Social Cognitive and Affective Neuroscience, 2020, 1–10

doi: 10.1093/scan/nsaa072
Original Manuscript

The role of right temporoparietal junction in processing social prediction error across relationship contexts

BoKyung Park,[1] Dominic Fareri,[2,] Mauricio Delgado,[3] and Liane Young[1]

[1]Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA 02467, USA, [2]Derner School of Psychology, Adelphi University, Garden City, NY 11530, USA and [3]Psychology Department, Rutgers University-Newark, Newark, NJ 07102, USA

Correspondence should be addressed to BoKyung Park, Department of Psychology and Neuroscience, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. E-mail: parkanj@bc.edu

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 (i.e. prediction error). Combining neuroimaging and computational modeling, we test the hypothesis that brain regions associated with theory of mind, especially right temporoparietal 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 across trials. Overall, participants engaged in less impression updating for friends *vs* strangers. Decreased rTPJ activity in response to a friend's negative behavior (taking money) was associated with reduced negative updating and increased positive ratings of the friend. However, to the extent that participants did update their impressions (more negative ratings) of friends, this behavioral pattern was explained by greater prediction error and greater rTPJ activity. These findings suggest that rTPJ recruitment represents the integration of prediction error signals and the capacity to overcome people's motivation to maintain positive impressions of friends in the face of conflicting evidence.

Key words: impression updating; motivated cognition; social prediction error; theory of mind

Introduction

People update their impressions of others in the face of new information to navigate an ever-changing social environment. Yet, as prior research has documented, when new information is inconsistent with people's pre-existing impressions, people often resist updating (Asch, 1946; Vonk, 1994). For example, information that is congruent with one's pre-existing impression is endorsed more (Crocker *et al.*, 1983; Fareri *et al.*, 2012), while inconsistent information is often attributed to situational and conditional factors (Wright and Mischel, 1988; Vonk, 1994). Resistance to impression updating has been especially well established in the context of pre-existing social relationships. A number of studies have demonstrated that people are motivated to positively perceive others who are close to them (Taylor and

Received: 22 August 2019; **Revised:** 6 May 2020; **Accepted:** 29 May 2020

Brown, 1988; Murray, 1999; Hughes and Beer, 2012). In particular, people are more likely to attribute close others' negative behaviors to external causes (Taylor and Koivumaki, 1976), consistent with 'motivated cognition'. That is, reasoning processes can be shaped by people's desire for certain conclusions (Kunda, 1990; Stevens and Fiske, 1995).

Related literature on intergroup cognition suggests that motivated impression maintenance may 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 temporoparietal 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 error (PE)—the difference between what people expect of others and what people actually observe (Koster-Hale and Saxe, 2013). Neuroimaging studies investigating the underlying mechanisms of impression updating have identified a role for regions that have been implicated in ToM. While some recent evidence implicates these regions in updating in contexts that do not directly evoke mentalizing—i.e. associative learning between a social agent and an object (Lockwood *et al.*, 2018) or self-relevant processing (Wittmann *et al.*, 2016)—the vast majority of work has established an association between these regions and impression updating in social domains (Hampton *et al.*, 2008; ; Mende-Siedlecki *et al.*, 2013a,b; Hill *et al.*, 2017; Thornton and Mitchell, 2018). Importantly, increased activation in 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, this region may be involved in motivated impression maintenance insofar as it is responsible for diminished representation of social PE (Hughes *et al.*, 2017). This study aims to test this proposal empirically by combining neuroimaging and computational modeling.

An important alternative account is that update resistance does not reflect motivated cognition but instead rational processing. For example, given that people typically possess strong prior knowledge about close others, it may be rational to rely on the strength of these priors rather than update based on new, prior-inconsistent information (Hahn and Harris, 2014; Gershman, 2019). Although both motivated and rational accounts predict update resistance in the case of close others, the current work provides an initial exploration of computational and neural processes that lead to updating or update resistance. First, we examined whether updating would be best accounted for by an interaction between initial impressions and trial-by-trial estimations of player value. Second, we speculated that an association between decreased (*vs* increased) rTPJ activity and update resistance in the case of close others would indicate motivated (*vs* rational) cognition. As we have proposed elsewhere (Kim *et al.*, 2020; Park *et al.*, in press), the rational discounting account suggests that processing prior-inconsistent information but failing to update on its basis may involve enhanced mentalizing (and generating auxiliary hypotheses) to explain it away (Gershman, 2019).

