How theory-of-mind brain regions process prediction error across relationship contexts

Journal:	*Social Cognitive and Affective Neuroscience*
Manuscript ID	Draft
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Park, BoKyung; Boston College, Psychology; Fareri, Dominic; Adelphi University, Psychology Delgado, Mauricio; Rutgers University, Psychology Young, Liane; Boston College, Psychology
Keywords:	Impression updating, Motivated cognition, Social prediction error, Theory of mind

How theory-of-mind brain regions process prediction error across relationship contexts

BoKyung Park

Boston College

Dominic Fareri

Adelphi University

Mauricio Delgado

Rutgers University-Newark

Liane Young

Boston College

Word Count: 4935

Address correspondence to:

BoKyung Park

Department of Psychology, Boston College

140 Commonwealth Avenue

Chestnut Hill, MA 02467

Email: parkanj@bc.edu

Phone: 617-552-0240

Overcoming motivated impression updating

Abstract

How do people update their impressions of close others? Although people may be motivated to maintain their positive impressions, they may also update their impressions, when their expectations are violated, leading to prediction error. Combining neuroimaging and computational modeling, we test the hypothesis that brain regions associated with theory of mind (ToM), especially right temporo-parietal junction (rTPJ), underpin both motivated impression maintenance and impression updating evoked by prediction error. Participants had money either given to or taken away from them by a friend or a stranger, and were then asked to rate each partner on trustworthiness and closeness on each trial. Overall, participants engaged in less impression updating for friends versus strangers. Decreased rTPJ activity in response to a friend's negative behavior (taking money) was associated with reduced negative updating and positive ratings of the friend. However, to the extent that participants did update their impressions and deliver more negative ratings of their friend, this behavioral pattern was explained by greater prediction error and greater rTPJ activity. These data suggest that rTPJ recruitment represents the integration of prediction error signals and the capacity to overcome people's motivation to maintain positive impressions of their friends in the face of conflicting evidence.

Keywords: Impression updating; Motivated cognition; Social prediction error; Theory of mind

In everyday life, people update their impressions of others in the face of new information to navigate an ever-changing social environment. However, the processes that guide impression updating may be systematically different for close and distant others. Extensive work demonstrates that people are motivated to positively perceive others who are close to them (Hughes & Beer, 2012; Murray, 1999; Taylor & Brown, 1988). People are more likely to attribute close others' negative behaviors to external causes (Taylor & Koivumaki, 1976), consistent with "motivated cognition"—people's reasoning processes can be shaped by their desire for certain conclusions (Kunda, 1990; Stevens & Fiske, 1995). Related literature on intergroup cognition suggests that motivated impression maintenance might be underpinned by reduced activation of brain regions for processing information about mental states, or theory of mind (ToM). Specifically, failures to negatively update impressions about ingroup members are accompanied by reduced activation in temporo-parietal junction (TPJ), pointing to a process of discounting new negative information about ingroup members to maintain pre-existing positive impressions (Hughes et al., 2017).

Yet, there are instances in which people do in fact update their impressions of close others. Previous research suggests that impression updating occurs when people experience social prediction errors (PE)—differences between what people expect of others and what people actually observe (Koster-Hale & Saxe, 2013). Neuroimaging studies investigating the underlying mechanisms of impression updating have identified a role for ToM regions, suggesting that these regions may help encode prediction error within social domains (Mende-Siedlecki, et al., 2013a; Mende-Siedlecki, et al., 2013b; Thornton & Mitchell, 2018). For example, increased activation in right TPJ (rTPJ) has been observed after participants see people behave in ways that are inconsistent with their initial impressions, with neural activity covarying with the degree of impression updating (Mende-Siedlecki, et al., 2013a). Thus, one role of ToM regions in motivated impression maintenance may involve diminished

representation of social prediction error (Hughes et al., 2017). The present study aims to test this proposal empirically, by combining neuroimaging and computational modeling.

In the current work, participants were asked to bring a friend to the scan session, where they then observed their friend and a stranger perform positive and negative behaviors, i.e., giving money to and taking money from the participant in the context of a modified, iterated dictator game. This approach allowed us to directly compare the processes underlying impression updating of close others (friends) and distant others (strangers). We specifically focused on rTPJ, a key region for ToM (Molenberghs et al., 2016; Saxe, 2009) and social prediction error (Koster-Hale & Saxe, 2013).

We tested four key hypotheses. (1) We expected that people would be motivated to maintain their impressions of their friends, especially when learning new negative information. (2) We predicted this behavioral pattern to be accounted for by *decreased* rTPJ activity in response to friends' negative behavior (taking money), suggesting that people do not integrate this new information. (3) We hypothesized that, to the extent that people do update their impressions of their friends, this pattern could be explained by a social prediction error account. (4) We hypothesized that impression updating and social prediction errors would be supported by *increased* rTPJ activity.

To test the third and fourth hypotheses, we applied a computational modeling approach as suggested in prior work (Kliemann & Adolphs, 2018) for characterizing mechanisms of social learning (Kishida & Montague, 2012; Fareri et al., 2015; Stanley, 2016; Siegel et al., 2018). We constructed a computational model aimed at predicting participants' ratings of the players (friend and stranger) through the interaction of a) their initial impressions of each player and b) the value assigned to the player based on experiences during the game. We expected that participants would differentially update evaluations of friends and strangers based on experienced prediction errors, and that increased rTPJ activity would be associated with greater prediction error.

Method

Participants

Thirty right-handed, neurologically and psychologically healthy native English speakers took part in this study, bringing a close, same-sex friend with them to the scan session[1]. Six participants were excluded from further analyses due to excessive head movement (three participants), a structural abnormality (one participant), an expectation that their friend would take all $20 (one participant), and completing only 25% of the trials (one participant). Participants who completed equal to or over half of all trials, after runs with excessive head movement (> 3mm) were removed, were included in the final sample, leaving 24 participants in total (14 females; age M = 20.00, S.D. = 1.67). All procedures were approved by the Institutional Review Boards at Boston College and the Massachusetts Institute of Technology.

Social judgment task

We created the "Social Judgment Task", a modified version of the Dictator Game (Forsythe et al., 1994; Kahneman et al., 1986). In this game, three people occupy two roles: two "Player 1s" and one "Player 2". Player 1s take turns on each trial and play the same Player 2 (the participant). At the beginning of each trial, both players receive $20, and Player 1 can freely give some money to or take some money from Player 2 in $5 increments. Player 2 passively observes Player 1's decision. Participants were told that the assignment of roles was random, but, in reality, participants always played as Player 2 and observed pre-programmed decisions of Player 1.

After observing how much Player 1 gave or took, participants rated the extent to which they found Player 1 (friend or stranger) on a given trial to be trustworthy ("trustworthiness"), or

[1] To ensure that participants were sufficiently close to their friend, we recruited only participants who chose circle 4 or more on the 7-point Inclusion of Other in the Self (IOS) scale (Aron et al., 1992), representing participants' closeness with their friend (M of the final sample = 5.54, S.D. = 1.02).

how close they felt to Player 1 ("closeness"). Thus, the game matrix consisted of a 2 (Player: Friend, Stranger) X 2 (Valence: Taking, Giving) X 2 (Task: Closeness, Trustworthiness) design, in addition to varying the amount given to and taken from the participant (low [$5, $10], high [$15, $20]). We measured perceived closeness and trustworthiness to cover different dimensions of relationships, i.e., how participants evaluate the relationship, and how participants evaluate an individual's moral character.

At the beginning of each trial, participants viewed the name of Player 1 (2s; friend or stranger) for that trial (Figure 1), followed by the type of rating they would be making (2s; trustworthiness or closeness) and a jittered fixation (2-6s). Participants then observed how much money Player 1 gave to or took from them (2s), followed by another fixation (2-6s), and made a trustworthiness rating or a closeness rating by moving an indicator on an 8-point scale. Trials were divided by a jittered fixation (2-6s). Participants were told that one of the trials would be randomly selected, and all players would receive the amount of money earned on that trial in addition to their base compensation. Pre-registered hypotheses and methods are available at [https://aspredicted.org/blind.php?x=bi5nd2].

Procedure

Participants arrived at the scan session with their friend, where they met a sex-matched confederate (the stranger), posing as another participant in the study. Pre-scan ratings were collected using questions asking all three individuals to evaluate how trustworthy they felt the other two people to be, and how close they felt to each of them. All three people were then presented with the game instructions together. The actual participants were escorted to the scanning area, instructed that they would play as Player 2, and asked to make trustworthiness and closeness ratings of Player 1 (friend or stranger) on each trial, with the values of their ratings kept secret from Player 1. After practicing the game for 8 trials, participants entered the scanner and completed 192 trials of the game (16 trials in each of 12 runs, total time = 74min 24sec) while functional scans were acquired.

After the Social Judgment Task, participants completed two runs of a ToM localizer task (10 trials in each run; total time = 9min 4sec) (Dodell-Feder et al., 2011). On each trial, participants read a vignette (10s) and judged whether a statement was true or false based on the vignette (4s), followed by a fixation (12s). Participants inferred either another person's mental states in the vignette ("belief" condition; e.g., "Lisa now believes that Jacob is sleeping") or physical representations of an object ("photo" condition; e.g., "Today the color of the blouse is white"). Twenty-one participants completed the ToM task (see Supplementary Section 1 for ToM analyses). Finally, participants exited the scanner and completed a post-scan survey not explored in the present paper. Participants were then debriefed and compensated.

While participants were in the scanning area, their friend was escorted to a separate room, given the same instructions as the MRI participants, and played the same game as Player 2 outside the scanner (see Supplementary Section 2 for behavioral responses of participants' friends). Again, as for the MRI participants, there were no real Player 1s, and participants' friends viewed the same pre-programmed behaviors as the participants.

FMRI acquisition and preprocessing

We used a 3T Siemens scanner outfitted with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. Thirty-two 3x3x3mm slices of gradient echo T2* weighted echo-planar images (EPI) provided whole brain coverage (TR = 2s, TE = 30ms, flip angle = 90°) for functional scans. Additionally, high-resolution anatomical scans were acquired (TR = 2.53s, TE = 1.69ms) while participants were looking at a blank screen.

We analyzed brain data using Analysis of Functional Neural Images (AFNI; AFNI_16.2.06 version) software (Cox, 1996). The first six functional scans before the task in each run were removed to compensate for magnet stabilization. All other images were slice-timing corrected (using the first slice as reference), de-obliqued, concatenated across runs, motion corrected (using the third volume as a reference and Fourier interpolation), spatially

smoothed (using a 3D isotropic Gaussian kernel of an 8mm full width at half maximum), normalized by the average activity over the entire task to generate percent signal change (PSC), and high-pass filtered (omitting frequencies < .01Hz, as described in Wu et al. 2014).

VOI analyses. Given that our a priori predictions focused on the associations between rTPJ and impression updating, we conducted volumes-of-interest (VOI) analyses (see Supplementary Section 3 for findings from whole brain analyses). A spherical VOI (radius = 8mm) centered on the coordinates derived from the ToM localizer task (Table S1) in the rTPJ [57, -58, 19] was constructed. We extracted average PSC from this VOI during the phase of the task during which participants observed Player 1's decision for each task condition. Sampling was delayed by 4s to account for the hemodynamic lag to peak (Knutson et al., 2007). Outliers that exceeded three standard deviations from mean activity were deemed likely to be caused by artifacts rather than signal and excluded from further analyses (M = 2.21 trials per each participant, ranging from 0-13). The same patterns remained without excluding any outliers. Additionally, one trial in which participants responded within 50ms in the Social Judgment Task was dropped from further analyses.

