on neural correlates of mentalizing about the cooperative intentions of the first player, which benefits the second player.

We predicted that aMPFC and TPJ would exhibit a pattern consistent with their suggested roles in perspective-taking. In particular, we expected that the risk manipulation, motivating participants to take the perspective of the outcomes for the other, would result in a shift in attention from self to the other and thus would be associated with changes in TPJ activity (Lamm et al., 2007). On the other hand, we expected that the aMPFC would be more engaged by the benefit manipulation, because this manipulation motivated the participants to consider their own outcomes and the cooperative intentions of others (McCabe et al., 2001; Gallagher et al., 2002; Hampton et al., 2008).

We expected that the ACC and rDLPFC would also be sensitive to risk and benefit manipulations and would exhibit a pattern consistent with a role in overcoming selfish impulses (Rilling et al., 2002, 2007; Knoch et al., 2006). Therefore, we expected that these regions were most engaged when the participants reciprocated in situations where the incentive to reciprocate was low (low-benefit condition). Finally, we predicted that the insula would be sensitive to situations, which involved violations of one's own behavioral norms (Montague and Lorenz, 2007; King-Casas et al., 2008). Therefore, we expected a pattern of activation partly overlapping with activation observed in ACC and rDLPFC. In the insula, we expected increased activation when reciprocating in both low-benefit and low-risk conditions.

Finally, we expected that the need and/or engagement of the affective and control regions would also be dependent on the internal motivations to reciprocate. As such, the individual differences in reciprocal behavior in the current task were related to scores on the SVO questionnaire (van Lange, 1999), which is a personality variable that indicates how people evaluate outcomes for themselves and others. This questionnaire has shown significant external validity in a variety of settings (McClintock and Allison, 1989; De Dreu and van Lange, 1995; van Lange et al., 1997). Prosocial personalities were expected to reciprocate more often than the prosself personalities (Kramer et al., 1986). We posited that the activity in regions, which are associated with affective processes, would also correlate with individual differences in SVO. The insula and striatum were predicted to be sensitive to individual predispositions to reciprocate or defect reflecting differences in social norms and preferences. By the same token, we expected that prosocial participants would show less activity

in the control network (DLPFC, ACC) when reciprocating than the proself individuals and that proself participants would show more activation in the control network when reciprocating.

3.2 Methods

3.2.1 Participants

Twenty-two healthy right-handed paid volunteers (11 female, 11 male; age 18–22, $M = 19.7$, $SD. = 1.3$) participated in the fMRI experiment. Four of the participants were excluded from the analysis, because there were missing cases in one or more conditions (i.e. only reciprocal choices or only defect choices, see supplementary data). Subsequent fMRI analyses were based on the remaining 18 participants (nine female, nine male; age 18–22, $M = 19.7$, $SD. = 1.4$). All participants reported normal or corrected-to-normal vision and an absence of neurological or psychiatric impairments. All participants gave informed consent for the study, and all procedures were approved by the Leiden University Department of Psychology and the medical ethical committee of the Leiden University Medical Center. In accordance with Leiden University Medical Center policy, all anatomical scans were reviewed by the radiology department following each scan. No anomalous findings were reported. Standard intelligence scores were obtained from each participant using the Raven's Progressive Matrices test. All participants had average or above average IQ scores ($M = 116.12$, $SE = 1.98$).

3.2.2 Task

Trust Game. During the fixed choice Trust Game (Berg et al., 1995; Malhotra, 2004), participants were instructed that in an earlier phase of the study, other individuals had been assigned the roles of first player and that they would complete the second phase of the study in the role of second player. They were instructed that they were not playing directly with first players, but that they played with the implementation of answers of first players which were gathered in the previous part of the experiment. They were explained that their decisions would have consequences for the first player and that the payment of all participants would take place after completion of the experiment.

Each round, participants were paired with a different, anonymous player to exclude reputation effects or strategy use, and the other players were matched for gender. For those trials where the first players had decided to trust, the participant was presented with two options: reciprocate or defect. If the participant decided to defect, the participant would maximize his/her own gains

and the first player would receive less money than in the no-trust option. In case the participant reciprocated, the money was shared almost equally and both players received more money compared to the no-trust option, but the second player received less money compared to when he/she would have defected (see Figure 3.1A). Participants were instructed that at the end of the experiment the computer would randomly select the outcome of five trials, and the sum of these trials would determine the pay-off for the participant and for the first players. Consequently, their decisions had implications for both their own pay-off as well as that of the other players.

Each trial started with a 3 s display of the choice alternative for the first player, followed by the trust or no-trust decision of the first player. For those trials on which the first player chose not to trust, the no-trust decision was visually presented for 3 s. For those trials on which the first player chose to trust, the defect and reciprocate options were presented, and participants were instructed to make their decision by pressing the middle or index finger of the right hand. Participants were instructed to respond within a 5 s window (see Figure 3.1B). The 5 s decision-display was followed by a 3 s display of their choice.

Risk for the trustor (high vs low) and benefit for the trustee (high vs low) were manipulated separately (Malhotra, 2004) (see Figure 3.1A). The risk manipulation determined the risk for the first player. In the high-risk condition, the first player could lose a large amount of money by trusting the participant in case the second player chose to defect. In contrast, in the low-risk condition, the first player could lose only a small amount of money by trusting the second player. The benefit manipulation determined the benefit for the second player when being trusted. In the low-benefit condition, the difference between money gained by player 2 when being trusted relative to not being trusted was small. In contrast, in the high-benefit condition, the increase of money for the second player by being trusted was large. The risk and benefit manipulations were based on the Malhotra (2004) paradigm.

