The Neural Bases of Framing Effects in Social Dilemmas

Julian Macoveanu
Thomas Ramsøy
Martin Skov
Hartvig R. Siebner
Toke R. Fosgaard*

Abstract

Human behavior in social dilemmas is strongly framed by the social context, but the mechanisms underlying this framing effect remains poorly understood. To identify the behavioral and neural responses mediating framing of social interactions, subjects underwent functional Magnetic Resonance Imaging while playing a Prisoners Dilemma game. In separate neuroimaging sessions, the game was either framed as a cooperation game or a competition game. Social decisions where subjects were affected by the frame engaged the hippocampal formation, precuneus, dorsomedial prefrontal cortex and lateral temporal gyrus. Among these regions, the engagement of the left hippocampus was further modulated by individual differences in empathy. Social decisions not adhering to the frame were associated with stronger engagement of the angular gyrus and trend increases in lateral orbitofrontal cortex, posterior intraparietal cortex, and temporopolar cortex. Our findings provide the first insight into the mechanisms underlying framing of behavior in social dilemmas, indicating increased engagement of the hippocampus and neocortical areas involved in memory, social reasoning and mentalizing when subjects make decisions that conform to the imposed social frame.

Keywords: Social reasoning, prisoners dilemma, fMRI, framing

JEL code: C90

*Department of Food and Resource Economics, University of Copenhagen, Rolighedsvej 25, 1958 Frederiksberg C, Denmark. Phone: +45 61687102, Email: tf@ifro.ku.dk

Introduction

Social interaction between humans often constitutes a dilemma. While in many situations cooperation between two or more agents leads to increased benefits for everyone, the individual often can benefit more from selfish behavior rather than mutual cooperation. The propensity to put oneself first takes into account how one´s own selfish actions may trigger sanctions by the interaction partner Fehr & Gintis 2007; Fehr & Schmidt, 1999). Further, one may first exploit the cooperative behavior of the other group members before engaging in cooperative interactions (Baumgartner et al., 2009). In other words, when engaging in social interactions, one's self-interest is weighted against the concerns for the intentions and behavior of others, and how the interaction partners may respond to one's actions.

The factors that drive decisions in such social dilemmas have been studied in economic and psychological experiments for decades (Fehr & Gintis, 2007; Fehr Schmidt 1999; Baumgartner et al., 2009; Spitzer et al., 2007; Zelmer, 2003). A key finding is that actors' decisions are influenced by how the social dilemma is presented (Pruitt, 1967; Andreoni, 1995; Deutsch, 1958; Liberman, Samuels & Ross, 2004), often referred to as "framing" (Pelphrey, Morris & McCarthy, 2004). This violation of the so-called description invariance principle is puzzling (Camerer & Thaler, 1995), as preferences should not change relative to how the options are presented. Although it is still under debate how framing shifts decision preferences in social dilemma situations, it is possible that framing modulates neural processes associated with mentalizing, i.e., the capacity to infer the mental states of others (Dufwenberg, Gächter, & Hennig-Schmidt, 2011). Indeed, social dilemmas require actors to predict the intentions of others to deduce how they will respond to specific acts, and several behavioral studies have already suggested that mentalizing plays a crucial role in solving framed social dilemmas. For example, framing has been shown to affect how cooperative a subject expects other group members to be (Dufwenberg, Gächter, & Hennig-Schmidt, 2011), and to shape the moral assessment of free riding behavior (Cubitt et al., 2011), yet the exact psychological mechanisms underlying these effects remain to be identified. Notably, a recent report by Chang and Sanfey (2011) show

that subjects' expectations about other's intentions in the Ultimatum Game provided a stronger explanatory power on their actual decisions compared to alternative explanations, such as inequality aversion.

Mentalizing in social dilemmas

If mentalizing is a driving force behind framing the behavior in social dilemmas, different measures of mentalizing should indeed reflect differences in the degree with which people are affected by framing. Here, our aim was to assess the effects of framing in social dilemmas by employing both behavioral and neurobiological measures of mentalizing. We conducted two related studies using an iterated version of the Prisoners' Dilemma (PD) game (Axelrod & Hamiloton, 1981; Gibbons, 2006) to probe whether subjects engage psychological and neural processes related to mentalizing when choosing between conflicting choices in two different contextual frames (i.e., collaboration versus competition frame). We first conducted a behavioral study focusing on overt measures of mentalizing, including asking the subjects what they thought the other players would do. In a second experiment, we used functional Magnetic Resonance Imaging (fMRI) to study the neural underpinnings of the framing effect while subjects played the PD game in the framed context of collaboration and competition. We were particularly interested in identifying neural activity associated with social decisions that adhered to the social context imposed by the frame. Here, we hypothesized that decisions that were aligned with the imposed frame (conformity), as opposed to decisions opposing the frame (non-conformity), would be associated with increased activation in the mentalizing network, specifically the superior temporal region (gyrus and sulcus), precuneus and medial prefrontal cortex (Baumgartner et al., 2011; Mitchell, 2009; Yarkoni et al., 2011; Knowch et al., 2008). We also expected increased engagement of the hippocampal formation, as prior research has implicated this region in different kinds of framing effects (Eichenbaum, Yonelinas & Ranganath, 2007; McClure et al., 2004). Specifically, we assert that there is a fundamental difference between situations in which people are directly affected by contextual information and when they are not. Being affected by a frame is

believed to be related to the involvement of a particular kind of heuristic that has a strong and direct effect on behavior. Conversely, not being affected by a frame could be the result of either the frame not triggering this heuristic, or the person being able to dispense from this response via higher-order control functions such as executive control. Together, this implies that situations in which our behavior is affected by a social frame will be related to a stronger engagement of brain regions involved in social functions.

In the experiments, subjects played an iterated PD game with a “stranger matching” procedure (Andreoni & Croson, 2008) where they decided whether to cooperate or not. In each trial, a grid was presented on a screen showing the financial outcomes of four combinations based on whether they chose to cooperate or defect, respectively. The grid provided information about their own potential economic gains (player A) as well as those of their interaction partner (player B). The dominant strategy for each player is to not cooperate (Gibbons, 2006). However, the outcomes were structured such that if both players chose to cooperate, their joint outcome was maximized. Each trial was pseudo-randomly framed as either a “Competition” game or a “Cooperation” game with the aim of swaying the subject to view the interaction with the other player as being either antagonistic or cooperative. Only the label of the game changed across frames (see Figure 1).

In the behavioral study, subjects performed two tasks: in a decision phase, they decided between option A and B which was a choice between cooperation and defection. In a subsequent belief phase, the subjects were asked to guess their opponent's choice (indicating cooperation or defection). The subjects were informed that the monetary outcomes of two randomly selected trials would be used to calculate how much they would be paid after the game. To ensure motivated efforts during the belief phase, the subjects were also informed that they would receive 10 DKK (≈2 USD) for each correct belief assessment drawn from two random trials.