Computational modeling analyses

In order to more formally examine the mechanisms underlying impression updating of friends and strangers on a trial-by-trial basis, we employed a computational modeling approach, based on previous work from our group (Fareri et al., 2012; Fareri et al., 2015). We sought to examine specifically whether participants would update their valuation of friends and strangers, with value defined by whether that player gave or took money in the Social Judgment Task; we employed a simple Rescorla-Wagner update rule (Rescorla & Wagner, 1972) to update player value. We then tested whether the value assigned to a player could be used to predict participants' subsequent ratings of friends and strangers using a simple linear function and minimized the sum squared error between our model prediction and participants' actual trial-by-trial ratings.

We formalized five models to test mechanisms of impression updating. Specifically, we examined whether MRI participants: 1) treated all interactions with both players similarly (Baseline model); 2) represented positive (giving money) and negative outcomes (taking money) differently (Loss Gain [LG] Model); 3) represented outcomes as a function of player but not with respect to outcome valence (Friend Stranger [FS] model); 4) represented positive and negative outcomes differently depending on player identity (Loss Gain-Friend Stranger [LGFS] model). In all of these models, we predicted ratings on a trial-by-trial basis using an intercept term and individual slopes for participants' pre-scan ratings and for the updated player value. A fifth and final model (Dynamic LGFS) was identical to the LGFS model, except that we tested the hypothesis that participants' ratings could be predicted with an interaction between participants' pre-scan ratings (initial impressions) and the current value of a player based on experience during the game. We calculated model fits for all models using Bayesian Information Criteria (Akaike, 1974), which strictly penalizes models for increasing numbers of free parameters. Model estimation was performed using tools maintained as part of the Cosanlab Toolbox (https://github.com/ljchang/CosanlabToolbox). Model fits and relevant parameters were compared using non-parametric Wilcoxon Signed Rank tests.

Baseline model. We first formalized a baseline model with 4 free parameters to test the hypothesis that participants would treat all experiences in the task similarly when updating their ratings of the players. We initialized the value V of each player i at 0, and used a simple Rescorla-Wagner prediction error rule to update the player value on each trial (t) after players' decisions to give or take a given amount of money $\boldsymbol{\gamma}$ were revealed. We implemented a single learning rate $\boldsymbol{\alpha}$ in this model for all possible combinations of players and outcomes to reflect participants' not differentiating giving or taking money as a function of player (Eq. 1):

$$\boldsymbol{V_i}(t+1) = \boldsymbol{V_i}(t) + \boldsymbol{\alpha} * (\boldsymbol{\gamma}(t) - \boldsymbol{V_i}(t))$$

where the difference between the experienced outcome $\boldsymbol{\gamma}$ and the value of a player i on a trial t represents prediction error (∂). Initial ratings and player value were then used to separately update ratings of trustworthiness (y) and closeness (j) (Eq. 2):

$$\hat{\boldsymbol{R}}_{iyj} = \boldsymbol{b}_0 + \boldsymbol{b}_1\boldsymbol{R}_{iyj} + \boldsymbol{b}_2\boldsymbol{V}_i$$

LG model. We next formalized a model with 5 free parameters to test the hypothesis that participants' ratings of friends and strangers could be predicted by updating their value from positive (giving) and negative (taking) outcomes separately (Fareri et al., 2012). In this model, the value associated with each player (P_i) was again initialized at 0. We implemented two learning rate parameters here to capture differential updating based on positive or negative social outcomes (y) (Eq. 3):

$$\boldsymbol{V_i}(t+1) = \boldsymbol{V_i}(t) + \boldsymbol{\alpha_y} * (\boldsymbol{\gamma}(t) - \boldsymbol{V_i}(t)).$$

Ratings were then predicted by the same linear model as described in Eq.2.

FS model. To test the hypothesis that participants would update the value of a partner differently, but not differentiate the valence of the outcome, we formalized another model using separate learning rates for each player (i: friend and stranger) but not differentiating outcome valence (Eq. 4):

$$\boldsymbol{V_i}(t+1) = \boldsymbol{V_i}(t) + \boldsymbol{\alpha_i} * (\boldsymbol{\gamma}(t) - \boldsymbol{V_i}(t))$$

LGFS Model. We combined our LG and FS models in order to test the hypothesis that participants would use positive and negative social outcomes to differentially update the value friends and strangers. We formalized a model with 7 free parameters and again set the initial expected value for each player at 0 and allowed for the updating of player value with separate learning rates for each possible combination of player (i) and outcome (y) (Eq. 5):

$$\boldsymbol{V_i}(t+1) = \boldsymbol{V_i}(t) + \boldsymbol{\alpha_{iy}} * (\boldsymbol{\gamma}(t) - \boldsymbol{V_i}(t)).$$

Dynamic LGFS Model. Last, we tested an additional version of our LGFS Model, which differed only in the linear model used to predict participants' ratings on a trial-by-trial basis. We

added an interaction term scaled by an additional free parameter to test the idea that participants' ratings of trustworthiness and closeness would be best predicted as an interactive effect of the pre-scan ratings of each player and the trial-by-trial value computed for each player. Our linear model was thus constructed as follows (Eq. 6):

$$\hat{R}_{iyj} = b_0 + b_1 R_{iyj} + b_2 V_i + b_3 V_i * R_{iyj}$$

We estimated parameters for each participant for all models using the fmincon optimization function in MATLAB in order to minimize the sum squared error between what the model would predict as a participant's trustworthiness or closeness rating of each player ($\hat{R}_{iyj}$) and their actual rating (R_{iyj}) on a given trial (Eq. 7):

$$SSE = (\hat{R}_{iyj} - R_{iyj})^2.$$

Results

Hypothesis 1: Participants update less for friends versus strangers

We conducted a linear mixed-effects regression on participants' trial-by-trial ratings, with Player (friend, stranger), Valence (taking, giving), and Task (closeness, trustworthiness), and the interactions between these factors as fixed effects and individual participants as random effects. Since including Participant Gender and trial-by-trial amount did not change our findings, we dropped them from further analyses (see Supplementary Sections 4-5 for findings with Participant Gender and Amount in the model).

First, we found main effects of Player (b = 1.69, S.E. = .02, t = 72.34, p < .001), Valence (b = .57, S.E. = .02, t = 24.66, p < .001), and Task (b = -.14, S.E. = .02, t = -6.16, p < .001), indicating higher ratings for 1) friend (M = 6.49, S.E. = .10) versus stranger (M = 3.12, S.E. = .10); 2) giving (M = 5.38, S.E. = .10) versus taking (M = 4.23, S.E. = .10); and 3) trustworthiness (M = 4.95, S.E. = .10) versus closeness (M = 4.66, S.E. = .10). A Player X Valence interaction (b = -.10, S.E. = .02, t = -4.45, p < .001) revealed that the difference in

ratings between giving and taking was greater for strangers (M = 1.36, S.E. = .07) than friends (M = .94, S.E. = .07) (Figure 2A). These effects were not modulated by any other factors (See Supplementary Section 6 for other effects).

To further investigate the extent to which participants *changed* their ratings between trials, we subtracted ratings on a given trial from ratings on the previous trial respectively for friend-closeness, friend-trustworthiness, stranger-closeness, and stranger-trustworthiness conditions, taking the absolute value of these scores as an index of trial-by-trial updating. We conducted a linear mixed-effects regression, with Player (friend, stranger), Valence (taking, giving), and Task (closeness, trustworthiness), and the interactions between these factors as fixed effects and individual participants as random effects.

We found main effects of Valence (b = .04, S.E. = .02, t = 2.04, p = .041) and Task (b = -.09, S.E. = .02, t = -5.08, p < .001), indicating that the participants updated more when Player 1 gave money (M = 1.13, S.E. = .19) versus took money (M = 1.05, S.E. = .19), and updated more for trustworthiness (M = 1.18, S.E. = .19) versus closeness (M = 1.00, S.E. = .19). Critically, we found a significant main effect of Player (b = -.20, S.E. = .02, t = -11.63, p < .001), such that the participants updated less for friend (M = .89, S.E. = .19) versus stranger (M = 1.29, S.E. = .19) overall. These effects were not modulated by any other factors.

Hypothesis 2: Less negative ratings in the friend-taking condition were accompanied by decreased rTPJ activity

To examine whether the degree to which participants resisted updating was associated with rTPJ activity, we conducted a linear mixed-effects regression on participants' trial-by-trial ratings with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and their trial-by-trial rTPJ activity, and interactions between these factors as fixed effects, while individual participants were included as random effects.

While reduced rTPJ activity in general was marginally associated with more positive ratings (b = -.09, S.E. = .05, t = -1.86, p = .063), this effect was qualified by a Player X Valence

X rTPJ interaction (b = .09, S.E. = .05, t = 1.81, p = .071). Although the predicted interaction was marginal, because we had a specific pre-registered hypothesis regarding the association between rTPJ activity and ratings during the friend-taking condition (https://aspredicted.org/blind.php?x=bi5nd2; H4a), we conducted follow-up analyses. As predicted, decreased rTPJ was associated with less negative ratings in the friend-taking condition (b = -.22, S.E. = .10, 95% CI = [-.41, -.03]) (Figure 2B). RTPJ activity was not associated with ratings in any other conditions (friend-giving: b = -.07, S.E. = .10, 95% CI = [-.26, .12]; stranger-taking: b = .06, S.E. = 10, 95% CI = [-.13, .25]; stranger-giving: b = -.14, S.E. = .10, 95% CI = [-.32, .05]). These effects were not modulated by other factors (see Supplementary Section 7 for other main effects and interactions).

Hypothesis 3. Impression updates for friends are explained by social prediction error

In order to examine the mechanisms by which updating occurred, we employed a computational modeling approach in which we attempted to predict participants' trial-by-trial ratings as a function of their pre-scan ratings and the trial-by-trial value assigned to a partner based on whether they gave or took money. Non-parametric Wilcoxon signed rank tests demonstrated that participants' ratings were best predicted by our LGSP Dynamic model (BIC = -18.56), which fit participants' data significantly better than our Baseline model (BIC = 71.12; z = 3.6, p < .0005), as well as all other models [LG model (BIC: 73.64; z = 3.63, p <.0005); SP model (BIC: 75.04; z = 3.49, p < .0005); LGSP model (BIC = -1.91; z = 3.54, p <.0005) (Figure 3A)]. We also examined differences in estimated learning rates for updating partner value (Fig. 3B); consistent with results of the mixed-effects analysis above, we found that participants overall exhibited lower learning rates when updating the value of their friends relative to strangers (z = -2.14, p < .02), and, interestingly, a non-parametric repeated measures ANOVA (Friedman test) revealed a significant Player x Valence interaction on learning rates (χ^2 = 10.60, df = 3, p < .02). Durbin-Conover pairwise comparisons indicated that learning rates were higher

for positive ($p < .005$) and negative ($p < .02$) outcomes from strangers, relative to positive outcomes from friends.

