Supplementary Materials for

The role of right temporo-parietal junction in processing social prediction error across relationship contexts

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

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 mPFC	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, mPFC; medial prefrontal cortex.

Supplementary Section 2. Analyses with participants who completed Theory-of-Mind (ToM) task

We conducted volumes-of-interest (VOI) analyses with participants who completed the Theory-of-Mind (ToM) task (N = 21). A spherical VOI (radius = 8mm) centered on the coordinates derived from each participant's ToM localizer task in the rTPJ was constructed. We extracted average PSC from each participant's 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).

First, to test whether prediction error signal derived from the Dynamic LGSP 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 = -.04, S.E. = .01, t = -3.21, p = .001) revealed that participants showed overall lower rTPJ activity in response to their friend (M = -.03, S.E. = .02) than to a stranger (M = .04, S.E. = .02). An interaction between Player X Prediction Error (b = -.03, S.E. = .01, t = -2.36, p = .019) showed that more negative prediction error in the friend condition was associated with greater rTPJ activity (b = -.04, S.E. = .02, t = -2.02, p = .043, 95% CI = [-.07, -.001], while prediction error and rTPJ activity was not associated in the stranger condition (b = .02, S.E. = .02, t = 1.27, p = .206, 95% CI = [-.01, .05]). More importantly, this effect was modulated by a significant Player X Valence X Prediction Error interaction (b = .04, S.E. = .01, t = 3.11, p = .002) indicating that more negative prediction error in the friend-taking condition was associated with increased rTPJ activity (b = -.08, S.E. = .03, t = -2.95, p = .003, 95% CI = [-.13, -.03]). In addition, more positive prediction error in the stranger-taking condition was associated with more increased rTPJ activity (b = .05, S.E. = .02, t

= 2.38, p = .018, 95% CI = [.01, .10])[1]. Prediction error signal from other conditions did not significantly track rTPJ activity (friend-giving b = .005, S.E. = .02, t = .21, p = .837, 95% CI = [-.04, .05]); stranger-giving (b = -.01, S.E. = .02, t = -.44, p = .662, 95% CI = [-.06, .04]).

Second, 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.

We found a marginal Player X rTPJ interaction (b = -.12, S.E. = .06, t = -1.88, p = .061). Decreased rTPJ was associated with less negative ratings in the friend condition (b = -.24, S.E. = .09, t = -2.68, p = .007, 95% CI = [-.42, -.07]). RTPJ activity was not associated with ratings in the stranger condition (b = -.005, S.E. = .09, t = -.05, p = .957, 95% CI = [-.18, .17]).

Although the Player X Valence X rTPJ interaction was only directional (b = .10, S.E. = .06, t = 1.51, p = .130), we proceed with examining the simple effects of this interaction to see if we could find corresponding patterns with the analyses using all participants (see the main paper). Confirming findings reported in the main paper, rTPJ activity only significantly tracked participants' ratings in the friend-taking condition (b = -.38, S.E. = .13, t = -2.88, p = .004, 95% CI = [-.63, -.12]). RTPJ activity was not associated with ratings in any other conditions (friend-giving: b = -.11, S.E. = .13, t = -.87, p = .384, 95% CI = [-.35, .14]; stranger-taking: b = .05, S.E. = .13, t = .42, p = .677, 95% CI = [-.20, .31]; stranger-giving: b = -.06, S.E. = .12, t = -.51, p = .607, 95% CI = [-.31, .18]). Thus, overall, we found converging patterns with rTPJ activity extracted from each participant's ToM localizer tasks as well as with rTPJ activity extracted from a group ToM localizer map.

[1] However, given that this association was absent in the full data, we do not want to overinterpret this effect.

Supplementary section 3. 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; $p < .001$).[2]

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.[3] 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, $t = 9.80$, $p < .001$), 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 = -1.50, S.E. = .31, $t = -4.82$, $p < .001$; pre-scan closeness 8: b = -1.32, S.E. = .36, $t = -3.67$, $p = .001$). 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, $t = 11.43$, $p < .001$; Response X Pre-scan closeness 8: b = .36, S.E. = .04, $t = 8.77$, $p < .001$; 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

[2] 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, $p < .001$)

[3] 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 = .66, S.E. = .03, 95% CI = [.61, .71]) (Figure S1).