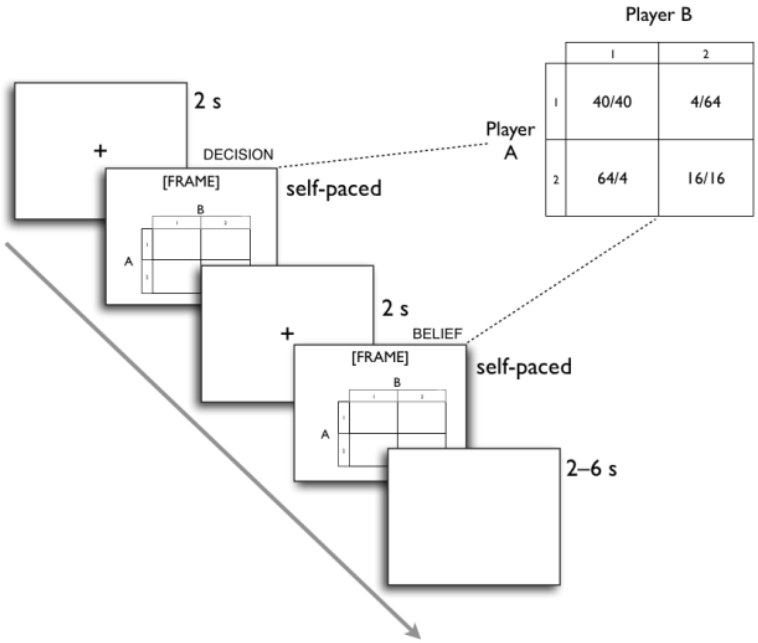

Figure 1. *Experimental design of the behavioral study (A), consisting of (1) a Decision phase showing the full decision matrix (2) a Belief elicitation phase. Both were organized as self-pace. The participant always acted as player A, and used the right hand index and middle fingers to respond with option A and B, respectively. The decision was self-paced and the decision screen disappeared after the decision was made. Our behavioral results demonstrate a clear effect of framing on cooperation and defection rates (B).*

In the neuroimaging experiment, a separate group of subjects underwent whole-brain fMRI while performing the PD game. The PD set-up was identical to the behavioral study except for two notable differences (see Figure 2). We eliminated the explicit belief phase as we were more interested to map the neural correlates of covert mentalizing when making the actual decision. We also separated the presentation of the frame information ("Cooperation" vs "Competition") from the decision phase to dissociate brain activity generally associated with the framing context from brain activity related to decision making in a social dilemma. Subjects were matched using a validated procedure (Baumgartner et al., 2009) and told that they had been matched with partners from the previous behavioral study, but were not provided any

information about how their partners had responded. Subjects participating in study 2 were matched with some of the subjects who completed study 1. This means that the decisions of subjects from study 1 might be used twice if they are matched with test persons from study 2. After completion of study 1, we sought approval from subjects participating to possibly use their decisions in a subsequently experiment. The subjects who agreed were also paid for the second set of decisions.

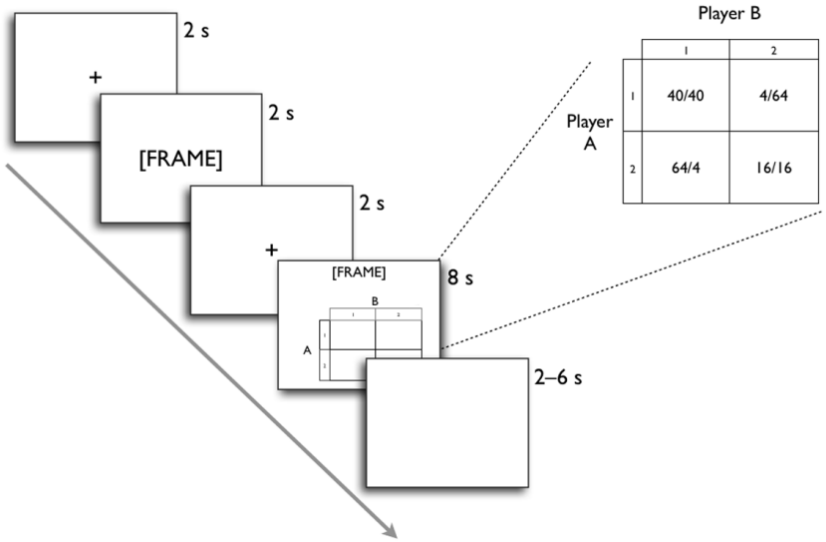

Figure 2. *Experimental design of the fMRI study (A), consisting of (1) a Framing phase, where subjects either saw the words 'Cooperation' or 'Competition'; (2) a Decision phase showing the full decision matrix wherein subjects had 8 seconds to decide (the decision screen was present throughout the 8 seconds, regardless of the actual choice). Slide duration is shown in seconds.*

Materials and Methods

Study 1 – Relationship between framed decisions and mentalizing

The protocol for both studies was approved by the local ethics committee (KF 01–131/03), and adhered to the standards expressed in the Declaration of Helsinki.

Participants. Thirty subjects (16 men/14 women; age mean/std = 24.7/9.8)

were recruited from the Copenhagen region through randomized sampling from a database of volunteers at the Department of Economics, University of Copenhagen. All subjects completed the experiment in one session, which lasted approximately 45 minutes. The study took place at the Laboratory for Experimental Economics, the Department of Economics, University of Copenhagen.

Experimental paradigm. Before playing the game, the subjects received instruction and training for the task. Subjects were told that they would be matched with a new player for each repetition of the game. Thus, subjects knew that in each repetition of the game they were randomly matched with a new player. Throughout the experiment subject would not be able to see the responses of the other player. Thus, no feedback was given between the repetitions. Before the game, the subjects also went through additional paper and pen tests to assess their cognitive, emotional and interpersonal traits (more details below).

Subjects first received standardized written instructions, which explained that they would be matched with a new person for each trial ("stranger matching") and that they would receive a fixed payment for attending the session. Subjects were also told that four randomly selected trials would be used for additional payment (see below). Each subject was positioned in front of a computer at an average viewing distance of 60 cm. Stimulus presentation and response recording (response and response time) were performed using E-Prime (PST Tools Inc., www.pstnet.com) in a Windows XP environment. During the game, each trial either displayed the word "Cooperation" or "Competition" above the payoff matrix (see Figure 1). Subjects were explicitly instructed that these words were the names of the games, and that their partner received the same information. This approach followed the tradition of behavioral economics, in which subjects were not instructed on any exact interpretation of the frame information, but rather provided the choice environment, and within this environment make their choice. (e.g., Liberman, Samuels & Ross, 2004; Dufwenberg, Gächter, & Hennig-Schmidt, 2011; Chang & Sanfey, 2011). The choice of having the frame information present during the choice was also made to minimize the possibility that participants forgot the frame.