The computer played a fixed strategy that was based on behavior of participants in previous studies (van den Bos, et al., 2010). In total, the task consisted of 43 high risk-high benefit trials (25 trusted, 18 not-trusted), 44 high risk-low benefit trials (23 trusted, 21 not trusted), 48 low risk-high benefit trials (35 trusted, 13 not-trusted) and 53 low risk-low benefit trials (42 trusted, 11 not-trusted). Consequently, for each participant, the task consisted of 188 rounds in total, with 125 trusted trials, which required a decision from the participant. The trials were divided over five blocks, each block lasted ~8.5 min. The trials were presented in pseudo-random order with a jittered

interstimulus interval (min. = 1.1s, max. = 9.9s, mean = 3.37s) optimized with OptSeq2 [surfer.nmr.mgh.harvard.edu/optseq/, developed by Dale (1999)].

Social Value Orientation. All participants completed the SVO questionnaire. The SVO is a brief measure of allocation choices between self and other and has shown significant external validity in a variety of settings. The questionnaire consists of nine tables or ‘decomposed games’ [for more details, see van Lange (1999)]. In these decomposed games, the participant determines the outcome for both himself and a hypothetical other.

The three different decompositions correspond to three different types of SVOs: (1) a cooperative orientation, reflecting a preference for joint outcomes, (2) an individualistic orientation, reflecting a preference for own outcomes and (3) a competitive orientation, reflecting a preference for a large positive difference between own and other outcomes. When participants make six or more consistent choices in nine games, they are classified as belonging to one of three types of SVO: cooperative, individualistic or competitive. In prior studies, cooperative participants have been categorized as a ‘prosocial’ group, and individualistic and competitive participants have been categorized as a ‘proself’ group. The reason for the latter categorization is based on the observation that both individualistic and competitive individuals value outcomes for self higher than outcomes for others (van Lange, 1999).

Task Procedure. Prior to the experiment, participants received oral instructions and completed a practice session (20 trials). The stimuli and timing of the practice sessions were the same as in the fMRI experiment. The Raven SPM and SVO questionnaire (Van Lange, 1999) were administered after the scanning session. The total duration of the experiment was ~2 h.

MRI Procedure. Data were acquired using a 3.0T Philips Achieva scanner at the Leiden University Medical Center. Stimuli were projected onto a screen located at the head of the scanner bore and viewed by participants by means of a mirror mounted to the head coil assembly. First, a localizer scan was obtained for each participant. Subsequently, T2*-weighted EPI (TR = 2.2 s, TE = 30 ms, 80x80 matrix, FOV = 220, 352.75-mm transverse slices with 0.28mm gap) were obtained during five functional runs of 232 volumes each. The first two scans were discarded to allow for equilibration of T1 saturation effects. A high resolution T1-weighted anatomical scan and a high resolution T2-weighted matched-bandwidth high-resolution anatomical scan (same slice prescription as EPI) were obtained from each participant after the functional runs. Stimulus

presentation and the timing of all stimuli and response events were acquired using E-Prime software.

fMRI Data Analysis. Data were preprocessed using SPM2 (Wellcome Department of Cognitive Neurology, London). The functional time series were realigned to compensate for small head movements. Translational movement parameters never exceeded 1 voxel (< 3 mm) in any direction for any subject or scan. Functional volumes were spatially smoothed using a 6mm full-width half-maximum Gaussian kernel. Functional volumes were spatially normalized to EPI templates. The normalization algorithm used a 12-parameter affine transformation together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 3mm cubic voxels. The MNI305 template was used for visualization and all results are reported in the MNI305 stereotaxic space (Cosoco et al., 1997), an approximation of Talairach space (Talairach and Tournoux, 1988).

Statistical analyses were performed on individual participants' data using the general linear model in SPM2. The fMRI time series data were modeled by a series of events convolved with a canonical hemodynamic response function (HRF). The start of the first player's choice display and the start of the second player's choice display (only for trust trials) of each trial were modeled as zero-duration events. The second player's choice display condition was divided in trust and no-trust choices and the trust choices were divided into reciprocate and defect decisions. Finally, those choices were further divided in four experimental conditions (high vs low risk, high vs low benefit). These trial functions were used as covariates in a general linear model, along with a basic set of cosine functions that highpass filtered the data and a covariate for run effects. The least-squares parameter estimates of height of the best-fitting canonical HRF for each condition were used in pairwise contrasts. The resulting contrast images, computed on a subject-by-subject basis, were submitted to group analyses. At the group level, contrasts between conditions were computed by performing one-tailed t-tests on these images, treating participants as a random effect. Mean reciprocity levels were used in regression analyses to test for brain-behavior relations. We applied AlphaSim (Ward, 2000) to calculate the appropriate threshold significance level and cluster size. A significance threshold of $p < 0.05$, corrected for multiple comparisons was calculated by performing 10 000 Monte Carlo simulations in AlphaSim resulting in an uncorrected threshold of $p < 0.001$, requiring a minimum of 12 voxels in a cluster.

Region-of-Interest (ROI) Analyses. ROI analyses were performed to further characterize sensitivity to risk and benefit manipulations. Averaging the signal across voxels, as is done in ROI analyses, captures the central tendency and tends to reduce uncorrelated variance. Thus, ROI analyses have greater power than whole-brain statistical contrasts to detect effects that are present across a set of voxels. ROI analyses were performed with the Marsbar toolbox in SPM2 (Brett et al., 2002; <http://marsbar.sourceforge.net/>).

The contrast used to generate functional ROIs based on a priori hypotheses was that of all choices > fixation, unless otherwise specified in the text. Functional maps were masked with anatomical masks from the Marsbar toolbox. For all ROI analyses, effects were considered significant at an α of 0.008, based on Bonferroni correction for multiple comparisons ($p = 0.05/0.06$ ROIs (aMPFC, rTPJ, rDLPFC, ACC, anterior insula and ventral striatum), unless reported otherwise. For each ROI, the center of mass is reported.