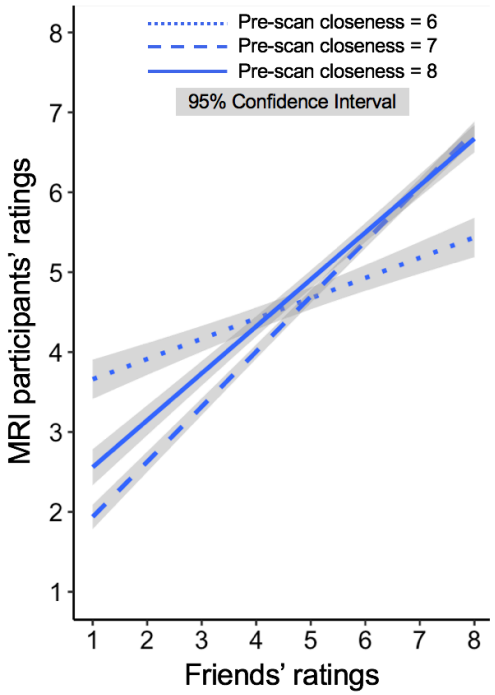

Figure S1. Associations between the MRI participants' ratings and their friends' ratings during the Social Judgment Task, divided by the MRI participants' pre-scan closeness ratings that they gave to their friends.

Supplementary section 4. Whole-brain analyses

a. Methods

We conducted an exploratory whole-brain analysis on the time points when participants viewed the decision of Player 1 (see Figure 1 in the main paper) with the prediction error signal as a regressor. A general linear model (GLM, ordinary least-squares regression) with thirteen 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) PE (model-derived values), and the interactions between (5) Player X Task, (6) Player X Valence, (7) Player X PE, (8) Task X PE, (9) Valence X PE, (10) Player X Task X PE, (11) Player X Valence X PE, and (12) Player X Task X Valence X PE. 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. The average correlation of regressors in the fMRI design matrix was M = .05, S.D. = .02, ranging from .03 - .08.

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 $p < .01$, and then cluster thresholded, cluster size > 166 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), extending into ventral and dorsal striatum, in response to the Valence X PE regressor (Table S2; Figure S2). This effect suggests that sgACC activity was increased along with more positive PE in the giving condition, and more negative PE in the taking condition.

Table S2. 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						
	L Precuneus	-25	-83	37	3.34	185
Prediction Error	No voxels survived					
Player X Task	No voxels survived					
Player X Valence						
	L Middle temporal gyrus	-53	-50	-8	4.24	375
	R Precuneus	14	-64	15	3.81	236
	R Cingulate Cortex	8	-43	39	3.69	229
Player X Prediction Error	No voxels survived					

Task X Prediction Error						
	R Superior temporal gyrus	47	-20	-2	-3.69	924
	L Middle Temporal Gyrus	-56	-47	-11	-3.97	743
	R Superior temporal gyrus (extended into R insula)	44	22	-25	-3.71	288
Valence X Prediction Error						
	R Anterior cingulate (extended into striatum)	**2**	**2**	**-8**	**5.00**	**236**
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						
	L Middle Frontal Gyrus	-49	13	29	-3.85	335
	R Postcentral Gyrus	32	-34	49	-3.38	210

Note. Uncorrected $p < .01$, cluster > 166 continuous voxels, corrected $p < .05$, regions of interest in bold. R; right, L; left.

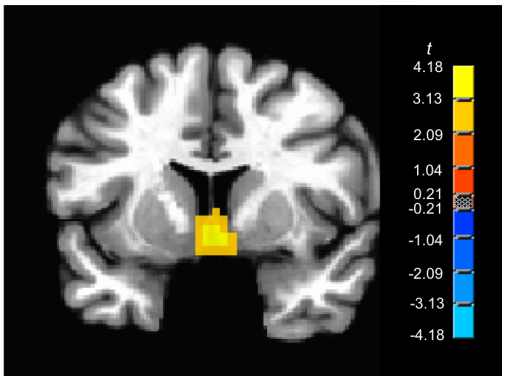

Figure S2. *Subgenual anterior cingulate cortex (sgACC) activity modulated by prediction error.* (A) Whole-brain analysis revealed that participants showed increased sgACC activity, extended into striatum, in response to the interaction between valence and prediction error signal. $p < .01$ uncorrected, cluster > 166 continuous voxels, $p < .05$ corrected.