Between each trial, a fixation cross was shown for two seconds. During the decision phase, the subjects responded by pressing one of two buttons to indicate their choice; in effect whether they chose to "cooperate" or "defect." The paradigm did not put any constraints on the response times. The same buttons were used in the "Belief" phase to indicate the anticipated choice of their opponent. Subjects were informed that they would be playing against a real human being to increase the level of engagement in the game.

The experiment consisted of 56 trials, which were presented in a fixed pseudo-randomized order, i.e., the order of the trials were randomized during the experiment design phase but shown in the same order to all participants. There was an equal number of cooperation and competition trials. To avoid routine behavior and increase motivation and task performance, the payoff structure was variably and pseudo-randomly scaled across trials (20%, 40%, 60%, 80%, and 100% scaling of maximum payoff). In the 100% scaling, the payoff when the opponent cooperated and the subject defected was 5 DDK (≈1 USD) and 80 DDK (≈13.5 USD), respectively; mutual cooperation earned them 50 DDK (≈8.5 USD) each; mutual defection earned them 20 DDK each (≈3.5 USD) each. While the absolute value of the payoff varied, the relative value remained unchanged. The applied scaling was balanced across the frames and varied in a pseudo-randomized manner. The screenshot presented in Figure 1A is an example of an 80% scaling. Each person's choice option was labeled as "1" and "2", respectively, and subjects made their choice by pressing the "1" and "2" button on the PC keyboard. As a further step to ensure the subjects' full attention, the spatial order of the location of the choice options was changed in 25% of the trials. This manipulation was used to avoid automatic button pressing for any given decision. An overview of the experimental design is illustrated in Figure 1a.

Behavioral data analysis. We first tested whether previously reported between-subject framing effects in social dilemmas were also present when using a within-subject experimental design. We applied Wilcoxon sign rank tests to compare the individual degree of cooperation, the degree of beliefs regarding the interaction partner (cooperation or no cooperation), and the average response time used. This

enabled us to test whether the type of frame biased the frequency distribution of subject´s decision to cooperate or not or whether the subject´s decision to cooperate or not was associated with a bias in the subject´s belief regarding the cooperation of the partner. To evaluate the importance of relations between our main variables (decision, conformity, belief, and frame) we also applied a few other tests. In particular, we used a Tetrachoric correlation test, a Pearson chi-square test, and a Mann-Whitney ranksum test. The tetrachoric correlation test is designed to evaluate correlations of binary variables (Digby 1983), whereas the Pearsons' chi square test the distributions of two categorical variables (Pearson 1900), and the Mann-Whitney test is a non-parametric test which we use to compare the cooperation across frames (Mann & Whitney 1947). The analyses were performed in the statistical software package STATA (StataCorp LP, www.stata.com). Significance level was set at $p<0.05$. Group data are reported as mean ± one standard deviation, if not specified otherwise.

Study 2 – Neural correlates of framed decisions in social dilemmas

Participants. Fourteen subjects (3 male/ 11 women; mean±std age=30.5±3.4) were recruited from the Copenhagen region using the same methods as in Study 1, whilst ensuring that enrolled subjects had not participated or heard of Study 1. All subjects signed an informed consent and a standardized self-report scheme on medical history and other relevant information. Exclusion criteria included a history of or current psychiatric or neurological illness, or suspicion thereof, and factors not compatible with MR scanning (e.g. claustrophobia, pace maker, magnetic ligands in the body). No subjects were excluded on such grounds. The fMRI study took place at the Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre. To assess individual differences in empathic ability, subjects completed an eight-item version of the Empathy Quotient (EQ-8) assessment, an assessment of empathic ability (Loewen, Lyle & Nachshen, 2009), from which we calculated each individual's EQ score.

Experimental paradigm. The trial structure of the experimental task is illustrated in Figure 2 and differed in two aspects from the trial structure used in the

behavioral experiment (study 1). Subjects had only to decide, but were not required to report their beliefs about the partner´s mode of interaction. Further, the framing cue was always presented before the presentation of the grid to establish the frame ("Cooperation" or "Competition") before the decision phase. As in the behavioral study, subjects received instructions that these were names of the games, and that their partner received the same information, but were not instructed about the meaning of these labels. The trial structure was as follows: First a fixation cross appeared in the middle of the screen for 1 second, then subjects were presented with the frame information ("Cooperation" or "Competition") for 2 seconds. Thereafter, a fixation cross appeared in the middle of the screen for 1 second, which was followed by the grid which displayed the players' options, similar to Study 1. This grid was displayed for 8 seconds. As in Study 1, the response options were labelled "1" and "2". The subjects were asked to make their choice while the grid was displayed, using their index and middle finger to press buttons on a response box to respond "1" and "2", respectively. Each trial was separated by a black screen, which was pseudo-randomly jittered for 2 to 6 seconds. All in all, the fMRI session lasted for 18 minutes comprising a total number of 64 trials. As in Study 1, the trials were programmed in a fixed pseudorandom order., using an event-related paradigm. All subjects saw the same order of trials, with equal number of trials (n=32) in each condition. As in Study 1, 25% of trials had a different variation in the choice options, allowing us to avoid habitual motor responses. This choice was made both to avoid pure automatization of choice, and to have the two studies as much aligned as possible.

With regard to the social decision, participants received the same instructions as subjects who had participated in the behavioral study. However, they were not told to express their beliefs regarding the cooperation style of the interaction partner.

In study 2 we implemented a matching procedure developed by Baumgartner et al. (2009). Subjects were told that they would be randomly assigned to players from a previous behavioral study (that is study 1), but they would not be informed about how these players had actually responded. In practice we implemented this procedure by asking subjects in study 1 whether we could use their decisions from study 1 in a later

experiment and pay them again according to the outcome of their (re-used) decisions and the decisions made by participants in the later study. All participants in study 1 agreed to have their decisions reused.

Behavioral analysis. Behavioral data were analyzed and reported as described in Study 1, for details see Study 1. To further probe individual differences in behavior and brain responses, we included the EQ score as a covariate in the analysis

Image acquisition and analysis. Subjects were scanned using a Siemens Magnetom Trio 3T MR scanner (Erlangen, Germany) with an eight-channel head coil (Invivo, FL, USA). Consistent head placement within the scanner was ensured by orienting the head to predefined reference marks on the scanner head coil. Movement was minimized by applying cushions to fix the head in position. A scout scan was run to define the field of view (FOV) for the subsequent scans. The subjects were first scanned using a structural T1-weighted MPRAGE (Magnetization Prepared Rapid Acquisition Gradient Echo) scan with a voxel dimension of 1x1x1 mm3, FOV=256 mm, matrix 192x256x256, TR/TE/TI = 1540/3.93/800 ms, and a flip-angle of 9°. During task performance, the subjects were scanned with a T2* weighted Blood Oxygenation Level-Dependent (BOLD) fMRI protocol using an EPI sequence with the parameters TR/TE = 2430/30 ms, 64 x 64 matrix, a flip angle aligned to the AC-PC line (approximately 12°), 42 slices and a voxel size of 3x3x3 mm3 with no inter-slice space.