3.3 Results

3.3.1 Behavioral data

Trust Game. On average, participants reciprocated half of the trials ($M = 51\%$), but there were large individual differences in behavior ($SD = 18\%$, min. = 22%, max. = 78% see supplementary results). To investigate whether there were effects of the risk and benefit manipulations on reciprocity decisions, we performed a repeated measures ANOVA with risk (high vs. low) and benefit (high vs. low) as within-subject factors. As expected, high risk for the first player resulted in more reciprocal choices (59%) than low risk for the first player (43%) (main effect risk, $F(1,18) = 26.85$, $p < 0.001$) and high benefit for the second player resulted in more reciprocal choices (61%) than low benefit for the second player (40%) (main effect benefit $F(1,18) = 22.03$, $p < 0.001$). In addition, there was a significant risk x benefit interaction [$F(1,18) = 9.92$, $p < 0.01$]. This interaction demonstrated that the difference between high- and low benefit reciprocal choices was larger for low risk trials (high benefit: 58%, low benefit: 27%) than for high-risk trials (high benefit: 64%, low benefit: 53%). Thus, when the risk to trust was high for the first player, participants focused less on their own benefit when deciding to reciprocate. Finally, there were no differences in mean reaction times for defect ($M = 1.77$ s, $SE = 0.13$) vs. reciprocate ($M = 1.76$ s, $SE = .12$) choices [$t(21) = 0.044$, $P = 0.96$].

Social Value Orientation. Classification of participants by SVO (Van Lange, 1999) resulted in 8 prosself and 10 prosocial-oriented individuals. The SVO was a strong predictor of reciprocal behavior in the Trust Game as administered in

the scanner session. A t-test for reciprocity level demonstrated that prosocial individuals reciprocated significantly more ($M = 62\%$, $SD = 11\%$) than proself individuals [$M = 39\%$, $s.d. = 10\%$; $t(1,16) = 3.72$, $p < 0.002$]. When reciprocity levels in the Trust Game were divided based on a median split analysis, the low-reciprocity group consisted of all eight proself classified participants and one prosocial classified participant. The high-reciprocity group consisted of only prosocial classified participants. Thus, performance in the current version of the Trust Game had high external validity as demonstrated by a high correlation with SVO.

3.3.2 fMRI data

Whole Brain Results - Main effects. To examine the neural correlates of reciprocity, we examined neural activity for reciprocate and defect choices for those trials on which the participant was trusted. The comparison of defect choices $>$ reciprocate choices revealed activity in the aMPFC (BA 32; Figure 3.2A, Table 3.1) and the primary visual cortex (MNI 6,-93, 12), whereas the opposite contrast (reciprocate $>$ defect) resulted in significant activation only in primary visual cortex (MNI 9, -63, 12). It should be noted that defect and reciprocate alternatives were always displayed on the same location of the screen, which may explain the consistent activation in the visual areas for the separate contrasts.

Regression analysis. The second set of contrasts aimed at revealing individual differences in neural activation by adding average reciprocity level as a predictor variable to a regression analysis. This analysis revealed a positive correlation between levels of reciprocity and BOLD activity for defect $>$ reciprocate choices in the dorsal ACC, bilateral anterior insula, right TPJ (rTPJ) and precuneus (Figure 3.2B). Those individuals who generally showed prosocial behavior by reciprocating more often also showed increased activation in these areas when defecting. In contrast, those individuals who reciprocated less often showed more activation in these areas when reciprocating (see also supplementary results). Thus, these areas were sensitive to the less frequently chosen alternative, regardless of whether the less frequent alternative was to reciprocate or to defect.

There were no regions that showed a negative correlation between reciprocity and BOLD activation for defect $>$ reciprocate at a $p < 0.001$ threshold. However, lowering the threshold to an uncorrected threshold of $p < 0.05$ revealed a negative correlation between reciprocity and the defect $>$ reciprocate contrast in the ventral striatum. Here, individuals who reciprocated more often showed increased activation when reciprocating, and individuals

who reciprocated less often showed less activation when reciprocating (see supplementary results for performance correlations).

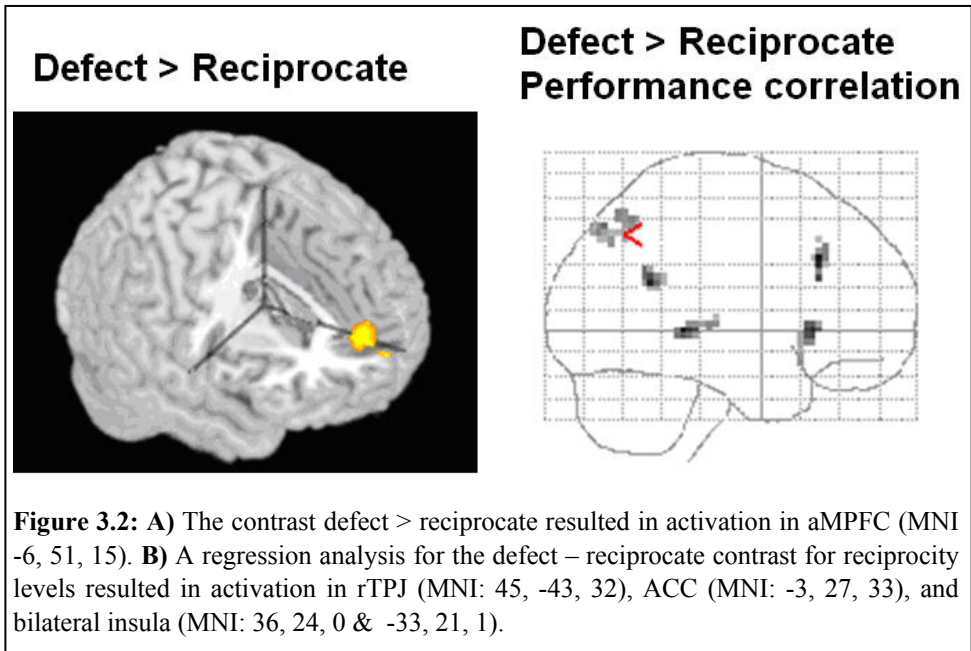


Table 3.1: Brain Regions revealed by whole brain contrasts and regressions analysis