Hypothesis 4: Greater prediction error was associated with increased rTPJ activity

To test whether prediction error signal derived from the LGSP Dynamic model accounted for trial-by-trial rTPJ activation during the Social Judgment Task, we conducted a linear mixed-effects regression on participants' trial-by-trial rTPJ activation with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and their trial-by-trial Prediction Error values, and interactions between these factors as fixed effects, while individual participants were included as random effects. A main effect of Player (b = -.03, S.E. = .01, t = -2.55, p = .011) revealed that participants showed overall lower rTPJ activity in response to their friend (M = -.02, S.E. = .02) than to a stranger (M = .04, S.E. = .02). More importantly, there was a significant Player X Valence X Prediction Error interaction (b = .03, S.E. = .01, t = 2.02, p = .043) indicating that more negative prediction error in the friend-taking condition was associated with increased rTPJ activity (b = -.07, S.E. = .03, 95% CI = [-.13, -.01]), while prediction error signal from other conditions did not significantly track rTPJ activity (friend-giving b = -.01, S.E. = .03, 95% CI = [-.06, .04]); stranger-taking (b = .04, S.E. = .03, 95% CI = [-.01, .09]); stranger-giving (b = -.01, S.E. = .03, 95% CI = [-.07, .04]) (Figure 4). This effect was not modulated by any other factors (see Supplementary Section 8 for exploratory whole-brain prediction error analyses).

Discussion

The present research sought to examine impression updating for close friends and strangers, paying special attention to the contributions of motivated cognition and prediction error. By providing participants with the opportunity to revise their evaluations about both a friend and a stranger, observing both partners giving money to them and taking money from them, we found that: (1) participants engaged in less updating overall for friends versus

strangers, reflecting participants' motivation to protect their positive impressions of their friends; (2) this behavioral effect was related to participants' rTPJ activity, such that, when a friend took money, diminished rTPJ recruitment was associated with less negative (and more positive) ratings of friends; (3) participants updated their ratings of their partners using a combination of both their initial impressions of friends and strangers and trial-by-trial experiences with each player, derived from a prediction error update rule; (4) rTPJ activity was associated with prediction error evoked by a friend's negative behavior—in the condition in which a friend took money from the participant, greater prediction error corresponded to greater rTPJ activity.

These findings suggest that people's failure to negatively update their impressions of a close friend is accounted for by reduced prediction error, driven in part by people's positive prior impressions of their friends. Our modeling analyses revealed that impression updating was driven by a linear interaction of prior impressions of friends and strangers, along with the value assigned to a player as updated with separate learning rates for giving and taking money for each player. Consistent with prior work on learning about partner trustworthiness (Chang et al., 2010), this dynamic model of impression updating better captured participants' behavior, compared to simpler models that assumed participants would not account for partner identity, outcome valence, or participants' initial impressions of friends and strangers. Thus, participants' prior impressions of their friend versus the stranger shaped their subsequent evaluations about them over the course of the experiment, contributing to reduced updating for friends versus strangers. When participants overcame this bias to effectively updating their impressions of their friends, this process was driven by enhanced prediction error signaling.

The neuroimaging analyses reveal that rTPJ is a central hub in impression updating, particularly for close others. Over the course of the experiment, reduced rTPJ responses to a friend's taking money were associated with both more positive ratings of the friend and decreased prediction error. These patterns indicate that participants' failure to represent unexpected negative information about a friend contributed to participants' maintenance of their

positive impressions of the friend. Meanwhile, when rTPJ was successfully recruited, participants were more likely to negatively rate their friend, reflecting the contribution of prediction error.

These findings provide a new perspective that may help to make sense of mixed results from previous work on the role of mentalizing regions in moral judgment and impression updating. Specifically, while some studies have found that *increased* activity in ToM regions is associated with favorable evaluation of ingroup members (Baumgartner et al., 2012; Rilling et al., 2008), other studies have found that *decreased* ToM activity is associated with more positive ingroup perception (Dungan et al., 2016; Hughes et al., 2017; Kliemann et al., 2008; Park et al., 2017). We suggest that ToM regions, and especially rTPJ, may function as the hub for coordinating a response based either on motivated cognition or prediction error-based updating; the response may be guided by context, which differs across studies. Some contexts might encourage people to engage in more control-demanding processes to explain away evidence that is inconsistent with one's prior impression (FeldmanHall & Shenhav, 2019; Gershman, 2019), which might be reflected by increased mentalizing, whereas other contexts might motivate people to disregard others' negative behavior, associated with decreased mentalizing (Kim et al., 2019; Park et al., in press).

While we did not have strong hypotheses regarding the associations between rTPJ activity and prediction error signals evoked by friends versus strangers, we found that rTPJ activity was uniquely associated with prediction error during the condition in which a friend took money. Participants may have devoted more mentalizing effort to resolving prediction error in this condition, in attempting to understand their friends' intent behind their taking behavior (Zaki et al., 2016). Interestingly, in spite of the relationship between rTPJ and prediction error in the friend taking condition, we did not observe enhanced learning rates for friends after negative outcomes were revealed, suggesting some impediment to integrating negative prediction error for updating.

In line with extant literature implicating components of the reward circuit (e.g., medial prefrontal cortex, striatum) in encoding prediction error and social learning (Garrison et al., 2013; Fareri et al., in press), we also saw correlates of a social reward prediction error signal that did not differentiate friends and strangers in subgenual anterior cingulate cortex and anterior caudate nucleus (Supplementary Section 8). This pattern is consistent with a role for these regions in encoding social outcome value in trust games (Fareri et al., 2012, 2015; Fouragnan et al., 2013; Vanyukov et al., 2019), and when learning about the generosity of others (Hackel et al., 2015). Thus, it is possible that reward-related regions encode a more generic prediction error signal across social and non-social contexts. However, one limitation of our work and these studies is that social outcomes are conflated with monetary outcomes, and so one possibility for future work is exploring whether prediction errors evoked by friends and strangers are processed in dissociable networks from monetary outcomes (c.f., Behrens et al., 2008; Stanley 2016).

Forming, maintaining, and revising impressions of other agents is critical for social interaction. Building on literature on motivated and intergroup cognition, we present initial findings on the underlying neural processes and computational factors that support motivated impression maintenance as well as impression updating, for both close and distant others. By enhancing our understanding of these mechanisms, this research can serve as the groundwork for supporting accurate person perception and perhaps leveraging motivated cognition to strengthen existing relationships.

Funding

This work was supported by a grant from the John Templeton Foundation [5107321] awarded to Liane Young.

Acknowledgements

We thank E. Alai and M. Kronitz for their research assistance; M. Kim, J. Hirschfeld-Kroen, and K. Jiang for their comments; and members of the Boston College Morality Lab for feedback on earlier versions of this manuscript.

References

Akaike H. (1974) *A New Look at the Statistical Model Identification.* In: Parzen E., Tanabe K., Kitagawa G. (Eds). *Selected Papers of Hirotugu Akaike*. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY

Aron, A., Aron E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596-612.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*, 1452-1469.

Behrens, T. K. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*, 245-249.

Chang, L.J., Doll, B., B., van t'Wout, M., Frank, J., M., & Sanfey, A., G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2): 87-105.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162–73.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage, 55*, 705-712.

Dungan, J. A., Stepanovic, M., & Young, L. (2016). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience, 11*(8), 1183-1192.

Fareri, D. S., Chang, L., J., & Delgado, M., R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fareri*, D. S., Chang*, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience, 35*(21), 8170-8180.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (in press). Neural mechanisms of social learning.

In M. S. Gazzaniga, G. R. Mangun, and D. Poeppel (Eds.). *The Cognitive eurosciences*. MIT Press.

FeldmanHall, O. & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour, 3*, 426-435.

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*(3), 347–69.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience, 33*(8), 3602-3611.

Garrison, J., Erdeniz, B., & Done, J. (2013) Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews. 37*(7), 1297-1310.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin and Review, 26*(1), 13-28.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience* 18, 1233–1235. doi:10.1038/nn.4080.

Hughes, B. L. & Beer, J., S. (2011). Orbitofrontal Cortex and Anterior Cingulate Cortex are modulated by motivated social cognition. *Cerebral Cortex, 22*, 1372-1381.

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience, 12*(1), 49-60.

Kahneman, D., Knetsch, J.L., Thaler, R. (1986). Fairness as a constraint on profit seeking: entitlements in the market. *The American Economic Review, 76*(4), 728–41.

Kim, M., Park, B., & Young, L. (2019). *The psychology of rational versus non-rational belief updating*. Manuscript invited, Trends in Cognitive Science.

Kishida, K. T. & Montague, P. R. (2012). Imaging models of valuation during social interaction in

humans. Biological Psychiatry, 72(2), 93-100.

Kliemann, D. & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology, 24*, 1-6.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*, 2949-2957.

Knutson, B., Rick, S., Wimmer, G.E., Prelec, D., Loewenstein, G. (2007). Neural predictors of purchases. *Neuron, 53*(1), 147–56.

Koster-Hale, J. & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836-848.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-198.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013a). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *The Journal of Neuroscience, 33*(50), 19406-19415.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013b). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623-631.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews, 65*, 276-291.

Murray, S. L. (1999). The quest for conviction: motivated cognition in romantic relationships. *Psychological Inquiry, 10*(1), 23-34.

Park, B., Blevins, E., Knutson, B., & Tsai, J. L. (2017). Neurocultural evidence that ideal affect match promotes giving. *Social Cognitive and Affective Neuroscience, 12*(7), 1083-1096.

Park†, B., Kim†, M., & Young, L. (in press). An examination of accurate versus “biased” mentalizing in moral and economic decision-making. In Gilead, M. & Ochsner, K. N. (Eds). *The Neural Basis of Mentalizing.* New York: Springer.

Rescorla, R, & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the

effectiveness of reinforcment and non reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64-99). New York: Appleton-Century-Crofts.

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *Neuroimage, 41*, 1447-1461.

Saxe, R. (2009). Theory of mind (neural basis). In Banks, W. (Ed.), *Encyclopedia of Consciousness*. Cambridge, MA: MIT Press.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*, 750-756.

Stanley, D. A. (2016). Getting to know you: General and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience, 11*(4), 525–536. doi:10.1093/scan/nsv145.

Stevens, L. & Fiske, S. (1995). Motivation and cognition in social life: A social survival perspective. *Social Cognition, 13*(3), 189-214.

Taylor, S. E. & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin,103*(2), 193-210.

Taylor, S.E. & Koivumaki, J. H. (1976). The perception of self and others: acquaintanceship, affect, and actor-observer differences. *Journal of Personality and Social Psychology, 33*(4), 403-408.

Thornton, M. A. & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*, 3505-3520.

Vanyukov, P. M., Hallquist, M. N., Delgado, M., Szanto, K., & Dombrovski, A. Y. (2019). Neurocomputational mechanisms of adaptive learning in social exchanges. *Cognitive, Affective, & Behavioral Neuroscience*, 1-13. doi:10.3758/s13415-019-00697-0.

Wu, C.C., Samanez-Larkin, G.R., Katovich, K., Knutson, B. (2014). Affective traits link to reliable neural markers of incentive anticipation. *NeuroImage, 84*, 279–89.

Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K., & Shohamy, D. (2016). Social Cognition as Reinforcement Learning: Feedback Modulates Emotion Inference. *Journal of Cognitive Neuroscience, 28*(9), 1270–1282. doi:10.1162/jocn_a_00978.

**Equal contribution; †Co-correspondence*

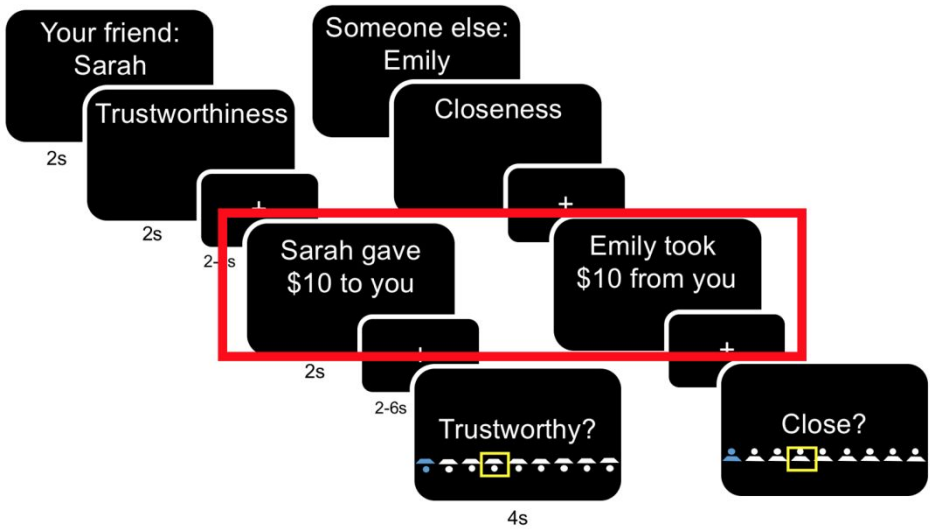

Figure 1. *Representative trials of Social Judgment Task.* Participants viewed the name of Player 1 (2s); the rating that they would make in the trial (2s); and a jittered fixation cross (2-6s). Then the decision of the Player 1 appeared on the screen (2s). After a jittered fixation cross (2-6s), participants were allowed to make their ratings (4s). Each trial was divided by another jittered fixation cross (2-6s). For fMRI analyses, we focused on the phase when participants viewed ostensible Player 1's decision, marked in the red box.

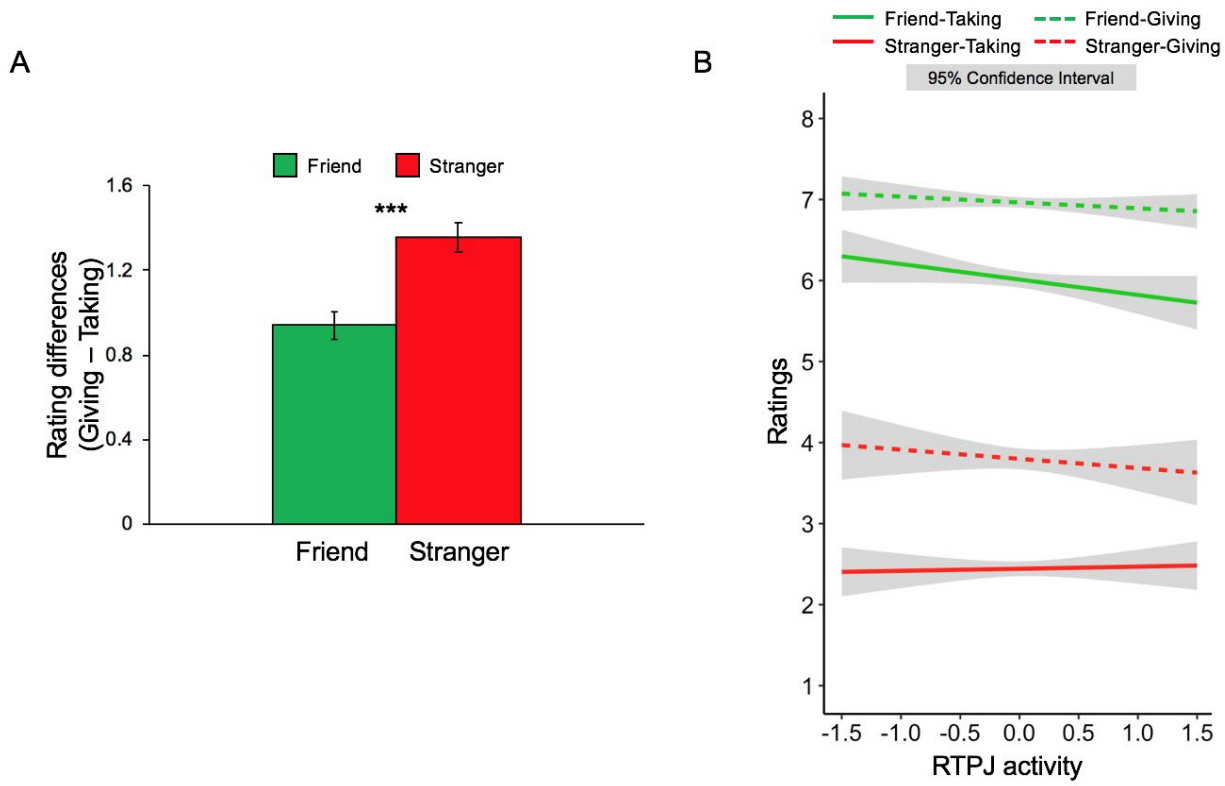

Figure 2. (*A) Differences in participants' ratings during giving versus taking conditions.* Participants differentiated giving and taking conditions more for a stranger than for their friend. ***$p < .001$. *(B) Association between rTPJ activity and social judgment ratings*. Participants' trial-by-trial rTPJ activity significantly tracked ratings when their friend took money from them. The greater rTPJ activity participants showed in response to their friend's taking behavior, the lower ratings they ended up giving to their friend in the given trial.

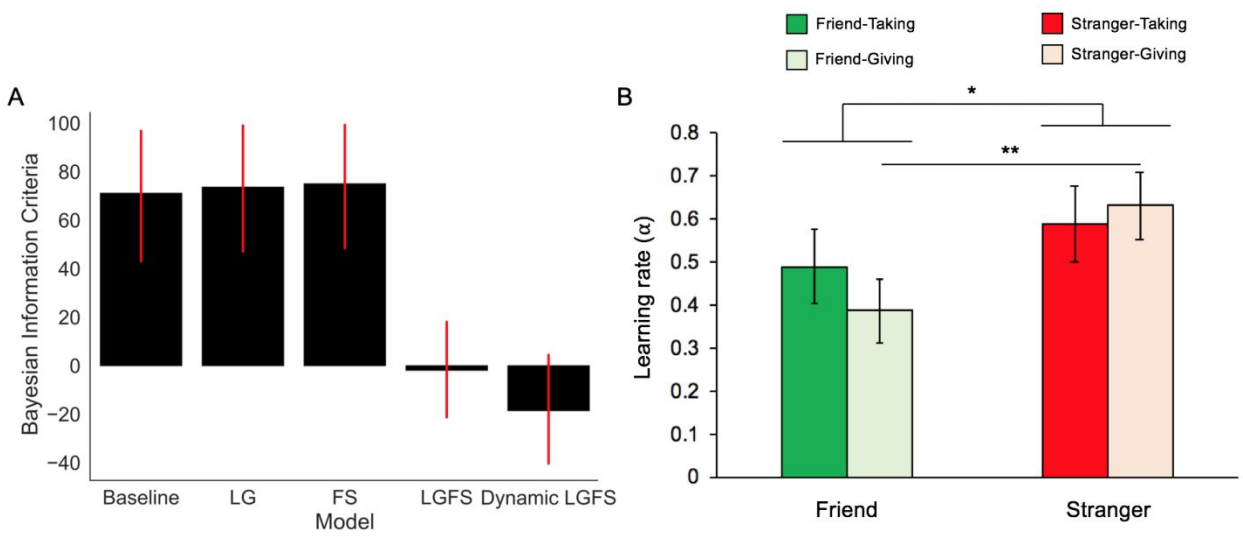

Figure 3. (A) *Bayesian Information Criteria (BIC) values for models.* From left, BIC values of Baseline model, Loss Gain (LG) model, Friend Stranger (FS) model, Loss Gain-Friend Stranger (LGFS) model, and Dynamic Loss Gain-Friend Stranger (Dynamic LGFS) model. (B) *Learning rate (α) values divided by conditions*. Participants showed lower learning rates in the friend condition than in the stranger condition overall. Pairwise comparisons revealed that it was driven by participants' lower learning rates in the friend-giving condition than in the stranger-giving condition. *$p < .05$, **$p < .01$.

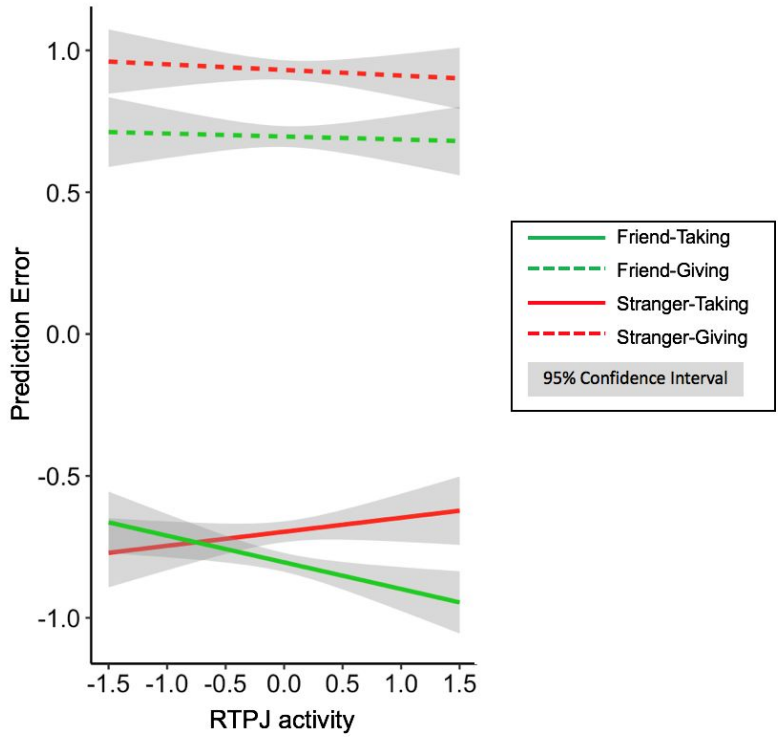

Figure 4. *Association between rTPJ activity and prediction error*. Participants' trial-by-trial rTPJ activity was significantly associated with their prediction error signal evoked when their friend took money from them. The more negative the prediction error participants experienced in response to their friend's taking behavior, the greater rTPJ activity they showed.

Supplementary Materials for

How theory-of-mind brain regions process prediction error across relationship contexts

Park, B., Fareri, D., Delgado, M., and Young, L.