Preprocessing and analysis were performed using SPM8 (Wellcome Department of Imaging Neuroscience, London). Images were realigned without smoothing. The EPI image series was co-registered to each individual's AC-PC aligned structural image using mutual information, trilinear interpolation without warping, and was subsequently checked manually. Images were normalized to the MNI template and smoothed using a Gaussian kernel with a full-width at half-maximum of 8 mm. The first 2 volumes of each session were discarded to allow for T1 equilibration effects.

The individual fMRI time series were analyzed using multiple regression analysis (General Linear Model, GLM) with separate event regressors for the frame

information phase, the decision making phase (when subjects were presented with the response options) and a separate regressor for motor responses (modeled at the time of the button press, duration = 0). Rest was not modeled independently. The regressors were convolved with a canonical hemodynamic response function. We also accounted for artifacts caused by head movement, pulse and respiration by including an additional 24 nuisance regressors in the first-level analyses.

For the decision making phase, we modeled the combinations of a two (framing: Cooperation, Competition) by two (behavior: defect, cooperate) factorial design. For the construction of regressors of interest, we let α be the cooperation frame, β be the competition frame, and x and y as cooperate and defect behaviors, respectively. Thus, we constructed four regressors of interest at the first-level analysis: α_y, α_x, β_x and β_y. To study the general relationship between frame conformity and brain activation, we analyzed the instances in which subjects aligned their choice to the frame, using the following model:

$$[(\alpha_x - \alpha_y) + (\beta_y - \beta_x)]$$

In addition, our study design allowed us to explore the neural correlates of frame nonconformity, i.e., when the subject´s decision violated the mode of social interaction as imposed by the frame. To study the neural engagement during nonconform choices we analyzed both instances of nonconform behaviors, i.e., $[(\alpha_y - \alpha_x) + (\beta_x - \beta_y)]$.

After testing a priori hypotheses, a post-hoc analysis was run to test for an additional modulatory role of individual empathy scores according to the EQ-8 questionnaire (Loewen, Lyle & Nachshen, 2009). We tested the univariate regression of empathy scores on neural engagement for the conformity > non-conformity contrast, and then on the non-conformity > conformity contrast.

Finally, we studied the effects of framing on neural activation by analyzing only the framing phase. Here, we modeled the cooperation frame and competition frame as independent regressors. Paired T-tests were run for competition>cooperation

and cooperation > competition.

Regarding frame conformity, we had specific neuroanatomical hypothesis. We reasoned that for frame conformity, framing should increase mnemonic processing and engage brain regions involved in mentalizing. Thus, we defined regions of interest (ROI) based on previous regions found to be activated during mnemonic (Eichenbaum, Yonelinas & Ranganath, 2007; McClure et al., 2004; De Martino et al., 2006) and mentalizing processes (Iacoboni et al., 2004; Baumgartner et al., 2011; Mitchell, 2009; Yarkoni et al., 2011). Each ROI was constructed using a 10 mm sphere centered in the peak voxel coordinate of the a priori region. For the hippocampus, the coordinates were (x, y, z = -24, -24, -20 and 20, -20, -16), and for the mentalizing network our choice of specific regions were guided by a review of the neural bases of mentalizing (Mitchell, 2009) and using specific coordinates from a Neurosynth meta-review based on 33 neuroimaging studies under the term "mentalizing" (see http://neurosynth.org/terms/mentalizing/studies). Thus, we used each reported peak voxel as our coordinate of interest when defining the center of each sphere, and then used a 10 mm sphere around this center. The mentalizing regions included (0, -52, 36) for precuneus; (-52, -56, 20 and 52, -56, 20) for the bilateral temporo-parietal junction; (56, 2, -20 and -56, -4, -20) for the bilateral superior temporal region; (0, 48, -16) for the ventromedial PFC; and (0, 32, 52) for the dorsomedial PFC. All ROI analyses used the aggregated activation across each individual ROI.

Clusters from frame conformity and linear regression analysis were considered significant at $p<0.05$ after correction for multiple comparisons using family-wise error correction (FWE) within the predefined ROIs (height threshold prior to small volume correction was $p<0.001$ uncorrected). Activated regions outside the predefined ROIs were considered significant at $p<0.05$ FWE corrected across the whole brain, reporting only on the cluster-level p-values. We also considered as trend activations voxels $p<0.001$ uncorrected for multiple comparisons at the cluster-level.

Results

Study 1 – Relationship between framed decisions and mentalizing

Subjects earned 139 DKK (≈26 USD) on average, including an attendance fee of 50 DKK. In the decision phase, subjects took an average of 9.94 (0.41) seconds to decide. Conformity ($\alpha_x + \beta_y$) did not differ in response times compared to nonconformity ($\alpha_y + \beta_x$; Wilcoxon signed rank z=-1.64, p=0.101). Similarly, in the second phase – the belief phase – the average response time was 2.78 (0.12) seconds, and he frame manipulation did not significantly affect response time. (Wilcoxon signed rank z=-0.46, p=0.649).

Table 1a provides an overview of the choices and expected choices (belief) of others and Figure 3 illustrates the individual conformity rates across the frames. Overall the cooperation rate was 29% in the competition frame, compared to 61% in the cooperation frameA Wilcoxon signed rank test (z =4.07, p<0.001) reveals that the individual cooperation rates were significantly different across the frames. The cooperation rates corresponds to an overall conformity rate of 61% in the cooperation rate but 71% in the competition frame. The individual conformity rates are illustrated in Figure 3 and it is evident that there exist large variation as to how frame conform choices subjects chose. Some subjects always conform with the frame, whereas other never do. Framing also significantly affected subjects' belief regarding the behavior of the interaction partner (player B): On average, the expectation that player B's decisions would be cooperative was 50% in the competition frame, but was 65% in the cooperation frame (Wilcoxon signed rank test: z=3.43, p=0.001). There was a significant relationship between this belief and the subject's own choices, as illustrated in Table 1b. (Tetrachoric rho = 0.43, p<0.001) $x^2(1)=127.691$, p<0.001): Independently of the frame, 55% of defection choices were accompanied by the belief that player B would also defect. Conversely, only 28% of cooperation choices were accompanied by a belief that player B would defect.