Anatomical region	L/R	Volume (mm)	Z	MNI coordinates		
				x	y	z
Main effect of Choice						
Defect > Reciprocate						
Paracingulate cortex, VMPFC	L	666	5.84	-6	51	15
Visual Cortex	L/R	1006	6.06	6	-93	12
Reciprocate > Defect						
Visual Cortex	L/R	720	4.43	9	-63	3
Regression Defect > Reciprocity			Z			
Positive corr. avg. reciprocity						
Anterior Cingulate Cortex	L/R	917	4.10	-3	27	33
Anterior Insula	R	371	4.06	36	24	0
Anterior Insula	L	286	3.97	-33	21	1
Temporal Parietal Junction	R	862	4.06	45	-43	32
Precuneus	L	423	3.32	-24	-72	45
Thalamus	R	223	3.91	6	-30	0
Negative corr. avg. reciprocity**						
Ventral Striatum	R	171	1.63	14	12	-5

MNI coordinators for main effects, peak voxels reported at $p < .001$, at least 10 contiguous voxels.** peak voxel reported at $p < .05$

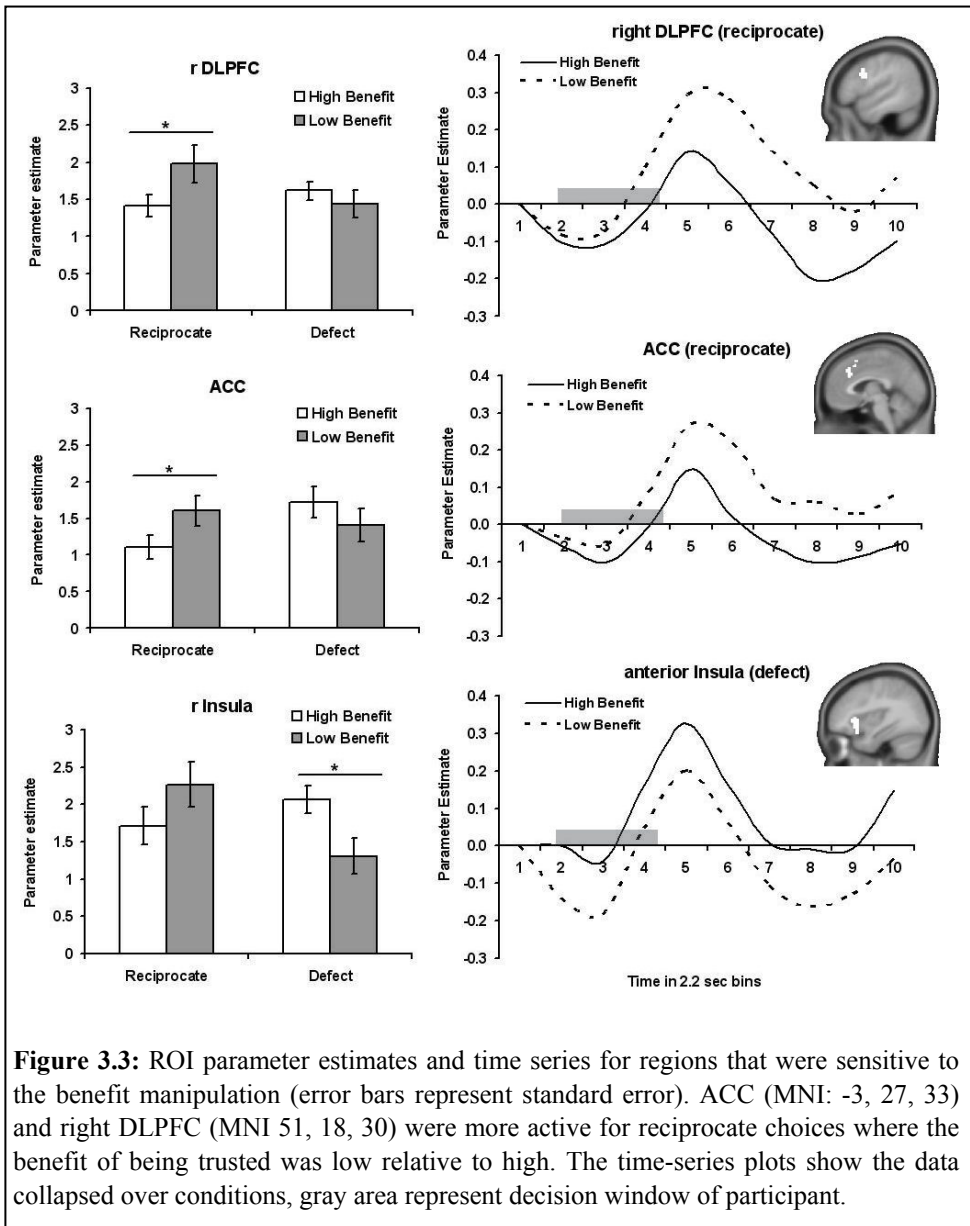
ROI analyses. ROI analyses were performed to further characterize sensitivity to risk and benefit manipulations. For these analyses, we focused on six a priori defined regions: aMPFC, rTPJ, rDLPFC, ACC, anterior insula and ventral striatum. rDLPFC, ACC and ventral striatum were derived from the all choices > fixation contrast. Not all regions were revealed by this contrast; therefore, aMPFC was selected based on the defect > reciprocate contrast, and the right TPJ and right insula were derived from the regression analyses.