Supplementary section 5. Findings when Participant Gender was entered in the model

a) Effect of conditions on ratings and updates

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 = 71.77, p < .001), Valence (b = .58, S.E. = .02, t = 24.91, p < .001), and Task (b = -.13, S.E. = .02, t = -5.50, 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 = 5.40, S.E. = .11) versus taking (M = 4.24, S.E. = .11); and 3) trustworthiness (M = 4.94, S.E. = .11) versus closeness (M = 4.69, S.E. = .11). A Player X Valence interaction (b = -.10, S.E. = .02, t = -4.50, p < .001) revealed that the difference in ratings between giving and taking was greater for strangers (M = 1.37, S.E. = .07) than friends (M = .95, S.E. = .07).

Participant Gender qualified the main effects of Player (b = .09, S.E. = .02, t = 3.97, p < .001), Valence (b = -.12, S.E. = .02, t = -5.07, p < .001), and Task (b = -.11, S.E. = .02, t = -4.92, p < .001), indicating that females differentiated their friend (M = 6.51, S.E. = .14) and the stranger (M = 2.98, S.E. = .14) more than males (friend M = 6.46, S.E. = .16; stranger M = 3.30, S.E. = .16), while males differentiated taking (M = 4.18, S.E. = .16) and giving (M = 5.58, S.E. = .16) conditions more than females (taking M = 4.29, S.E. = .14; giving M = 5.21, S.E. = .14). Moreover, females gave greater ratings in the trustworthiness condition (M = 4.99, S.E. = .14) than in the closeness condition (M = 4.51, 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 = .697). 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 = 12.95, p < .001), showing that participants gave more positive ratings to the stranger in the trustworthiness condition (M = 3.57, 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 = 6.31, S.E. = .11), ps < .001. This effect was modulated by Participant Gender (b = .14, S.E. = .02, t = 6.13, p < .001), manifested more with female participants (friend closeness M = 6.72, S.E. =.15; friend trustworthiness M = 6.31, S.E. = .15; stranger closeness M = 2.30, S.E. = .15; stranger trustworthiness M = 3.67, S.E. = .15; closeness versus trustworthiness comparisons ps < .001) than with male participants (friend closeness M = 6.61, S.E. = .17; friend trustworthiness M = 6.32, S.E. = .17; p = .003; stranger closeness M = 3.13, S.E. = .17; stranger trustworthiness M = 3.48, 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 = -2.27, p = .023). Participants differentiated the closeness and trustworthiness ratings more in the giving condition (closeness M = 5.21, S.E. = .11; trustworthiness M = 5.58, S.E. = .11, p < .001) than in the taking condition (closeness M = 4.16, S.E. = .11; trustworthiness M = 4.31, S.E. = .11, p = .022).

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 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.

We found main effects of Valence (b = .04, S.E. = .02, t = 2.07, p = .039) and Task (b = -.09, S.E. = .02, t = -5.14, p < .001), indicating that the participants updated more when Player 1 gave money (M = 1.16, S.E. = .19) versus took money (M = 1.09, S.E. = .19), and updated more for trustworthiness (M = 1.22, S.E. = .19) versus closeness (M = 1.03, S.E. = .19). Critically, we found a significant main effect of Player (b = -.20, S.E. = .02, t = -11.42, p < .001), such that the participants updated less for friend (M = .92, S.E. = .19) versus stranger (M = 1.33, S.E. = .19) overall. These effects were not modulated by any other factors, including Participant Gender.

b) Effect of prediction error on updating

We conducted a linear mixed-effects regression on participants' degrees of updating with Participant Gender (male, female), Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and the size of trial-by-trial PE values, and interactions between these factors as fixed effects, while individual participants were included as random effects. The first trial from each participant was excluded to avoid any bias from the initialized values. A significant main effect of Player (b = -.10, S.E. = .04, t = -2.59, p = .010) showed that participants updated more for the stranger (M = 1.31, S.E. =.18) than for their friend (M = .94, S.E. = .18) overall. A significant main effect of the absolute PE (b = .46, S.E. = .04, t = 11.83, p < .001) indicated that greater PE was associated with greater updating. These main effects were modulated by a significant Player X absolute PE interaction effect (b = -.11, S.E. = .04, t = -2.88, p = .004) revealed that PE was more tightly related to updating for the stranger (b = .57, S.E. = .05, 95% CI = [.46, .68]) than for friend (b = .35, S.E. = .05, 95% CI = [.25, .46]). Although the Player X Valence X absolute PE interaction did not reach the significant level (b = .05, S.E. = .04, t = 1.35, p = .178), simple effects showed that the association between the absolute PE and absolute updating was the smallest in the friend-taking condition (b = .25, S.E. = .08, 95% CI = [.09, .41]) than in the other conditions (friend-giving b = .46, S.E. = .08, 95% CI = [.31, .61]; stranger-taking b = .57, S.E. = .08, 95% CI = [.42, .72]; stranger-giving b = .57, S.E. = .08, 95%

CI = [.41, .72]). These effects were not modulated by any other factors, including Participant Gender.