1. Theory-of-Mind (ToM) analyses and results
2. Behavioral responses of participants' friends
3. Whole-brain analyses
4. Findings when Participant Gender was entered in the model
5. Findings when Amount was entered in the model
6. Additional findings from the linear mixed-effects model examining the influence of conditions on participants' responses
7. Additional findings from the linear mixed-effects model examining the influence of rTPJ activity on participants' responses
8. Whole-brain analyses with prediction error signal

Supplementary Section 1. Theory-of-Mind (ToM) analyses and results

For the functional scans acquired during the ToM task, we conducted the analyses on the time points when participants looked at the vignettes and answered the questions. We constructed a GLM model including two regressors of interest: (1) marking the time points of each trial when participants were presented with the vignette and the question, and (2) contrasting the belief trial (+1) and the photo trial (-1). Eight regressors of no interest, sampling white matter activity, cerebrospinal fluid activity, and six head movement regressors, were also included. Other procedures were the same in the Social Judgment Task analyses (see Supplementary Section 3). Consistent with the previous literature (Decety & Cacioppo, 2012; Saxe, 2009; Saxe & Powell, 2006; Saxe, Carey, & Kanwisher, 2004; Saxe & Kanwisher, 2003; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004; Saxe & Wexler, 2005), participants showed robust activation in their ToM network in response to the belief condition compared to the photo condition (Table S1), including in bilateral TPJ, dmPFC, and precuneus.

Table S1. Activation in response to Theory of Mind task, belief vs. photo

Region	x	y	z	Peak Z	Voxels
LTPJ	-46	-64	22	6.95	1434
R Superior temporal gyrus/ RTPJ	57	-58	19	5.82	1199
L Superior frontal gyrus/ L dmPFC	-7	47	37	5.64	918
L Precuneus	-4	-58	28	6.28	572
L Middle frontal gyrus	-37	35	24	-5.13	292
R Middle frontal gyrus	44	35	24	-4.95	111
L Inferior temporal gyrus	-56	-50	-18	-4.84	87

R Medial frontal gyrus	5	52	-2	4.38	67
R Precuneus	29	-65	41	-4.20	51
L Precuneus	-22	-68	35	-4.23	41

Note. Uncorrected $p < .001$, $k > 33$, corrected $p < .05$, regions of interest in bold. R; right, L; left, TPJ; temporo-parietal junction, dmPFC; dorsomedial prefrontal cortex.

Supplementary section 2. Behavioral responses of participants' friends

While participants completed the scanning part of the study, their friends were escorted to a separate place and informed that they would play as the Player 2 in the game. Thus, all scan participants and friends actually played as the Player 2 despite having been told, incorrectly, that their friend and the stranger would play as the Player 1s.

The friends looked at the same pre-determined decisions of the ostensible Player 1s. The pseudo-randomized paradigm was yoked between the MRI participants and their friends that they came to the lab with; thus, they saw the same decisions of Player 1s in the same order.

We conducted a linear mixed-effects regression on the friends' trial-by-trial ratings during the Social Judgment Task, with Player (MRI participant, stranger), Valence (taking, giving), Task (closeness, trustworthiness) as fixed effects and the individual friends as random effects. Significant main effects of Player (b = 1.83, S.E. = .02, t = 84.11, p < .001), Valence (b = .60, S.E. = .02, t = 27.34, p < .001), and Task (b = -.26, S.E. = .02, t = -11.79, p < .001) revealed that the friends gave higher ratings 1) to the scan participants (M = 6.79, S.E. = .12) than to strangers (M = 3.13, S.E. = .12); 2) in the giving trials (M = 5.55, S.E. = .12) than in the taking trials (M = 4.36, S.E. = .12); and 3) on the trustworthiness scale (M = 5.22, S.E. = .12) than on the closeness scale (M = 4.70, S.E. = .12).

These main effects were qualified by different interaction effects. First, a significant Player X Task interaction effect (b = .32, S.E. = .02, t = 14.59, p < .001) revealed that the friends gave higher closeness ratings (M = 6.85, S.E. = .13) to the MRI participants than trustworthiness ratings (M = 6.73, S.E. = .13), p = .048, while they gave higher trustworthiness ratings (M = 3.70, S.E. = .13) to the stranger than closeness ratings (M = 2.55, S.E. = .13), p < .001. There was also a significant Player X Valence interaction (b = -.07, S.E. = .02, t = -3.19, p = .001) indicating that they differentiated the taking vs. giving conditions slightly more for the

stranger (taking M = 2.46, S.E. = .13; giving M = 3.79, S.E. = .13; $p < .001$) than for the MRI participants (taking M = 6.27, S.E. = .13; giving M = 7.32, S.E. = .13; $p < .001$).[1]

Additionally, we examined if trial-by-trial ratings made by MRI participants and their friend during the Social Judgment Task were also correlated to each other, and if they could be modified as a function of their pre-scan closeness. We ran a linear mixed-effects regression on the friends' trial-by-trial ratings, entering the MRI participants' trial-by-trial responses and the pre-scan closeness ratings the MRI participants gave to their friends (ranging from 6 to 8 out of an 8-point scale) as the fixed effects.[2] Individual pair identification codes (pairs of each MRI participant and each friend) were included as mixed effects. There was a significant main effect of MRI participants' responses (b = .31, S.E. = .03, $t = 9.78$, $p < .001$), indicating that MRI participants' responses were significantly related to the friends' responses in the Social Judgment Task, which was not surprising given that they saw the same predetermined Player 1's decisions. Compared to friends who received a 6 in the pre-scan closeness evaluation from their paired MRI participants, those who received a 7 or 8 gave lower ratings in general (pre-scan closeness 7: b = -1.51, S.E. = .31, $t = -4.83$, $p < .001$; pre-scan closeness 8: b = -1.32, S.E. = .36, $t = -3.65$, $p = .002$). More importantly, these effects were qualified by the significant interactions between MRI participants' responses and their pre-scan closeness ratings for their friends (Response X Pre-scan closeness 7: b = .41, S.E. = .04, $t = 11.42$, $p < .001$; Response X Pre-scan closeness 8: b = .36, S.E. = .04, $t = 8.76$, $p < .001$; compared to the baseline, Pre-scan closeness 6). Although all MRI participants' responses, regardless of their pre-scan closeness ratings for their friend, were significantly and positively correlated with the friends' ratings during the Social Judgment Task, responses of the MRI participants who gave 6 to their

[1] Additionally, a Valence X Task interaction, b = -.06, S.E. = .02, $t = -2.96$, $p = .003$, showed that the friends differentiated closeness and trustworthiness ratings more in the giving condition (closeness M = 5.23, S.E. = .13; trustworthiness M = 5.88, S.E. = .13; $p < .001$) than in the taking condition (closeness M = 4.17, S.E. = .13; trustworthiness M = 4.55, S.E. = .13, $p < .001$)

[2] Spearman's correlation tests revealed that the pre-scan closeness ratings for individuals' paired friend were significantly correlated with each other, $r_s = .53$, $p = .008$, indicating that the closer the MRI participants rated their friend to them initially, the closer their friend rated the MRI participants to them as well.

friend were less tightly associated with their friends' ratings (b = .31, S.E. = .03, 95% CI = [.25, .37]), compared to the responses of the MRI participants who gave 7 (b = .72, S.E. = .02, 95% CI = [.69, .76]) and those who gave 8 (b = .67, S.E. = .03, 95% CI = [.62, .72]) (Figure S1).

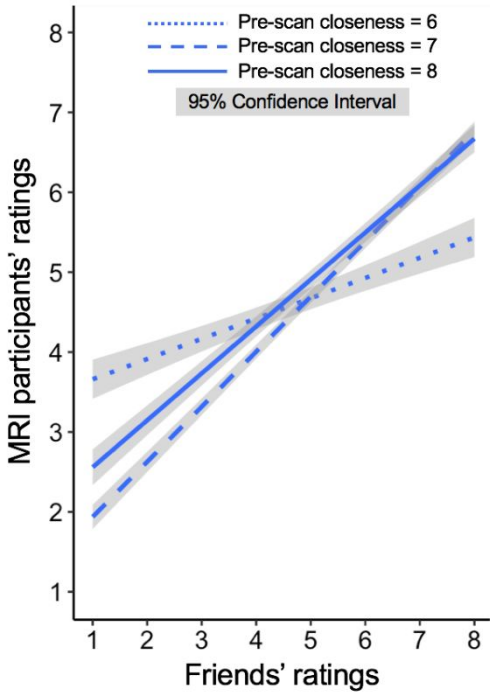

Figure S1. Associations between the MRI participants' ratings and their friends' ratings during the Social Judgment Task, divided by the MRI participants' pre-scan closeness ratings that they gave to their friends.

Supplementary section 3. Whole-brain analyses

a. Methods

We conducted a whole-brain analysis on the functional scans acquired during the Social Judgment Task. To simplify the model, amount (that Player 1 gave or took) was binned into two categories: low ($5, $10) and high ($15, $20). However, analyses with the original four levels of amount yielded similar results.

Primary analyses were conducted on the time points when participants viewed the decision of Player 1 (see Figure 1 in the main manuscript). We constructed a general linear model (GLM, ordinary least-squares regression) including thirteen regressors of interest. The first regressor marked the time points of each trial when participants observed the decision of Player 1. Other regressors marked: (1) Player 1 (Player: Friend = +1, Stranger = -1), (2) The type of rating participants were asked to make (Task: Closeness = +1, Trustworthiness = -1), (3) Valence of the decision (Valence: Giving = +1, Taking = -1), (4) The amount Player 1 gave or took (low = 1, high = 2), and the interactions between (5) Player X Task, (6) Player X Valence, (7) Player X Amount, (8) Task X Amount, (9) Valence X Amount, (10) Player X Task X Valence, (11) Player X Task X Amount, and (12) Player X Valence X Amount. To minimize the influence of physiological confounds, eight regressors of no interest were also included: six modeling head movement, one sampling white matter activity, and one sampling cerebrospinal fluid activity (Chang & Glover, 2009). Before they were submitted to the model, regressors of interest were convolved with a canonical gamma variate hemodynamic delay (Cohen, 1997). Linear regression t-statistic maps were converted to Z-scores, coregistered with structural maps, spatially normalized by warping to Montreal Neurological Institute space (linear to colin27T1_seg template), and resampled as 3mm cubic voxels.

A one-sample t-test was conducted with AFNI program 3dttest function to examine the group-level neural responses to each contrast. This t-test map was initially voxelwise thresholded, at $p < .001$, and then cluster thresholded, cluster size > 35 continuous $3mm^3$

voxels, to yield corrected maps for detecting whole-brain activity at $p < .05$ corrected. Cluster correction was performed using 3dClustSim as implemented in AFNI_16.2.06. We computed the smoothness of the residuals of participants' data at the single subject level using 3dFWHMx, implementing the spatial autocorrelation function, and used these smoothness estimates as inputs into 3dClustSim with 10000 iterations.

b. Findings

We found decreased activity in rTPJ in response to the Player X Valence X Amount regressor (Table S2; Figure S2A). This effect suggests that RTPJ activity was greater when friends took larger amounts and/or when strangers gave larger amounts, consistent with the social prediction error account (Koster-Hale & Saxe, 2013). A Player X Valence X Task X Amount linear mixed-effects model on rTPJ percent signal change (PSC) confirmed these findings, revealing a significant Player X Valence X Amount interaction (b = -.02, S.E. = .01, t = -2.70, p = .007). Participants showed increased RTPJ activity when their friend took larger amounts (M = .03, S.E. = .02) than smaller amounts (M = -.05, S.E. = .02), p = .008. Participants did not differentiate the amount that friends gave or that strangers gave or took, ps > .24. (Figure S2B).