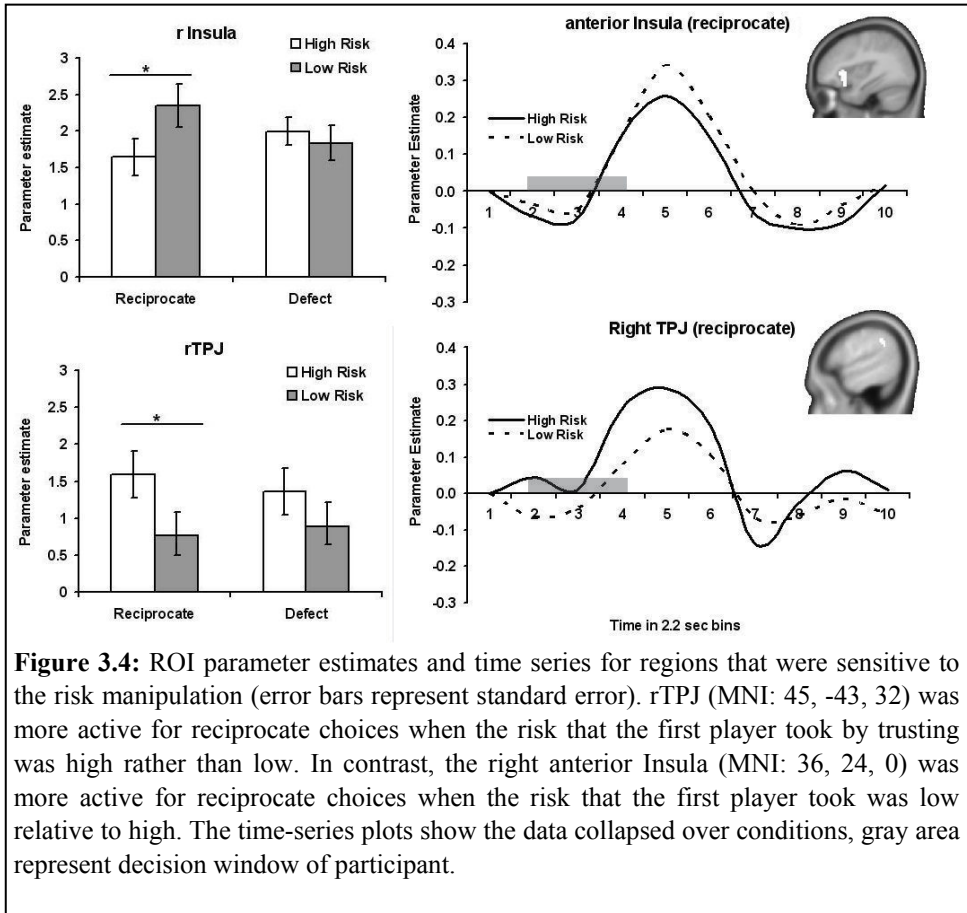
Because our hypotheses concerned the modulations of the neural correlates of reciprocal choices, we analyzed the effects of the risk and benefit manipulations for reciprocal choices. We used ANOVA to analyze BOLD differences that accompanied the choices to reciprocate and to characterize possible interactions with risk and benefit manipulations. These analyses revealed main effects of benefit in the ACC [$F(1, 17) = 5.46$, $p = 0.01$, Figure 3.3A] and the rDLPFC [$F(1, 17) = 9.98$, $p < 0.003$; Figure 3.3B]. These analyses demonstrated that there was greater activation in both the ACC and the rDLPFC when participants chose to reciprocate when the benefit for themselves was low relative to when the benefit for themselves was high. Thus, ACC and rDLPFC were more active when participants decided to reciprocate, even though the benefit of being trusted was low.

There was also a main effect of risk in the right TPJ [$F(1, 17) = 6.43$, $P = 0.01$, Figure 3.4A]. In this region, more activation was observed for reciprocate choices when the risk for the first player was high relative to when the risk for the first player was low. Finally, there was a main effect of risk in the right insula [$F(1, 17) = 8.80$, $P < 0.005$, Figure 3.4B], but opposite to the risk effect in the rTPJ, this region was more active when participants chose to reciprocate when the risk for the first player was low relative to when the risk for the first player was high. Thus, rTPJ was more active when participants decided to reciprocate and repaid the risk that was taken by the first player. In contrast, the right insula was more active when participants reciprocated despite the low need for repayment. Finally, there were no effects of risk or benefit for the aMPFC or the striatum.

Frequency Effects. Because the changes in activation can be influenced by frequency effects, we correlated activation in the ROIs with the frequency of different types of behavior to test whether the reported effects of risk and benefit can be explained by frequency differences. In addition, we added the frequency of behavior as a covariate of interest in ANCOVAs. Together, these effects showed that the risk and benefit effects were not correlated with frequency of choices, except for neural activation in the insula (see

supplementary data). That is, activation in the insula was highest for the least frequently occurring choices.