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 as random effects. A significant main effect of Player (b = -.04, S.E. = .01, t = -2.79, p = .005) 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 = 2.24, p = .025) 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 = -.0005, S.E. = .03, 95% CI = [-.05, .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 = -2.18, p = .030), 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 = -.06, 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 = [-.03, .08]; female giving b = -.04, S.E. = .03, 95% CI = [-.09, .02]).

d) 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 = -2.13, p = .033), a Player X Valence X rTPJ interaction (b = .10, S.E. = .05, t = 2.11, p = .035) revealed this effect to be more pronounced in the friend-taking condition (b = -.21, S.E. = .10, 95% CI = [-.40, -.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 = -.20, S.E. = .10, 95% CI = [-.38, -.01]).[4] RTPJ activity was not associated with ratings of the stranger in the taking condition (b = .06, S.E. = .10, 95% CI = [-.12, .25]) or ratings of their friend in the giving condition (b = -.07, S.E. = .10, 95% CI = [-.26, .12]). These effects were not modulated by any other factors, including Participant Gender.

[4] 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.

Supplementary section 6. 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 = 72.99, p < .001), Valence (b = .57, S.E. = .02, t = 24.57, p < .001), and Task (b = -.14, S.E. = .02, t = -6.14, p < .001), indicating that participants gave more positive evaluations 1) to their friend (M = 6.48, S.E. = .10) than to the stranger (M = 3.12, S.E. = .10), 2) in the giving (M = 5.37, S.E. = .10) than in the taking (M = 4.23, S.E. = .10) conditions, and 3) on the trustworthiness (M = 4.94, 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 = 11.46, p < .001), showing that participants gave more positive ratings when the amount was small (M = 4.52, S.E. = .11) than large (M = 3.95, 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 = 5.12, 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 = -2.29, p = .022) showed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.17, S.E. = .11; trustworthiness M = 5.56, S.E. = .11; p < .001) than in the taking condition (closeness M = 4.15, S.E. = .11; trustworthiness M = 4.32, S.E. = .11; p = .006). Additionally, there was a significant Player X Task interaction (b = .32, S.E. = .02, t = 13.84, 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 = 6.66, 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 = -.11, S.E. = .02, t = -4.56, p < .001), indicating that participants differentiated the taking (M = 2.45, S.E. = .11) and giving (M = 3.79, S.E. = .11) conditions more for the stranger than for their friend (taking M = 6.02, S.E. = .11; giving M = 6.95, S.E. = .11), although both taking versus giving comparisons were significant, ps < .001.

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), Task (closeness, trustworthiness), Amount (small, large), 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 = 2.05, p = .040), Task (b = -.09, S.E. = .02, t = -5.44, p < .001), and Amount (b = .25, S.E. = .02, t = 14.40, p < .001), indicating that the participants updated more when Player 1 gave money (M = 1.12, S.E. = .19) versus took money (M = 1.05, S.E. = .19), updated more for trustworthiness (M = 1.18, S.E. = .19) versus closeness (M = .99, S.E. = .19), and updated more when the amount was large (M = 1.33, S.E. = .19) than small (M = .84, S.E. = .19). A significant interaction between Valence X Amount (b = .04, S.E. = .02, t = 2.23, p = .026) indicated that participants updated more in the giving condition (M = 1.41, S.E. = .19) than in the taking condition (M = 1.26, S.E. = .19) when amount was large, p = .002, but they did not differentiate giving and taking conditions (giving M = .84, S.E. = .19; taking M = .84, S.E. = .19) when amount was small, p = .902. Critically, we still found a significant main effect of Player (b = -.20, S.E. = .02, t = -11.81, p < .001), such that the participants updated less for friend (M = .89, S.E. = .19) versus stranger (M = 1.29, S.E. = .19) overall. This effect was not modulated by any other factors, including the amount.