Table S2. Activation in response to the Social Judgment Task

Contrast	Region	x	y	z	Peak Z	Voxels
Player	No voxels survived					
Task	No voxels survived					
Valence	No voxels survived					
Amount						
	L dmPFC	-7	31	53	5.37	975
	(L Superior frontal gyrus)					

	L Cingulate gyrus	-4	-52	26	4.72	308
	L Inferior frontal gyrus (extended into L insula)	-34	18	-7	4.66	301
	R Thalamus	8	-26	4	4.86	248
	R Inferior frontal gyrus (extended into R insula)	35	30	-3	4.70	236
	RTPJ	44	-68	35	5.28	231
	R Middle temporal gyrus	57	-38	-7	3.90	99
	LTPJ	-46	-61	32	4.09	92
	R Middle frontal gyrus	38	6	42	4.12	57
	L Cingulate gyrus	-4	-18	34	4.15	39
Player X Task	No voxels survived					
Player X Valence						
	R Superior temporal gyrus/RTPJ	57	-39	20	4.18	132
	R Middle temporal gyrus	57	-57	2	3.95	60
Player X Amount	No voxels survived					
Task X Amount	No voxels survived					
Valence X Amount						
	R Lingual gyrus	11	-87	-17	-4.38	86
	R Fusiform Gyrus	47	-62	-26	-4.06	69
Player X Task X Amount						
	R Cingulate gyrus	8	-6	44	-4.01	205
	R Postcentral gyrus	26	-32	69	-4.19	116

	L Middle frontal gyrus	-22	-17	67	-3.91	93
Player X Valence X Amount						
	R Inferior parietal lobule/ RTPJ	57	-39	23	-4.24	45
Player X Task X Valence	No voxels survived					

Note. Uncorrected $p < .001$, cluster > 35 continuous voxels, corrected $p < .05$, regions of interest in bold. R; right, L; left, TPJ; temporo-parietal junction, dmPFC; dorsomedial prefrontal cortex.

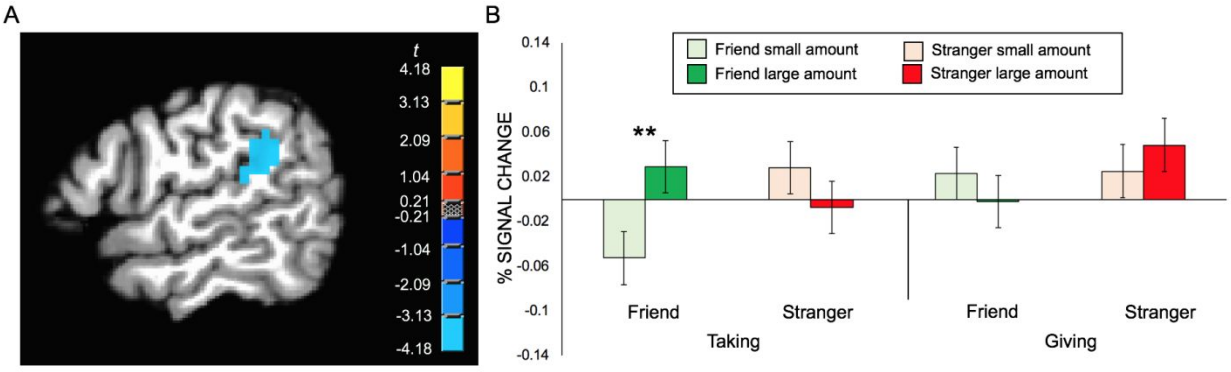

Figure S2. *RTPJ activity varied between conditions.* (A) Whole-brain analysis revealed that participants showed decreased rTPJ activity in response to the interaction between player, valence, and amount. $p < .001$ uncorrected, cluster > 35 continuous voxels, $p < .05$ corrected. (B) Subsequent PSC analyses showed that it was driven by participants' decreased activity in response to their friend's taking behavior, especially when their friend took smaller amounts. $^{**}p < .01$.

Supplementary section 4. Findings when Participant Gender was entered in the model

a) Effect of conditions on ratings

We conducted a linear mixed-effects regression on participants' trial-by-trial ratings, with Participant Gender (male, female), Player (friend, stranger), Valence (taking, giving), and Task (closeness, trustworthiness), and the interactions between these factors as fixed effects and individual participants as random effects.

First, we found main effects of Player (b = 1.67, S.E. = .02, t = 72.36, p < .001), Valence (b = .59, S.E. = .02, t = 25.41, p < .001), and Task (b = -.13, S.E. = .02, t = -5.63, p < .001), indicating higher ratings for 1) friend (M = 6.49, S.E. = .11) versus stranger (M = 3.14, S.E. = .11); 2) giving (M = 5.41, S.E. = .11) versus taking (M = 4.23, S.E. = .11); and 3) trustworthiness (M = 4.95, S.E. = .11) versus closeness (M = 4.69, S.E. = .11). A Player X Valence interaction (b = -.10, S.E. = .02, t = -4.53, p < .001) revealed that the difference in ratings between giving and taking was greater for strangers (M = 1.39, S.E. = .07) than friends (M = .97, S.E. = .07).

Participant Gender qualified the main effects of Player (b = .09, S.E. = .02, t = 3.90, p < .001), Valence (b = -.12, S.E. = .02, t = -5.25, p < .001), and Task (b = -.12, S.E. = .02, t = -5.06, p < .001), indicating that females differentiated their friend (M = 6.52, S.E. = .14) and the stranger (M = 2.99, S.E. = .14) more than males (friend M = 6.47, S.E. = .16; stranger M = 3.30, S.E. = .16), while males differentiated taking (M = 4.18, S.E. = .16) and giving (M = 5.59, S.E. = .16) conditions more than females (taking M = 4.29, S.E. = .14; giving M = 5.22, S.E. = .14). Moreover, females gave greater ratings in the trustworthiness condition (M = 5.00, S.E. = .14) than in the closeness condition (M = 4.50, S.E. = .14), p < .001, while males did not differentiate the rating tasks (trustworthiness M = 4.90, S.E. = .16; closeness M = 4.87, S.E. = .16, p = .704). However, regardless of the Participant Gender, the effects were in the same general direction

as the main effects; participants gave more positive ratings to their friend than to the stranger, and in the giving condition than in the taking condition, regardless of their gender.

Additionally, there was a significant interaction between Player X Task (b = .30, S.E. = .02, t = 13.04, p < .001), showing that participants gave more positive ratings to the stranger in the trustworthiness condition (M = 3.58, S.E. = .11) than in the closeness condition (M = 2.71, S.E. = .11), while they gave more positive ratings to their friend in the closeness condition (M = 6.66, S.E. = .11) than in the trustworthiness condition (M = 6.32, S.E. = .11), ps < .001. This effect was modulated by Participant Gender (b = .15, S.E. = .02, t = 6.28, p < .001), manifested more with female participants (friend closeness M = 6.72, S.E. =.15; friend trustworthiness M = 6.32, S.E. = .15; stranger closeness M = 2.29, S.E. = .15; stranger trustworthiness M = 3.68, S.E. = .15; closeness versus trustworthiness comparisons ps < .001) than with male participants (friend closeness M = 6.61, S.E. = .17; friend trustworthiness M = 6.33, S.E. = .17; p = .003; stranger closeness M = 3.13, S.E. = .17; stranger trustworthiness M = 3.47, S.E. = .17; p = .001). However, the effects were in the same direction for both male and female participants.

Lastly, there was a significant Valence X Task interaction (b = -.05, S.E. = .02, t = -2.31, p = .021). Participants differentiated the closeness and trustworthiness ratings more in the giving condition (closeness M = 5.22, S.E. = .11; trustworthiness M = 5.59, S.E. = .11, p < .001) than in the taking condition (closeness M = 4.15, S.E. = .11; trustworthiness M = 4.31, S.E. = .11, p = .019).

b) Effect of rTPJ activity on ratings

We conducted a linear mixed-effects regression on participants' trial-by-trial ratings with Participant Gender (male, female), Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and their trial-by-trial rTPJ activity, and interactions between these factors as fixed effects, while individual participants were included as random effects.

While reduced rTPJ activity in general was associated with more positive ratings (b = -.10, S.E. = .05, t = -2.05, p = .041), a Player X Valence X rTPJ interaction (b = .10, S.E. = .05, t

= 2.05, p = .040) revealed this effect to be more pronounced in the friend-taking condition (b = -.20, S.E. = .10, 95% CI = [-.39, -.02]), consistent with the main paper: reduced rTPJ activity in response to the friend's negative behavior (i.e., taking money) tracked with less negative (more positive) ratings of the friend. Although unexpected, we found that greater rTPJ activity was also associated with more negative ratings when the stranger gave money (b = -.19, S.E. = .10, 95% CI = [-.37, -.001]).[3] RTPJ activity was not associated with ratings of the stranger in the taking condition (b = .07, S.E. = .09, 95% CI = [-.12, .25]) or ratings of their friend in the giving condition (b = -.07, S.E. = .10, 95% CI = [-.25, .12]). These effects were not modulated by any other factors.

Additionally, there were significant main effects of Player (b = 1.67, S.E. = .02, t = 71.99, p < .001), Valence (b = .59, S.E. = .02, t = 25.47, p < .001), and Task (b = -.13, S.E. = .02, t = -5.72, p < .001); participants gave higher ratings 1) to their friend (M = 6.49, S.E. = .11) than to a stranger (M = 3.15, S.E. = .11); 2) in the giving condition (M = 5.41, S.E. = .11) than in the taking condition (M = 4.23, S.E. = .11); and 3) on the trustworthiness scale (M = 4.95, S.E. = .11) than on the closeness scale (M = 4.69, S.E. = .11).[4]

These main effects were qualified by a couple of interactions. First, a significant Valence X Task interaction (b = -.06, S.E. = .02, t = -2.37, p = .018) revealed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.22, S.E. = .11; trustworthiness M = 5.60, S.E. = .11, p < .001) than in the taking condition

[3] However, given that this effect was not predicted in advance, and the association between rTPJ activity and ratings in the stranger-giving condition became non-significant without Participant Gender in the model, we would not interpret this effect with too much attention.

[4] These main effects were modulated by Participant Gender. Participant Gender X Player (b = .09, S.E. = .02, t = 3.91, p < .001) interaction indicated that females differentiated their friend (M = 6.51, S.E. = .14) and the stranger (M = 2.99, S.E. = .14) more than males (friend M = 6.47, S.E. = .16; stranger M = 3.31, S.E. = .16); Participant Gender X Valence (b = -.12, S.E. = .02, t = -5.32, p < .001) interaction showed that males gave greater ratings in the giving condition (M = 5.60, S.E. = .16) than females (M = 5.22, S.E. = .14), although this difference was only directional, p = .085; Participant Gender X Task (b = -.12, S.E. = .02, t = -5.04, p < .001) interaction revealed that females differentiated closeness and trustworthiness ratings more (closeness M = 4.50, S.E. = .14; trustworthiness M = 5.00, S.E. = .14, p < .001) than males (closeness M = 4.88, S.E. = .16; trustworthiness M = 4.91, S.E. = .16, p = .653). However, regardless of Participant Gender, the directions of ratings were the same in general.