3.4 Discussion

The goal of this study was to investigate the neural correlates of reciprocity motives in brain regions that have previously been associated with mentalizing (aMPFC, rTPJ), reward and arousal (ventral striatum and insula) and inhibition of selfish impulses (ACC, rDLPFC). As expected, our behavioral results showed that participants reciprocated more when the first player took a high risk to trust and when the benefit of being trusted was high for the trustee, indicating that when reciprocating participants took into account both the consequences for the other as well as for themselves (Pillutla et al., 2003; van den Bos et al., manuscript submitted). Consistent with previous studies, our brain imaging data demonstrated that several brain regions worked together when individuals reciprocated trust and, in addition, provided more insight into how these regions were differentially sensitive to reciprocity motives.

First, separate analyses revealed that the two important areas of the mentalizing network, the aMPFC and rTPJ (Frith and Frith, 2003) have

separable functions in reciprocal behavior. Consistent with previous studies, the aMPFC was more active when participants defected compared to when they reciprocated (Gallagher et al., 2002; Decety et al., 2004). As such, the aMPFC was more active when the personal outcome of the decision was the greatest. This result is consistent with the hypothesis that the aMPFC is important for self-referential processing (Northoff et al., 2006; Ochsner, 2008) and with the interpretation that the aMPFC may have a general role in the evaluation or representation of reward information (Harris et al., 2007; van den Bos et al., 2007; Hampton et al., 2008). However, supplementary analyses revealed that the activation in aMPFC was not sensitive to the magnitude of personal gain (see supplementary data). Contrary to our predictions, there was no effect of the benefit manipulation on the activity in the aMPFC. Apparently, activation in the aMPFC is not directly sensitive to changes in cooperative intentions of the other player, but this region is sensitive to increases in personal outcome (defection). In future studies, it will be important to not only test motives for reciprocity, but also motives for defection.

In contrast to the aMPFC, the right TPJ was not sensitive to the type of choice but was sensitive to the risk manipulation when reciprocating. Activity in this area was higher when participants reciprocated when the risk was high rather than low. In the high-risk condition, the consequences of the participants' decision to reciprocate were fairly large for the first player compared to the low-risk condition. This finding indicates that, in line with our hypotheses, the rTPJ is involved in the shifting attention from the self to the other (Lamm et al., 2007) in order to distinguish between the consequences for self and other in a social decision-making paradigm (Lamm et al., 2007). This interpretation is consistent with a recently postulated hypothesis that argues that the rTPJ is involved in the reorientation of attention from self to other (Decety and Lamm, 2007; Mitchell, 2008).

Interestingly, our results also show that the activity in the rTPJ is sensitive to individual differences in SVO. That is, proself individuals showed more activation in the rTPJ when reciprocating, whereas prosocial individuals showed more activation in the rTPJ when defecting. Different processes may underlie these differences in neural activation for prosocials and proselfs, but one explanation may be that individuals with a prosocial orientation have their goals more aligned with those of the other, leading to less attention shifting when reciprocating, but more attention shifting when defecting (Decety and Hodges, 2006). These hypotheses should be further tested in future research.

The ventral striatum and insula were hypothesized to be sensitive to reward and arousal manipulations and were expected to be particularly sensitive to individual differences in reciprocal behavior. Indeed, regression analyses

demonstrated that activity in the striatum was higher for reciprocal choices than for defective choices for the prosocial participants (albeit at an unconservative threshold, but confirmed by unbiased ROI analyses, see supplementary results), whereas the proself participants showed the opposite pattern. The pattern of activation for the prosocial individuals is consistent with prior studies, which showed that cooperative choices are associated with ventral striatum activity (Fehr and Camerer, 2007). Even though the choice to reciprocate resulted in larger mutual gain, it also yielded a smaller monetary personal reward. Possibly, for prosocial individuals reciprocating in itself has a higher reward value whereas for proself individuals the personal gain has a higher reward value. This interpretation should be treated with caution, because it relies on reverse inferencing (Poldrack, 2006), but the results fit with a hypothesis postulated in a recent review analysis on other-regarding preferences (Fehr and Camerer, 2007). This hypothesis suggests that the ventral striatum represents the positive experienced utility of cooperation.

The insula was also sensitive to individual differences in SVO. However, the insula showed the opposite pattern of activity compared to the striatum. Furthermore, the insula showed sensitivity to the risk manipulation. The pattern of activation suggests that the insula is indeed sensitive to norm violations (King-Casas et al., 2008). That is, prosocial participants showed more activation in the insula when they defected (the unlikely alternative given their SVO), whereas the proself participants showed more activation in the insula when they reciprocated (again, the less likely option given their SVO). In addition, the insula was activated on those trials where participants chose to reciprocate when the risk that the first player took was low. In that case, there was less incentive to reciprocate than in the high risk situations. However, even though the choice to reciprocate occurred less frequently when the risk was low compared to when it was high, our supplementary analyses, using the frequency of the choice as covariate, revealed that these effects could not be attributed to a nonspecific effect of frequency. Together, these findings support the hypothesis that the insula is most active when a personal norm is violated (which can be a reciprocate norm for prosocial individuals or a defect norm for proself individuals) (Singer et al., 2006; Montague and Lohrenz, 2007). As such, the anterior insula have a more general role in social decision-making besides marking events as negative, such as pain, disgust or unfair offers (Sanfey et al., 2004; de Vignemont and Singer, 2006). Rather, the insula may be sensitive to the arousal associated with norm violations, which could also explain why the anterior insula are activated following other types of unexpected events such as a risk prediction error (Preuschoff et al., 2008). Alternatively, the insula

responses to violation of personal norms may serve as control signals, which mark social expectation violations (King-Casas et al., 2008).