Importantly, this relation was significantly affected by framing. When choosing to defect, subjects reported a significantly higher belief that player B would defect in the competition frame (59%) than in the cooperation frame (48%;

Tetrachoric rho = 0.18, p=0.001). No such relation was found for cooperative choices: The expected degree of cooperation was 71% in the competition frame and 73% in the cooperation frame (Tetrachoric rho = 0.02, p=0.795).

Table 1- Overview of behavioral results in Study 1

A. Cooperation rate, and belief about cooperation, across frames

	Both Frames	Competition frame	Cooperation frame
Cooperation rate*	44.9% (0.49)	29.2% (0.45)	60.6% (0.49)
Conformity rate¤		70.8% (0.45)	60.6% (0.49)
Belief of cooperation	57.0% (0.50)	49.5 % (0.50)	64.5% (0.48)

Numbers in parentheses are standard deviations
** Cooperation rate is the relative occurrence of cooperation choices*
¤ Conformity rate is the relative occurrence of conforming choices (aligning choice and frame)

B. Average belief of defection

	Both Frames	Competition frame	Cooperation frame
When defecting	55.3% (0.49)	59,5% (0.49)	47.7% (0.50)
When cooperating	27.9% (0.45)	28.6% (0.45)	27.5% (0.45)

Numbers in parentheses are standard deviations

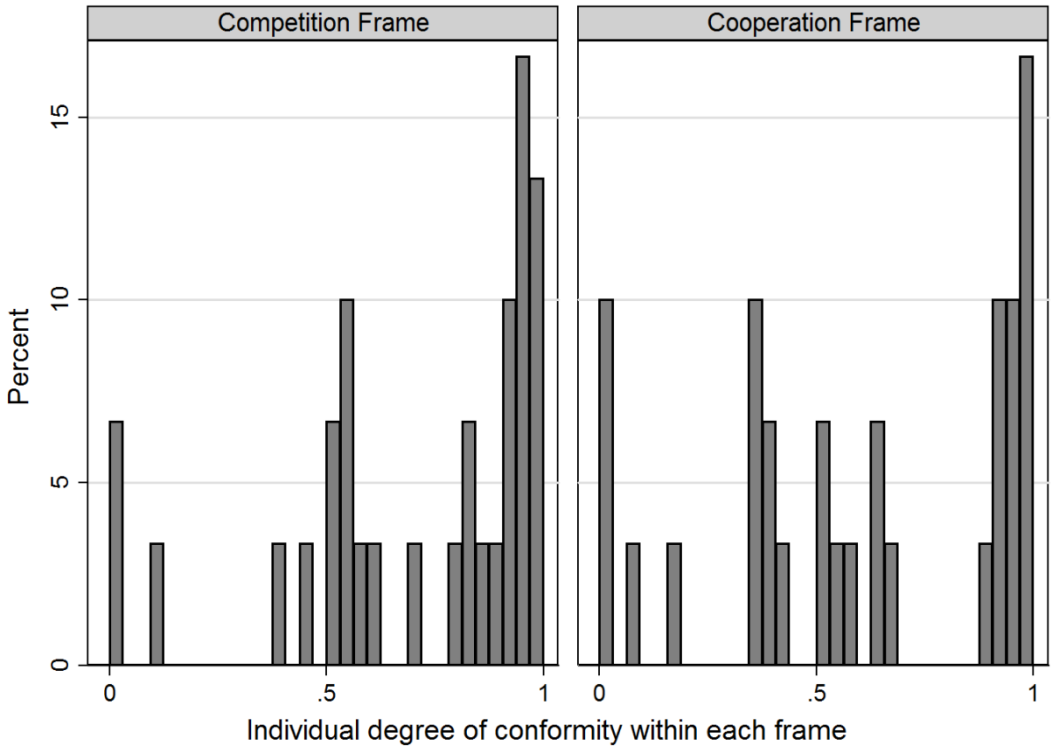

Figure 3. *The degree of conformity at the individual level in study 1, across the frames, ranging from no conformity (0.0) to full conformity (1.0). A decision is considered conform if the cooperative choices are taken in the cooperation frame, and the defection choices are taken in the competition frame. The individual measure is the degree to which the subject takes conform choices.*

Study 2 – Neural correlates of framed decisions in social dilemmas

Analyses of the behavioral data obtained during fMRI yielded a significant within-subject framing effect (Mann-Whitney test z=-6.82, p=0.000)) replicating the result that had been obtained in the behavioral experiment (study 1)[1]. In the competition frame, 31% of the decisions were cooperative, while in the cooperation frame, 54% of the decisions were cooperative. We found that there was a significantly relation between conform choices and the frames (Tetrachoric rho=-0.228, p<0.001). Overall,

[1] Comparing the individual levels of cooperation across frames a Mann-Whitney test also finds a significant difference (z=1.931, p=0.054), and similarly a Wilcoxon signrank test also find a significant difference (z = 1.759, p=0.079).

conformity rate was 68% in the competition frame as opposed to 54% in the cooperation frame. Individual conformity rates across the frames are illustrated in Figure 4, and in parallel with study 1 large variation is observed. The degree of conformity with the frame was found to vary substantially across subjects, ranging from 6.25% to 100%, where a score of 100% means that the subject always behaves according to the frame. The average response time of the decisions was 3.4±0.05 seconds, and did not differ between frames (Wilcoxon sign rank test: p=0.581) or decision types (Wilcoxon sign rank testp=0.424)[2]. Notably, the average response time in Study 2, where subjects had to respond within 8 seconds, was shorter than in Study 1 (9.94±0.04 seconds), where responses were self-paced.

Table 2- Overview of behavioral results in Study 2

Cooperation rate, and conformity rate, across frames

	Both Frames	Competition frame	Cooperation frame
Cooperation rate	43.1% (.50)	31.6% (0.47)	54.4% (0.50)
Conformity rate		68.4 % (0.47)	54.4% (0.50)

Numbers in parentheses are standard deviations
** Cooperation rate is the relative occurrence of cooperation choices*
¤ Conformity rate is the relative occurrence of conforming choices (aligning choice and frame)

Confirming our main hypothesis, frame conformity was associated with a significantly stronger bilateral activation of the hippocampal formation than nonconformity, and parts of the mentalizing network, including the lateral temporal cortex, precuneus, and the dorsomedial PFC (see Table 3). Notably, this activation pattern was present for both the competition frame and the cooperation frame,

[2] The reaction time was 3.45 seconds (1.71) in the competition frame, and 3.39 seconds (1.50) in the cooperation frame. Across the decision types, the reaction time was 3.46 seconds (1.50) for defection choices and 3.54 seconds (1.60) for cooperative decisions. The numbers in the parentheses are the standard deviations .

demonstrating that the engagement of these regions was associated with increased likelihood of conformity with the frame (Figure 3). Other regions of the mentalizing network, such as the temporo-parietal junction and the ventromedial PFC, did not show a relationship with conformity.