b) Effect of prediction error on updating

We conducted a linear mixed-effects regression on participants' degrees of updating with Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), Amount (small, large), and the size of trial-by-trial PE values (i.e., absolute value), and interactions between these factors as fixed effects, while individual participants were included as random effects. The first trial from each participant was excluded to avoid any bias from the initialized values. A significant main effect of Player (b = -.10, S.E. = .04, t = -2.69, p = .007) showed that participants updated more for the stranger (M = 1.25, S.E. =.18) than for their friend (M = .86, S.E. = .18) overall. A significant main effect of the absolute PE (b = .23, S.E. = .04, t = 5.35, p < .001) indicated that greater PE was associated with greater updating. These main effects were modulated by a few interactions. First, a significant Amount X absolute PE interaction (b = .18, S.E. = .04, t = 4.50, p < .001) indicated that PE was more tightly linked to updating when the amount was large (b = .41, S.E. = .05, 95% CI = [.31, .52]) compared to when the amount was small (b = .05, S.E. = .06, 95% CI = [-.08, .17]). Moreover, significant Player X absolute PE interaction effect (b = -.11, S.E. = .04, t = -2.61, p = .009) revealed that PE was more tightly related to updating for the stranger (b = .34, S.E. = .06, 95% CI = [.22, .45]) than for friend (b = .12, S.E. = .06, 95% CI = [.01, .24]).

These effects were further modulated by a significant Player X Valence X Amount X absolute PE interaction (b = .11, S.E. = .04, t = 2.77, p = .006). Weaker associations between absolute PE and absolute updating for friend versus stranger was most salient in taking condition, especially when amount was large (friend-taking-large b = .15, S.E. = .11, 95% CI = [-.07, .37]; significantly different from stranger-taking-large b = .67, S.E. = .10, 95% CI = [.48, .86], p < .001). Although the association between absolute PE and absolute updating was smaller for friend than for stranger in the taking-small amount conditions, this effect was not significant (friend-taking-small b = -.06, S.E. = .14, 95% CI = [-.33, .22]; stranger-taking-small b = .02, S.E. = .12, 95% CI = [-.22, .26], p = .674). Also, unexpectedly, the association between absolute PE and absolute updating was also smaller for the friend-giving-small amount

condition (b = -.06, S.E. = .12, 95% CI = [-.31, .18]) than in the stranger-giving-small amount condition (b = .29, S.E. = .12, 95% CI = [.04, .53]), p = .046. The associations between absolute PE and absolute updating for friend-giving-large amount and stranger-giving-large amount conditions were not significantly different (friend-giving-large b = .46, S.E. = .10, 95% CI = [.26, .66]; stranger-giving-large b = .36, S.E. = .12, 95% CI = [.14, .59]), p = .535.

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. = .02, t = -2.18, p = .030), 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 = -.04, S.E. = .02, t = -2.87, p = .004). Participants showed more increased rTPJ activity when their friend took large amounts (M = .04, S.E. = .06) than small amounts (M = -.10, S.E. = .04), p = .025. Participants did not differentiate the amounts in other conditions (friend-giving: large M = -.02, 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 = .14, 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 = 2.09, p = .037), 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 = [-.11, .01]) compared to other conditions (friend-large: b = .02, S.E. = .03, 95% CI = [-.05, .08]; 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 = 1.73, p = .084) 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 = [-.11, .03]; friend-giving: b = .004, S.E. = .03, 95% CI = [-.06, .06]; stranger-taking: b = .03, S.E. = .03, 95% CI = [-.03, .09]; stranger-giving: b = -.03, S.E. = .03, 95% CI = [-.10, .03]).

d) 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. While reduced rTPJ activity in general was associated with more positive ratings (b = -.09, S.E. = .05, t = -1.93, p = .054), this effect was still qualified by a marginal Player X Valence X rTPJ interaction (b = .09, S.E. = .05, t = 1.81, p = .070) as in the main paper. RTPJ activity was more tightly connected to participants' ratings in the friend-taking condition (b = -.19, S.E. = .10, 95% CI = [-.38, .001]) compared to other conditions (friend-giving: b = -.06, S.E. = .10, 95% CI = [-.24, .13]; stranger-taking: b = .05, S.E. = .10, 95% CI = [-.14, .23]; stranger-giving: b = -.17, S.E. = .09, 95% CI = [-.36, .02]), although none of the simple effects were significant after entering Amount in the model. This effect was not modulated by any other factors, including Amount.