Prior studies have suggested that cooperative behavior involves not only brain regions which are sensitive to mentalizing or reward representation, but also the control of impulses and actions. These studies have suggested that the ACC and the rDLPFC are important for regulating impulses to either defect or cooperate (Knoch et al., 2006; Rilling et al., 2007). Consistent with these earlier studies, in the current study, we showed that indeed the ACC and the rDLPFC were most active when social impulse control was required. In particular, ACC and rDLPFC were activated when participants reciprocated even though the benefit of being trusted was low. In other words, when the external incentive to reciprocate was low, the ACC and the rDLPFC were more engaged in reciprocal decisions. Inspection of the figures shows that the pattern of results observed for the insula follows a similar pattern as observed for ACC and rDLPFC, regions thought to be important for cognitive control (Ridderinkhof et al., 2004) and inhibition of self-oriented impulses (Knoch et al., 2006). It should be noted that, in this study, we could not distinguish between brain activity related to the actual choice and the appraisal of this choice. Thus, it is possible that ACC and rDLPFC activation is associated with the decision phase and the insula activation with the appraisal phase. These are important questions to test in future research.

Furthermore, activation in the ACC but not the rDLPFC, was also modulated by SVO. In prosocial individuals, the ACC was more active when reciprocating than when defecting, whereas in prosocial individuals, the ACC was more active when defecting than when reciprocating. One explanation for its role in both overriding the tendency to defect when the benefit is low, and the modulation of defecting vs. reciprocating depending on SVO, may be associated with the experience of response conflict (Botvinick et al., 1999). Importantly, activation in ACC and rDLPFC was not correlated with the frequency of making specific choices, arguing against the possibility that the effects can be explained by non-specific frequency effects.

Conclusion

Together, the results of this study demonstrated that several brain regions are differentially sensitive to reciprocity motives. We demonstrate that even though several brain areas are sensitive to individual differences in SVO (ACC, insula, rTPJ), these regions are differentially sensitive to the risk and benefit manipulations. The combined interpretation of sensitivity to SVO and modulation by risk and benefit manipulations allowed for advanced inference of the putative roles of these regions in reciprocal behavior. Our analyses revealed

the different motives for reciprocity, the risk for the trustor and the benefit for the trustee could be dissociated on the neural level. This study suggests a number of directions for future research as well as testable hypotheses. The differential involvement of the reported regions in reciprocal exchange demonstrates that neuroimaging methods may provide insight in the neural correlates of behavioral differences between individuals. It is possible that similar social interaction tasks could be used to explore social processing in a variety of populations, including developmental populations as well as individuals who fail to take the intentions of others into account.

3.5 Supplementary material

Supplementary Table 3.1: Percentage of average reciprocity in the four conditions. Standard errors between brackets.

	High Benefit	Low Benefit
High Risk	64% (.06)	53% (.06)
Low Risk	58% (.06)	27% (.04)

Supplementary Table 3.2: percentage of cooperative choices per factor level and participant. Participants in red are excluded from further fMRI analysis because there were either no observations in the reciprocate or defect condition.

PPN	High Risk High Benefit	High Risk Low Benefit	Low Risk High Benefit	Low Risk Low Benefit
101	9 (36%)	11 (48%)	17 (49%)	15 (36%)
102*	43 (100%)	44 (100%)	48 (100%)	5 (12%)
103	19 (76%)	17 (74%)	17 (49%)	21 (50%)
104	22 (88%)	19 (83%)	32 (91%)	26 (62%)
105	17 (68%)	11 (48%)	22 (63%)	18 (43%)
106	14 (56%)	8 (35%)	15 (43%)	23 (55%)
107	13 (52%)	7 (30%)	23 (66%)	6 (14%)
108	22 (88%)	20 (87%)	29 (83%)	12 (29%)
109*	0 (0%)	0 (0%)	0 (0%)	0 (0%)
110	8 (32%)	13 (57%)	4 (11%)	6 (14%)
111	23 (92%)	8 (35%)	29 (83%)	10 (24%)
112	12 (48%)	9 (39%)	9 (26%)	5 (12%)
113*	43 (100%)	44 (100%)	45 (94%)	25 (48%)
114	21 (84%)	17 (74%)	23 (66%)	27 (64%)
115	22 (88%)	12 (52%)	26 (74%)	10 (24%)
116	22 (88%)	19 (83%)	31 (89%)	11 (26%)
117	21 (84%)	16 (70%)	30 (86%)	9 (21%)
118	20 (80%)	8 (35%)	29 (83%)	6 (14%)
119	13 (52%)	10 (43%)	14 (40%)	5 (12%)
120*	0 (0%)	0 (0%)	0 (0%)	0 (0%)
121	7 (28%)	7 (30%)	6 (17%)	5 (12%)
122	5 (20%)	8 (35%)	12 (34%)	11 (26%)

Relation reciprocity level and activation in ROIs

The relation between individual differences in reciprocity and BOLD responses are further illustrated based on ROI values in Supplementary Figure 3.1. Based on the results of the regressions analysis the following reported results are all from the defect > reciprocate contrast. Subsequent post hoc analyses revealed that the reported effects were also significant for the reciprocate > fixation contrast but not for the defect > fixation contrast. Thus, these effects were primarily driven by differences in neural activation associated with reciprocity choices.

The patterns reported in Suppl. Figure 1 demonstrate a positive correlation between reciprocal behavior and activation in the anterior Insula (bilateral), the ACC and the right TPJ ($r = .79, p < .001$, $r = .59, p < .005$ and $r = .65, p < .001$ respectively). These patterns demonstrate that the individuals who reciprocated less (pro-self oriented) recruited these areas more when reciprocating compared to defecting whereas the prosocial individuals recruited these areas more when defecting compared to reciprocating.