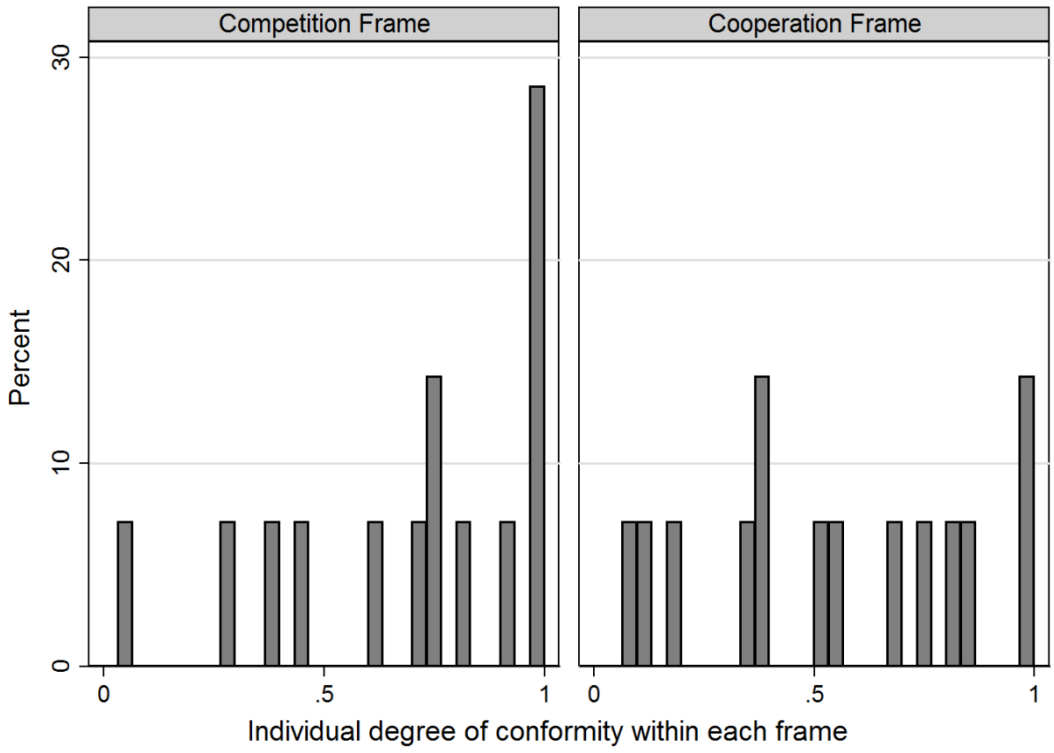

Figure 4. *The degree of conformity at the individual level in study 2, across the frames, ranging from no conformity (0.0) to full conformity (1.0). A decision is considered conform if the cooperative choices are taken in the cooperation frame, and the defection choices are taken in the competition frame. The individual measure is the degree to which the subject takes conform choices.*

To test for additional modulation of individual differences in mentalizing, we ran a post-hoc analysis in which each subject's score on a self-report empathy questionnaire was included as a covariate, first for the conform > non-conform contrast and the subsequently for the converse contrast (non-conform > conform). By testing the effect of individual differences in empathy score on our a priori brain

regions – the hippocampal formation and the mentalizing network (lateral temporal cortex, precuneus and dorsomedial PFC), only the left hippocampus showed a significant effect between empathy and neural activation, demonstrating a significant

Table 3 – Results from the analysis of a priori Regions of Interest during choice conforming to the frame (contrast: conformity > non-conformity).

Region	Hemisphere	Coordinates	Z	p
Lateral Temporal Cortex	R	56, 2, -20	4.81	0.002 FWE
	L	-56, -4, -20		n.s.
Temporo-parietal junction	R	52, -56, 20		n.s.
	L	-52, -56, 20		n.s.
Precuneus	medial	0, -52, 36	4.31	0.001 FWE
Ventromedial PFC	medial	0, 48, -16		n.s.
Dorsomedial PFC	medial	0, 32, 52	3.25	0.022 FWE
Hippocampus	R	30, -34, -8	4.28	0.001 FWE
	L	-26, -28, -14	3.68	0.014 FWE

Abbreviations: FWE = Family-Wise Error correction; n.s. = non-significant result. All results are reported as whole-ROI p-values where each ROI uses a priori coordinates as the center and a 10 mm sphere making up the ROI.

positive relationship between the level of empathy and engagement of the left hippocampus (-34, -30, -16, z=3.63, p<0.05 FWE corrected within predefined region of interest). This effect was only found for the conform > non-conform contrast, while no effect was found for the non-conform > conform analysis.

Exploratory analyses were run to test for additional activations during frame conformity, by using a whole-brain analysis and considering p-values at 0.05 FWE significant, and p<0.001 uncorrected as trend significant, both at cluster-level. The results are summarized in Table 4.

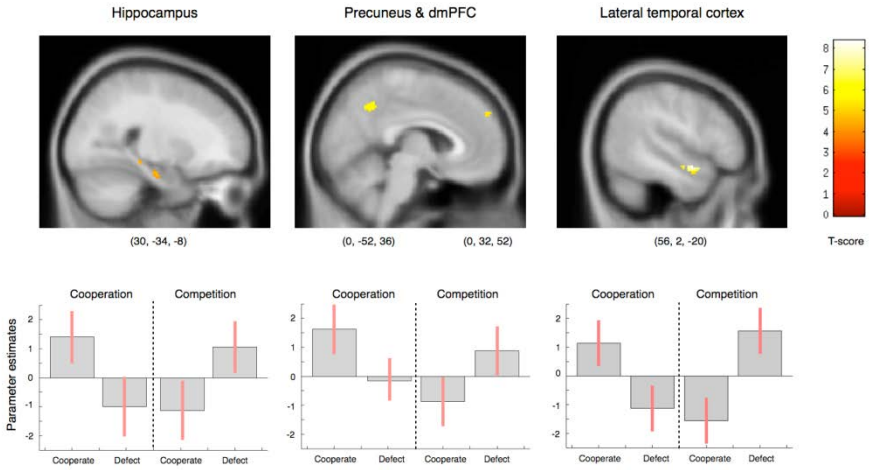

Figure 5. *Brain activity related to frame conformity in the PD game. Upper panel. Statistical parametric map of brain regions showing increased regional activation during frame conformity as opposed to frame-nonconformity. For display purposes, the maps are thresholded at p<0.001 uncorrected. Decisions conforming to the social frame caused stronger engagement of the hippocampal formation, precuneus, lateral temporal gyrus, and the dorsomedial PFC. Lower panel. Parameter estimates of the effect size for each decision condition are plotted for the voxel displaying regional peak activation for frame conformity decisions. Numbers on top indicates selected ROI center voxel. Bars correspond to the mean value and error bars indicate the 90% confidence interval of the mean.*

Table 4 – Whole-brain results from brain regions involved in decisions to conform or not to conform to the frame.