(closeness M = 4.15, S.E. = .11; trustworthiness M = 4.31, S.E. = .11, p = .016). Additionally, there was a significant Valence X Task X rTPJ activity interaction (b = .10, S.E. = .05, t = 2.06, p = .039) revealing that rTPJ activity particularly tracked the trustworthiness ratings (b = -.23, S.E. = .10, 95% CI = [-.42, -.04]) in the giving condition. RTPJ activity was not associated with other ratings (closeness in giving condition: b = -.02, S.E. = .09, 95% CI = [-.21, .16]; trustworthiness in taking condition: b = .02, S.E. = .09, 95% CI = [-.16, .21]; closeness in taking condition: b = -.16, S.E. = .10, 95% CI = [-.35, .03]). Additionally, an interaction between Player X Task (b = .30, S.E. = .02, t = 13.09, p < .001) showed that participants gave higher ratings to friend-closeness (M = 6.66, S.E. = .11) than to friend-trustworthiness (M = 6.32, S.E. = .11), while they gave greater ratings to stranger-trustworthiness (M = 3.59, S.E. = .11) than to stranger-closeness (M = 2.71, S.E. = .11), ps < .001.[5] Finally, there was a significant Player X Valence interaction (b = -.11, S.E. = .02, t = -4.61, p < .001), indicating that participants differentiated between the taking and giving conditions more for the stranger (taking M = 2.45, S.E. = .11; giving M = 3.85, S.E. = .11) than for their friend (taking M = 6.01, S.E. = .11; giving M = 6.98, S.E. = .11), ps < .001. However, as reported in the main paper, this effect was qualified by a significant Player X Valence X rTPJ interaction.

c) Effect of prediction error on rTPJ

We conducted a linear mixed-effects regression on participants' trial-by-trial rTPJ activation with Participant Gender (male, female), Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and their trial-by-trial Prediction Error values, and interactions between these factors as fixed effects, while individual participants were included

[5] This effect was further qualified by a significant Participant Gender X Player X Task interaction (b = .14, S.E. = .02, t = 6.11, p < .001). Females gave significantly lower closeness ratings to the stranger (M = 2.29, S.E. = .15) than males (M = 3.13, S.E. = .17), p = .001, while there were no gender differences in other ratings, ps > .386 (closeness ratings for friend: female M = 6.71, S.E. = .15; male M = 6.62, S.E. = .17; trustworthiness ratings for friend: female M = 6.32, S.E. = .15; male M = 6.33, S.E. = .17; trustworthiness ratings for stranger: female M = 3.69, S.E. = .15; male M = 3.49, S.E. = .17). However, regardless of their gender, participants consistently gave their friend greater closeness ratings than trustworthiness ratings and gave the stranger greater trustworthiness ratings than closeness ratings.

as random effects. A significant main effect of Player (b = -.04, S.E. = .01, *t* = -2.76, *p* = .006) showed that participants showed less rTPJ activity in response to their friend (M = -.03, S.E. = .02) than to the stranger (M = .05, S.E. = .02). This effect was qualified by a significant Player X Valence X Prediction Error interaction (b = .03, S.E. = .01, *t* = 2.19, *p* = .029) indicating that more negative prediction error in the friend-taking condition was associated with increased rTPJ activity (b = -.08, S.E. = .03, 95% CI = [-.14, -.02]), while prediction error signal from other conditions did not significantly track rTPJ activity (friend-giving b = -.003, S.E. = .03, 95% CI = [-.06, .05]; stranger-taking b = .04, S.E. = .03, 95% CI = [-.02, .09]; stranger-giving (b = -.01, S.E. = .03, 95% CI = [-.07, .05]). Additionally, there was a significant interaction between Participant Gender X Valence X Prediction error (b = -.03, S.E. = .01, t = -2.10, p = .036), revealing that Prediction Error was more tightly connected with rTPJ activity for males in the taking condition, although the simple effect was not significant (b = -.05, S.E. = .03, 95% CI = [-.12, .01]; female taking b = .01, S.E. = .03, 95% CI = [-.04, .06]; male giving b = .02, S.E. = .03, 95% CI = [-.04, .08]; female giving b = -.04, S.E. = .03, 95% CI = [-.09, .02]).

Supplementary section 5. Findings when Amount was entered in the model

a) Effect of conditions on ratings

To examine the effect of trial-by-trial amounts on participants' ratings in the Social Judgment Task, we ran a linear mixed-effects model with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and Amount (small, large) as the fixed effects, and individual participants as random effects. Again, we could find the main effects of Player (b = 1.68, S.E. = .02, t = 73.47, p < .001), Valence (b = .57, S.E. = .02, t = 25.01, p < .001), and Task (b = -.14, S.E. = .02, t = -6.26, p < .001), indicating that participants gave more positive evaluations 1) to their friend (M = 6.49, S.E. = .10) than to the stranger (M = 3.12, S.E. = .10), 2) in the giving (M = 5.38, S.E. = .10) than in the taking (M = 4.23, S.E. = .10) conditions, and 3) on the trustworthiness (M = 4.95, S.E. = .10) than on the closeness (M = 4.66, S.E. = .10) scales. Importantly, the Valence main effect was qualified by Amount (b = .26, S.E. = .02, t = 11.27, p < .001), showing that participants gave more positive ratings when the amount was small (M = 4.51, S.E. = .11) than large (M = 3.96, S.E. = .11) in the taking condition, but gave more positive ratings when the amount was large (M = 5.62, S.E. = .11) than small (M = 5.14, S.E. = .11) in the giving condition, ps < .001, as one might expect to see. A significant Valence X Task interaction (b = -.05, S.E. = .02, t = -2.32, p = .020) showed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.18, S.E. = .11; trustworthiness M = 5.58, S.E. = .11) than in the taking condition (closeness M = 4.14, S.E. = .11; trustworthiness M = 4.32, S.E. = .11), ps < .001. Additionally, there was a significant Player X Task interaction (b = .32, S.E. = .02, t = 13.94, p < .001), indicating that participants gave more positive ratings to the stranger in the trustworthiness condition (M = 3.58, S.E. = .11) than in the closeness condition (M = 2.66, S.E. = .11), while they gave more positive ratings to their friend in the closeness condition (M = 6.67, S.E. = .11) than in the trustworthiness condition (M = 6.31, S.E. = .11), ps < .001. More importantly, we could still find the significant Player X Valence interaction (b = -.10, S.E. = .02, t = -4.50, p

< .001), indicating that participants differentiated the taking (M = 2.44, S.E. = .11) and giving (M = 3.80, S.E. = .11) conditions more for the stranger than for their friend (taking M = 6.02, S.E. = .11; giving M = 6.96, S.E. = .11), although both taking versus giving comparisons were significant, *ps* < .001.

b) Effect of rTPJ activity on ratings

We conducted a linear mixed-effects regression model on participants' trial-by-trial ratings with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), Amount (small, large), and their trial-by-trial rTPJ activity, and interactions between these factors as fixed effects, while individual participants were included as random effects. There were significant mains effect of Player (b = 1.68, S.E. = .02, *t* = 73.16, *p* < .001), Valence (b = .58, S.E. = .02, *t* = 25.06, *p* < .001), and Task (b = -.14, S.E. = .02, *t* = -6.26, *p* < .001); participants gave higher ratings 1) to their friend (M = 6.49, S.E. = .10) than to the stranger (M = 3.13, S.E. = .10); 2) in the giving condition (M = 5.38, S.E. = .10) than in the taking condition (M = 4.23, S.E. = .10); and 3) on the trustworthiness scale (M = 4.95, S.E. = .10) than on the closeness scale (M = 4.66, S.E. = .10).

These main effects were qualified by a couple of interactions. First, the effect of Valence was qualified by Amount (b = .26, S.E. = .02, *t* = 11.12, *p* < .001), showing that participants gave more positive ratings in the taking condition when the amount was small (M = 4.50, S.E. = .11) than large (M = 3.96, S.E. = .11), while they gave more positive ratings in the giving condition when the amount was large (M = 5.62, S.E. = .11) than small (M = 5.14, S.E. = .11), *ps* < .001. A significant Valence X Task interaction (b = -.05, S.E. = .02, *t* = -2.36, *p* = .018) revealed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.19, S.E. = .11; trustworthiness M = 5.58, S.E. = .11, *p* < .001) than in the taking condition (closeness M = 4.14, S.E. = .11; trustworthiness M = 4.32, S.E. = .11, *p* = .005). Additionally, there was a significant Valence X Task X rTPJ activity interaction (b = .10, S.E. = .05, *t* = 2.07, *p* = .038) revealing that rTPJ activity particularly tracked the trustworthiness ratings (b = -.22, S.E.

= .10, 95% CI = [-.41, -.03]) in the giving condition. RTPJ activity was not associated with other ratings (closeness in giving condition: b = .004, S.E. = .09, 95% CI = [-.18, .19]; trustworthiness in taking condition: b = .02, S.E. = .09, 95% CI = [-.17, .20]; closeness in taking condition: b = -.15, S.E. = .10, 95% CI = [-.34, .04]). Additional interaction between Player X Task (b = .32, S.E. = .02, t = 13.89, p < .001) showed that participants delivered higher ratings for friend-closeness (M = 6.67, S.E. = .11) than for friend-trustworthiness (M = 6.32, S.E. = .11), while they gave greater ratings to stranger-trustworthiness (M = 3.59, S.E. = .11) than to stranger-closeness (M = 2.66, S.E. = .11), ps < .001. Finally, there was a significant Player X Valence interaction (b = -.11, S.E. = .02, t = -4.62, p < .001), indicating that participants differentiated between the taking and giving conditions more for the stranger (taking M = 2.44, S.E. = .11; giving M = 3.81, S.E. = .11) than for their friend (taking M = 6.02, S.E. = .11; giving M = 6.96, S.E. = .11), ps < .001. More importantly, we found that this effect was qualified by a marginal Player X Valence X rTPJ interaction (b = .08, S.E. = .05, t = 1.78, p = .076) as reported in the main paper. RTPJ activity was more tightly connected to participants' ratings in the friend-taking condition (b = -.18, S.E. = .10, 95% CI = [-.37, .01]) compared to other conditions (friend-giving: b = -.05, S.E. = .10, 95% CI = [-.24, .14]; stranger-taking: b = .05, S.E. = .09, 95% CI = [-.14, .23]; stranger-giving: b = -.16, S.E. = .09, 95% CI = [-.35, .02]), although none of the simple effects were significant after entering Amount in the model. This effect was not modulated by Amount.

c) Effect of prediction error on rTPJ

We conducted a linear mixed-effects regression model on participants' trial-by-trial rTPJ activity with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), Amount (small, large), and their trial-by-trial Prediction Error values, and interactions between these factors as fixed effects, while individual participants were included as random effects. There was a significant main effect of Player (b = -.03, S.E. = .01, t = -2.20, p = .028), indicating that participants showed more decreased rTPJ activity in response to their friend (M = -.01, S.E.