In contrast, there was a negative correlation with reciprocity scores and activity in the ventral striatum ($r = -.64, p < .001$, see Supplementary Figure 3.1). These patterns show that the individuals who reciprocated more (prosocial individuals) activated these areas more when reciprocating,

Frequency effects

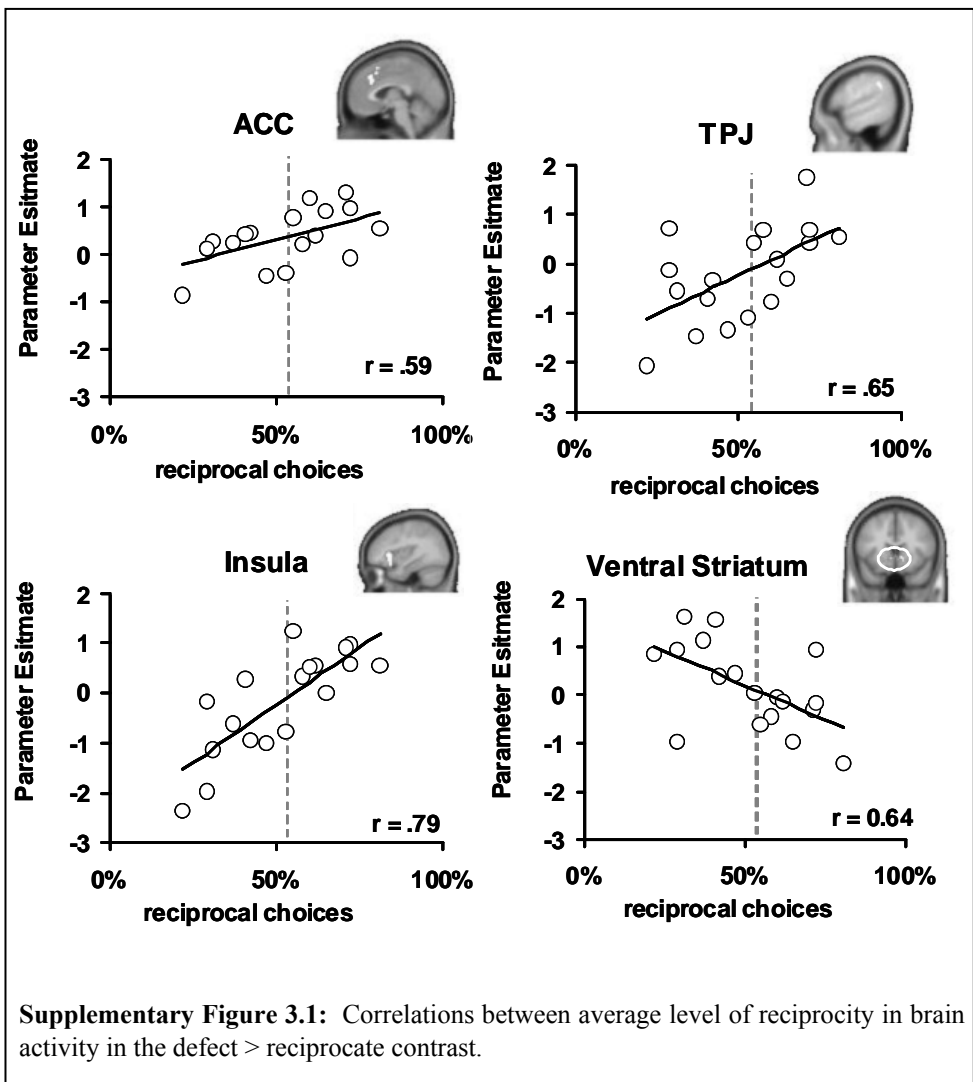
The results above show that the changes in activation can be influenced by frequency effects. To test whether the reported effects of Risk and Benefit correlated activation in the ROIs can be explained by frequency differences we performed several additional analyses.

First, we correlated the beta values for Low Benefit and High Benefit conditions with the frequency of these choices in the right DLPFC and the ACC. In all cases there was no significant correlation between the frequency of the choice and the beta values (all p 's $> .3$). The same analysis for the anterior insula resulted in a negative correlation between High Benefit Reciprocate frequency and beta values for neural activation in this condition ($r = -.588, p < .05$) and for Low Benefit Reciprocate frequency and neural activation in this condition ($r = -.719, p < .001$). This negative correlation indicates that the less frequent a choice was made, the higher the beta value for that choice.

Finally, we analyzed the effect of High and Low Risk in the anterior insula and TPJ in a similar manner. There were no correlations between frequency and beta values in the TPJ and Insula (all p 's $> .1$). This result was expected because activation in TPJ follows a pattern of activation opposite of frequency sensitivity.

Subsequently, we performed ANCOVAs with the frequency of the decisions as covariate. The results of these analyses show that all the effects reported in the original manuscript remain significant (at $p < .05$ threshold), furthermore the frequency covariates were never of significant influence (all p 's $> .1$). There was again one exception and that was the anterior insula; when testing for the effects of risk and benefit with frequency as covariate we found that the frequency of high benefit trials was significant ($F(1,13) = 4.972$, $p < .044$).

Together, these effects show that risk and benefit effects are not correlated with the frequency of choices, except for neural activation related to benefit, but not risk, manipulations in the insula.



Reward Magnitude

To further explore our interpretation of aMPFC function we further analyzed its sensitivity to reward magnitude. The small difference in the pay-off between trials made it possible to look at the neural correlates of the relative difference in gain for defect choices. For this analysis we divided the defect trials into either low or high gain trials (the difference only being one coin, or ten cents). For these analyses we contrasted the high gain defect choices with low gain defected choices. These analyses did not yield any significant effects at our $p < .001$ & 12 voxels threshold. Lowering the threshold for exploratory reason did not result in activation in aMPFC.