Region	Hemi-sphere	Coordinates	Voxels	Z	p
CONFORMITY (conform > non-conform)					
Caudate nucleus	L	-10, 17, 11	2142	6.38	0.001 FWE
Posterior insula	R	42, -14, 18	553	5.89	0.001 FWE
	L	-40, -28, 26	1258	5.79	0.001 FWE
Inferior frontal gyrus	R	32, 30, 16	1455	5.59	0.001 FWE
Occipital cortex	L	-20, -74, 12	565	5.53	0.001 FWE
Superior temporal cortex	R	50, 8, -4	1218	5.30	0.004 FWE
NON-CONFORMITY (non-conform > conform)					
Angular gyrus	L	-52, -74, 18	82	4.86	<0.05 FWE
Inferior frontal cortex	L	-20, 16, -18	73	4.23	<0.001 u.c.
Superior parietal cortex	L	-22, -66, 60	98	4.12	<0.001 u.c.
Temporopolar cortex	R	26, 8, -40	12	3.83	<0.001 u.c.

Abbreviations: FWE = Family-Wise Error correction; u.c. = uncorrected for multiple comparisons. All p-values are reported at cluster-level.

We also ran a post-hoc whole brain analysis to test for differences between the two kinds of frame congruency (congruency–cooperation vs congruency–competition). Here, we found no significant differences, even at a liberal threshold

($p<0.01$, uncorrected).

While framing facilitated frame conformity, subjects sometimes made decisions that violated the frame. This enabled us to test for brain regions showing increased neural activity associated with non-conforming decisions that would violate the expectation of the other player. This exploratory analysis yielded a significant engagement of the left angular gyrus in frame nonconformity as opposed to conformity decisions. Clusters in the inferior frontal cortex, superior parietal cortex, and temporopolar cortex displayed a similar trend towards stronger activation for nonconformity decisions ($p<0.001$, uncorrected; see Table 4).

Since participants in both Study 1 and 2 showed a relatively stronger framing effect in the competition frame than the cooperation frame, one possibility would be that mentalizing activation would be more prominent during the competition than the cooperation phase. To address this, we ran an exploratory analysis of the competition > cooperation conditions during choice. Here, we find that only the right supramarginal gyrus showed a significant difference (26, -30, 32, $Z=3.47$, $p<0.001$ uncorrected, 19 voxels). For the converse contrast (cooperation > competition during the choice phase) no brain region were significantly more engaged, even at a liberal threshold ($p=0.05$).

We also explored the impact of the framing cue on the neural activity elicited at the time that the cue was presented. Relative to the cooperation cue, the competition cue elicited a trend activation (whole-brain $p<0.001$, uncorrected, extent threshold = 5 voxels) in a number of brain regions, including the precuneus, caudate nucleus, anterior cingulate cortex and dorsolateral prefrontal cortex (see Figure S1 and Table 2). In contrast, no voxel in the brain showed stronger neural activation in response to the cooperation cue as opposed to the competition cue.

Discussion

This study provides the first combined behavioral and neurobiological account of the effects of framing in social dilemmas. Besides confirming the effect of framing on PD behavior (Pruitt, 1967; Andreoni, 1995; Deutsch, 2958; Liberman, Samuels & Ross,

2004; Cubitt et al., 2011; Park, 2000) our behavioral data also imply that subjects based their decisions, at least in part, on how they assumed the other actors would act. This finding supports the notion that mentalizing is an important psychological process during decision-making in social dilemmas (Frith & Singer, 2008). Importantly, this effect appeared to lead to different behavioral strategies in the two frame conditions: If the actor (player A) believes that the opposing player B will defect, player A will be compelled to defect. In the cooperation frame, the increased belief that the other player will collaborate prompts two types of responses. Most actors take the frame-congruent choice and follow the generally expected behavior and choose to cooperate. However, others choose to 'free-ride' on the expected cooperative choices made by the other player, and thus, to violate the implicit social expectancy of collaboration.

In agreement with the behavioral data, fMRI revealed a consistent activation of the hippocampal formation and regions known to be in mentalizing when subjects made decisions that conformed to the imposed frame. Besides confirming the significant effect that framing had on cooperation levels during game play, the fMRI results yielded increased activation of the bilateral hippocampal formation and the brain's mentalizing network for frame conform relative to nonconform decisions. In addition, the individual level of empathy was associated with increased activation in the left hippocampal formation during frame-coherent decisions, adding further support to the notion that this structure is significantly related to the effects of framing in social dilemmas. The hippocampal formation is known to play a key role in memory and associative functions (McClure et al., 2004), which suggests that framing effects are related to a stronger engagement of this mnemonic system. Accordingly, neuroimaging studies of decision-making and the influence of contextual information have implicated the hippocampus and surrounding regions (De Martino et al., 2006; Frith & Frith, 2006).

In addition to the hippocampus, brain regions known to be involved in mentalizing, such as the lateral temporal gyrus, precuneus, and the dorsomedial prefrontal cortex showed stronger activation for frame conformity as opposed to

nonconformity. The lateral temporal cortex and precuneus are thought to be particularly important to processing social information, and has been linked to mentalizing and social cognition in many imaging studies (Baumgartner et al., 2011; Mitchell, 2009; Yarkoni et al., 2011). Furthermore, the medial PFC has been implicated in social choice such as altruism, norm compliance and responses to undesirable social actions (Knoch et al, 2006; Knoch et al., 2008; Knoch et al., 2009). Our data extend the role of these structures to include the processing of contextual information and its influence on behavior in social dilemmas, yet further research is needed to disentangle the roles that these regions may have in this process.

When subjects acted in opposition with the frame, there was enhanced engagement of left angular gyrus. Similar trends towards increased activity for frame-nonconform decisions were found in the inferior frontal cortex, superior parietal cortex, and temporopolar cortex. Prior studies have demonstrated a role for the inferior parietal cortex and inferior frontal cortex in social decision-making, including social reasoning (Baumgartner et al., 2011), norm violation (Mitchell, 2009) and reward expectation (Baumgartner et al., 2009), and recent studies have implicated the angular gyrus in social reasoning and moral behavior (Yarkoni et al., 2011). This paper adds to this knowledge by including behavior in which an agent acts in opposition to contextual instructions, and in particular in situations in which the agent stands to benefit by violating expectations in social dilemmas.