Supplementary section 7. 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 = 13.60, 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 = 6.66, S.E. = .11) than in the trustworthiness condition (M = 6.31, S.E. = .11), ps < .001. Moreover, there was a Task X Valence interaction (b = -.05, S.E. = .02, t = -2.26, p = .024), showing that participants differentiated the closeness and trustworthiness ratings more in the giving condition (closeness M = 5.18, S.E. = .11; trustworthiness M = 5.56, S.E. = .11, p < .001) than in the taking condition (closeness M = 4.15, S.E. = .11; trustworthiness M = 4.32, S.E. = .11, p = .008).

Supplementary section 8. Directional Player by Valence by absolute Prediction Error effect on absolute updating

From the linear mixed-effects regression on participants' degrees of updating, entering Player (friend, stranger), Valence (taking, giving), Task (closeness, trustworthiness), and the size of trial-by-trial PE values and interactions between these factors as fixed effects and individual participants as random effects, there was a directional effect of Player X Valence X absolute PE interaction (b = .05, S.E. = .04, t = 1.36, p = .174). The simple effects showed that the association between the absolute PE and absolute updating was the smallest in the friend-taking condition (b = .24, S.E. = .08, t = 2.99, p = .003, 95% CI = [.08, .40]) than in the other conditions (friend-giving b = .44, S.E. = .08, t = 5.84, p < .001 , 95% CI = [.29, .59]; stranger-taking b = .58, S.E. = .07, t = 7.73, p < .001 , 95% CI = [.43, .72]; stranger-giving b = .57, S.E. = .08, t = 7.16, p < .001 , 95% CI = [.41, .72]). Specifically, associations between absolute PE and absolute updating was significantly smaller in the friend-taking condition than in the stranger-taking condition, p = .002, and was marginally smaller in the friend-taking condition than in the friend-giving condition, p = .080. Other conditions were not significantly different from each other, ps > .266 (Figure S3).

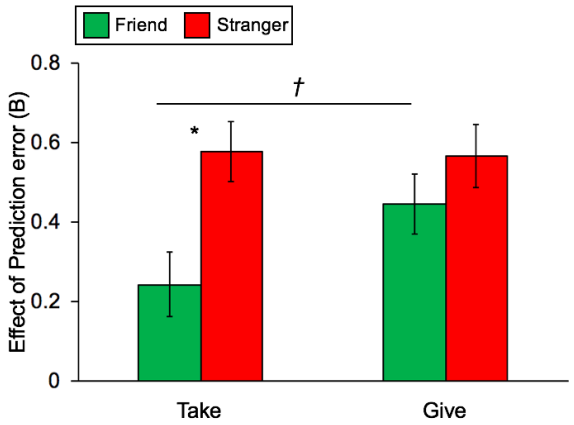

Figure S3. *Association between absolute prediction error (PE) and absolute updating.* Participants' trial-by-trial PE was significantly associated with the extent to which they updated their ratings across conditions. However this association was the smallest in the friend-taking condition. *$*p < .05$, †$p < .10$.*

Supplementary Section 9. 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 = 71.69, p < .001), Valence (b = .57, S.E. = .02, t = 24.29, p < .001), and Task (b = -.14, S.E. = .02, t = -6.13, p < .001); participants gave higher ratings 1) to their friend (M = 6.48, S.E. = .10) than to the stranger (M = 3.12, S.E. = .10); 2) in the giving condition (M = 5.37, 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 = -2.35, p = .019) revealed that participants differentiated closeness and trustworthiness more in the giving condition (closeness M = 5.18, S.E. = .11; trustworthiness M = 5.57, 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). Additionally, there was a significant Valence X Task X rTPJ activity interaction (b = .11, S.E. = .05, t = 2.24, p = .025) revealing that rTPJ activity particularly tracked the trustworthiness ratings (b = -.23, S.E. = .10, 95% CI = [-.42, -.03]) in the giving condition. RTPJ activity was not associated with other ratings (closeness in giving condition: b = .01, S.E. = .09, 95% CI = [-.17, .20]; trustworthiness in taking condition: b = .01, S.E. = .10, 95% CI = [-.17, .20]; 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 = 13.69, 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 = -4.47, 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 = 3.80, S.E. = .11) than for their friend (taking M = 6.02, S.E. = .11; giving M = 6.95, S.E. = .11), *ps* < .001. However, as reported in the main paper, this effect was qualified by a marginal Player X Valence X rTPJ interaction.

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.