= .02) than to the stranger (M = .05, S.E. = .02). It was modulated by a significant Player X Valence X Amount interaction (b = -.05, S.E. = .01, t = -3.09, p = .002). As shown in Figure S2, participants showed more increased rTPJ activity when their friend took large amounts (M = .05, S.E. = .06) than small amounts (M = -.11, S.E. = .04), p = .017. Participants did not differentiate the amounts in other conditions (friend-giving: large M = -.03, S.E. = .05; small M = .03, S.E. = .03; stranger-taking: large M = .01, S.E. = .05; small M = .05, S.E. = .03; stranger-giving: large M = .13, S.E. = .06; small M = .02, S.E. = .04; ps > .09). There was an additional interaction between Player X Amount X Prediction Error (b = .03, S.E. = .02, t = 2.24, p = .025), showing that Prediction Error was more tightly associated with rTPJ activity when the friend gave or took small amounts (b = -.05, S.E. = .03, 95% CI = [-.12, .01]) compared to other conditions (friend-large: b = .02, S.E. = .03, 95% CI = [-.04, .09]; stranger-small: b = .03, S.E. = .03, 95% CI = [-.03, .09]; stranger-large: b = -.03, S.E. = .03, 95% CI = [-.09, .03]), although none of these simple effects were significant. More importantly, we could still find the Player X Valence X Prediction Error interaction (b = .03, S.E. = .02, t = 1.72, p = .086) on rTPJ activity, although it became marginal. Prediction error was more tightly associated with rTPJ activity when their friend took money; when participants experienced more negative prediction error in response to their friend, they showed greater rTPJ activity (friend-taking b = -.04, S.E. = .03, 95% CI = [-.10, .03]; friend-giving: b = .004, S.E. = .03, 95% CI = [-.06, .06]; stranger-taking: b = .03, S.E. = .03, 95% CI = [-.02, .09]; stranger-giving: b = -.03, S.E. = .03, 95% CI = [-.09, .03]).

Supplementary section 6. Additional findings from the linear mixed-effects model examining the influence of conditions on participants' responses

From the linear mixed-effects regression on participants' trial-by-trial ratings, with Player (friend, stranger), Valence (taking, giving), and Task (closeness, trustworthiness) as fixed effects and individual participants as random effects, we found additional conditional effects on participants' ratings along with those reported in the main paper. First, there was a significant Player X Task interaction (b = .32, S.E. = .02, t = 13.70, p < .001), indicating that participants gave more positive ratings to the stranger in the trustworthiness condition (M = 3.58, S.E. = .11) than in the closeness condition (M = 2.66, S.E. = .11), while they gave more positive ratings to their friend in the closeness condition (M = 6.67, S.E. = .11) than in the trustworthiness condition (M = 6.32, S.E. = .11), ps < .001. Moreover, there was a Task X Valence interaction (b = -.05, S.E. = .02, t = -2.34, p = .020), showing that participants differentiated the closeness and trustworthiness ratings more in the giving condition (closeness M = 5.18, S.E. = .11; trustworthiness M = 5.58, S.E. = .11, p < .001) than in the taking condition (closeness M = 4.14, S.E. = .11; trustworthiness M = 4.32, S.E. = .11, p = .007).

Supplementary section 7. Additional findings from the linear mixed-effects model examining the influence of rTPJ activity on participants' responses

From the linear mixed-effects regression on participants' trial-by-trial ratings, with individual participants as random effects and Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and participants' trial-by-trial RTPJ activity as the fixed effects, we found factors that contributed to the participants' ratings in addition to those reported in the main paper.

There were significant mains effect of Player (b = 1.68, S.E. = .02, t = 72.24, p < .001), Valence (b = .58, S.E. = .02, t = 24.73, p < .001), and Task (b = -.15, S.E. = .02, t = -6.26, p < .001); participants gave higher ratings 1) to their friend (M = 6.49, S.E. = .10) than to the stranger (M = 3.12, S.E. = .10); 2) in the giving condition (M = 5.38, S.E. = .10) than in the taking condition (M = 4.23, S.E. = .10); and 3) on the trustworthiness scale (M = 4.95, S.E. = .10) than on the closeness scale (M = 4.66, S.E. = .10).

These main effects were qualified by a couple of interactions. First, a significant Valence X Task interaction (b = -.06, S.E. = .02, t = -2.43, p = .015) revealed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.18, S.E. = .11; trustworthiness M = 5.58, S.E. = .11, p < .001) than in the taking condition (closeness M = 4.14, S.E. = .11; trustworthiness M = 4.32, S.E. = .11, p = .006). Additionally, there was a significant Valence X Task X rTPJ activity interaction (b = .12, S.E. = .05, t = 2.40, p = .017) revealing that rTPJ activity particularly tracked the trustworthiness ratings (b = -.23, S.E. = .10, 95% CI = [-.42, -.04]) in the giving condition. RTPJ activity was not associated with other ratings (closeness in giving condition: b = .03, S.E. = .09, 95% CI = [-.16, .21]; trustworthiness in taking condition: b = .03, S.E. = .10, 95% CI = [-.16, .21]; closeness in taking condition: b = -.18, S.E. = .10, 95% CI = [-.37, .01]). Additionally, an interaction between Player X Task (b = .32, S.E. = .02, t = 13.79, p < .001) showed that participants delivered higher ratings for friend-closeness (M = 6.66, S.E. = .11) than for friend-trustworthiness (M = 6.31, S.E. = .11), while

they gave higher ratings to stranger-trustworthiness (M = 3.59, S.E. = .11) than to stranger-closeness (M = 2.66, S.E. = .11), *ps* < .001. Finally, there was a significant Player X Valence interaction (b = -.10, S.E. = .02, *t* = -4.46, *p* < .001), indicating that participants differentiated between the taking and giving conditions more for the stranger (taking M = 2.44, S.E. = .11; giving M = 3.80, S.E. = .11) than for their friend (taking M = 6.02, S.E. = .11; giving M = 6.96, S.E. = .11), *ps* < .001. However, as reported in the main paper, this effect was qualified by a significant Player X Valence X rTPJ interaction.

Supplementary Section 8. Whole-brain analyses with prediction error signal

a. Methods

We conducted an exploratory whole-brain analyses on the time points when participants viewed the decision of Player 1 (see Figure 1 in the main paper) with the prediction error signal regressors. A general linear model (GLM, ordinary least-squares regression) including fourteen regressors of interest was constructed.

The first regressor marked the time points of each trial when participants observed the decision of Player 1. Other regressors marked: (1) Player 1 (Player: Friend = +1, Stranger = -1), (2) The type of rating participants were asked to make (Task: Closeness = +1, Trustworthiness = -1), (3) Valence of the decision (Valence: Giving = +1, Taking = -1), (4) Prediction Error, and the interactions between (5) Player X Task, (6) Player X Valence, (7) Player X Prediction Error, (8) Task X Prediction Error, (9) Valence X Prediction Error, (10) Player X Task X Prediction Error, (11) Player X Valence X Prediction Error, and (12) Player X Task X Valence X Prediction Error. To minimize the influence of physiological confounds, eight regressors of no interest were also included: six modeling head movement, one sampling white matter activity, and one sampling cerebrospinal fluid activity (Chang & Glover, 2009). Before they were submitted to the model, regressors of interest were convolved with a canonical gamma variate hemodynamic delay (Cohen, 1997). Linear regression t-statistic maps were converted to Z-scores, coregistered with structural maps, spatially normalized by warping to Montreal Neurological Institute space (linear to colin27T1_seg template), and resampled as 3mm cubic voxels.

A one-sample t-test was conducted with AFNI program 3dttest function to examine the group-level neural responses to each contrast. This t-test map was initially voxelwise thresholded, at $p < .005$, and then cluster thresholded, cluster size > 98 continuous $3mm^3$ voxels, to yield corrected maps for detecting whole-brain activity at $p < .05$ corrected. Cluster correction was performed using 3dClustSim as implemented in AFNI_16.2.06. We computed

the smoothness of the residuals of participants' data at the single subject level using 3dFWHMx, implementing the spatial autocorrelation function, and used these smoothness estimates as inputs into 3dClustSim with 10000 iterations.

b. Findings

We found increased activity in subgenual anterior cingulate cortex (sgACC), extended into ventral and dorsal striatum, in response to the Valence X Prediction Error regressor (Table S3; Figure S3). Consistent with previous findings (Fareri et al., 2012), this effect suggests that activity in this area was greater with more positive prediction error in the giving condition and more negative prediction error in the taking condition.

Table S3. Activation in response to the Social Judgment Task when Prediction Error was entered in the model

Contrast	Region	x	y	z	Peak Z	Voxels
Player	No voxels survived					
Task	No voxels survived					
Valence	No voxels survived					
Prediction Error	No voxels survived					
Player X Task	No voxels survived					
Player X Valence						
	L Middle temporal gyrus	-49	-50	-8	4.47	289
Player X Prediction Error	No voxels survived					
Task X Prediction Error						
	L Fusiform gyrus	-43	-47	-18	-3.74	247
	R Superior temporal gyrus	51	-20	-2	-3.91	111

Valence X Prediction Error						
	L Postcentral gyrus	-7	-54	71	-3.55	149
	R Anterior cingulate (extended into striatum)	**2**	**2**	**-8**	**4.05**	**138**
Player X Task X Prediction Error	No voxels survived					
Player X Valence X Prediction Error	No voxels survived					
Player X Task X Valence X Prediction Error	No voxels survived					

Note. Uncorrected $p < .005$, cluster > 98 continuous voxels, corrected $p < .05$, regions of interest in bold. R; right, L; left.

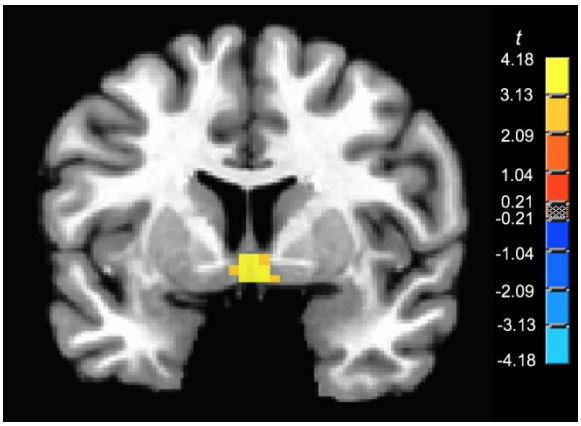

Figure S3. *Subgenual anterior cingulate cortex (sgACC) activity modulated by prediction error.* (A) Whole-brain analysis revealed that participants showed increased sgACC activity, extended into striatum, in response to the interaction between valence and prediction error signal. $p < .005$ uncorrected, cluster > 98 continuous voxels, $p < .05$ corrected.

References

Chang, C., Glover, G.H. (2009). Effects of model-based physiological noise correction on default mode network anti- correlations and correlations. *NeuroImage, 47*(4), 1448–59.

Cohen, M.S. (1997). Parametric analysis of fMRI data using linear systems methods. *NeuroImage, 6*(2), 93–103.

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology, 108*, 3068-3072.

Fareri, D. S., Chang, L., J., & Delgado, M., R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Koster-Hale, J. & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836-848.

Saxe, R. (2009). Theory of mind (neural basis). In Banks, W. (Ed.), *Encyclopedia of Consciousness*. Cambridge, MA: MIT Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87-124.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *Neuroimaging, 19*, 1835-1842.

Saxe, R., & Powell, L. J. (2006). It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692-699.

Saxe, R., & Wexler A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia. 43*(10), 1391-1399.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*, 1435-1446.