Finally, exploratory analyses revealed that the frame cue elicited stronger activation of the precuneus, caudate nucleus, anterior cingulate cortex and dorsolateral prefrontal cortex for competition as opposed to cooperation. Notably, since our inter-stimulus interval was 1 second and non-jittered, this exploratory analysis cannot reliably distinguish between the framing and decision phase, and the results can only be seen as tentative trends in framing effects. Nevertheless, when making the same comparisons during the decision phase, only the right supramarginal gyrus showed a stronger engagement during competition relative to cooperation framing. These observed trends may suggest that the regions already implicated in framing effects (McClure et al, 2004; Deppe et al., 2007), may be engaged already at

the time of framing and before decision options have been presented. Furthermore, these data support previous reports of a role for the right PfC in social decision-making (Knoch et al., 2006) as well as the engagement of anterior cingulate cortex in frame susceptibility (Deppe et al., 2007). The present results expand this knowledge showing that a contextual frame can be established instantly and independent of a presented response option and the type of decision made (i.e., cooperative or deceptive decisions). In contrast, the cooperation instruction was not associated with a specific neural response pattern compared to the competition frame. This suggests that framing a social dilemma as a competition will more strongly activate neural structures that are related to the context dependency of social decisions compared to cooperative framing. One might speculate that this is caused by the competitive frame being considered as being more important or challenging than the cooperation frame. Indeed, recent studies have demonstrated that cooperative behaviors may represent a social "default mode" of decision making in similar conditions (Loewen, Lyle & Nachschen, 2009; Rand, Greene & Nowak, 2012, but see also Rubinstein, 2007 for contradictory findings). By the same token, when expecting competition, subjects deviate from this social default and more strongly engage in expectation and mental calculation and mentalizing.

Taken together, our combined behavioral and neuroimaging data suggest that framing in social dilemmas works by invoking a social mnemonic heuristic where subjects choose their behavioral responses based on how they think their opponents will act. Our exploratory analysis of the framing stage hints at the possibility that such framing effects may occur even before decision options are perceived, although more studies are needed to confirm this assertion. As such, the study illustrates the complexity of decision-making in social dilemmas, in which humans either adhere to contextual cues, or choose to violate the tentative instruction embedded within those cues.

References

Andreoni J (1995) Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. The Quarterly Journal of Economics 110:1-21.

Andreoni J, and Croson R (2008) Partners versus strangers: Random rematching in public goods experiments. Handbook of experimental economics results 1:776-783.

Axelrod R, and Hamilton WD (1981) The evolution of cooperation. Science 211:1390-1396.

Baumgartner T, Fischbacher U, Feierabend A, Lutz K, and Fehr E (2009) The neural circuitry of a broken promise. Neuron 64:756-70.

Baumgartner T, Knoch D, Hotz P, Eisenegger C, and Fehr E (2011) Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat Neurosci 14:1468-74.

Camerer C, and Thaler RH (1995) Anomalies: Ultimatums, Dictators and Manners. The Journal of Economic Perspectives 9:209-219.

Chang, L. J., & Sanfey, A. G. (2011). Great expectations: Neural computations underlying the use of social norms in decision-making. Social Cognitive and Affective Neuroscience. doi:10.1093/scan/nsr094

Cubitt RP, Drouvelis M, Gächter S, and Kabalin R (2011) Moral judgments in social dilemmas: How bad is free riding? Journal of Public Economics 95:253-264.

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

Deppe M et al. (2007) Anterior cingulate reflects susceptibility to framing during attractiveness evaluation. Neuroreport 18:1119-23.

Deutsch M (1958) Trust and suspicion. The Journal of conflict resolution 2:265-279.

Digby P. G. N., Approximating the Tetrachoric Correlation Coefficient, Biometrics, Vol. 39, No. 3, pp. 753-757 Sep., 1983.

Dufwenberg M, Gächter S, and Hennig-Schmidt H (2011) The framing of games and the psychology of play. Games and Economic Behavior 73:459-478.

Eichenbaum H, Yonelinas AP, and Ranganath C (2007) The medial temporal lobe and recognition memory. Annu Rev Neurosci 30:123-52.

Fehr E, and Gintis H (2007) Human Motivation and Social Cooperation: Experimental and Analytical Foundations. Annual Review of Sociology 33:43-64.

Fehr E, and Schmidt KM (1999) A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114:817-868.

Frith CD, and Frith U (2006) The neural basis of mentalizing. Neuron 50:531-4.

Frith CD, and Singer T (2008) The role of social cognition in decision making. Philos Trans R Soc Lond B Biol Sci 363:3875-86.

Gibbons R (2006) A primer in game theory. Prentice-Hall Financial Times, New York.

Iacoboni M et al. (2004) Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. Neuroimage 21:1167-73.

Knoch D, Nitsche MA, Fischbacher U, Eisenegger C, Pascual-Leone A, and Fehr E (2008) Studying the neurobiology of social interaction with transcranial direct current stimulation--the example of punishing unfairness. Cereb Cortex 18:1987-90.

Knoch D, Schneider F, Schunk D, Hohmann M, and Fehr E (2009) Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proc Natl Acad Sci U S A 106:20895-9.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, and Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314:829-32. Liberman V, Samuels SM, and Ross L (2004) The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. Pers Soc Psychol Bull 30:1175-85.

Loewen PJ, Lyle G, and Nachshen JS (2009) An eight-item form of the Empathy Quotient (EQ) and an application to charitable giving. Online publication: http://bit.ly/Ox0JkX

Mann, H. B., and D. R. Whitney. 1947. On a test whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18: 50-60.

McClure SM, Li J, Tomlin D, Cypert KS, Montague LM, and Montague PR (2004) Neural correlates of behavioral preference for culturally familiar drinks. Neuron 44:379-87.

Mitchell, J. P. (2009). Inferences about mental states. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364(1521), 1309-16. doi:10.1098/rstb.2008.0318

Park ES (2000) Warm-glow versus cold-prickle: a further experimental study of framing effects on free-riding. Journal of Economic Behavior & Organization 43:405-421.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, Series 5 50: 157–175.

Pelphrey KA, Morris JP, and McCarthy G (2004) Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci 16:1706-16.

Piovesan, M., & Wengström, E. (2009). Fast or fair? A study of response times. Economics Letters, 105(2), 193-196. doi:10.1016/j.econlet.2009.07.017

Pruitt DG (1967) Reward structure and cooperation: the decomposed Prisoner's Dilemma game. Journal of Personality and Social Psychology; Journal of Personality and Social Psychology 7:21.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. Nature, 489(7416), 427-430. doi:10.1038/nature11467

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. The Economic Journal, 117(523), 1243-1259.

Spitzer M, Fischbacher U, Herrnberger B, Grön G, and Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185-96.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods, 8(8), 665-670..

Zelmer J (2003) Linear public goods experiments: A meta-analysis. Experimental Economics 6, 299-310.