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 tested four key hypotheses. (i) We expected that people would be more likely to maintain their impressions of their friend compared to their impressions of a stranger, especially when learning new negative information. (ii) We hypothesized that to the extent that people do update their impressions of their friends, this pattern could be explained by a social PE account. (iii) We predicted that increased rTPJ activity would encode social PE evoked by negative behaviors performed by participants' friends. (iv) Finally, we predicted that rTPJ activity would account for the degree to which participants update their impressions based on new information.

To test the second and third hypotheses, we applied a computational modeling approach as suggested in prior work (Kliemann and Adolphs, 2018) for characterizing mechanisms of social learning (Kishida and 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 (i) their initial impressions of each player and (ii) 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 PEs and that increased rTPJ activity would be associated with greater PE.

Method

Participants

Thirty right-handed, neurologically and psychologically healthy native English speakers took part in this study, bringing a close, same-gender 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 (>3 mm) 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 (Kahneman *et al.*, 1986; Forsythe *et al.*, 1994). 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.

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).

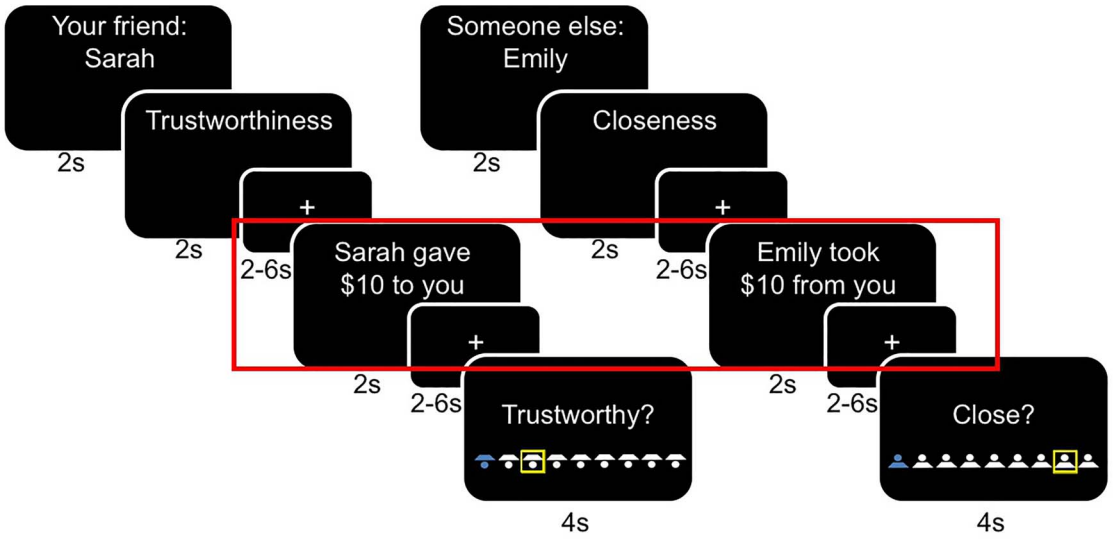

Fig. 1. Representative trials of Social Judgment Task. Participants viewed the name of Player 1 (2s); the rating that they would make in the trial (2s) and a jittered fixation cross (2–6s). Then, the decision of the Player 1 appeared on the screen (2s). After a jittered fixation cross (2–6s), participants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

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 how close they felt to Player 1 ('closeness'). Thus, the game matrix consisted of a 2 (Player: Friend, Stranger) × 2 (Valence: Taking, Giving) × 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 (4s). 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. Preregistered 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 gender-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 eight trials, participants entered the scanner and completed 192 trials of the game (16 trials in each of 12 runs, total time = 74 min 24 s) 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 = 9 min 4 sec) (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; Supplementary Section 2 for findings with only the participants who completed the ToM task). Finally, participants exited the scanner and completed a post-scan survey not explored in this article. 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 magnetic resonance imaging (MRI) participants, and played the same game as Player 2 outside the scanner (see Supplementary Section 3 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 analyses

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 3 × 3 × 3mm slices of gradient echo T2∗-weighted echo-planar images provided whole brain coverage [time repetition (TR) = 2s, echo time (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_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), deobliqued, 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 8 mm 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 <0.01 Hz, as described in Wu *et al.*, 2014) (See Supplementary Section 4 for whole-brain analyses).

Volume-of-interest analyses. Given that our a priori predictions focused on the associations between rTPJ and impression updating, we conducted volume-of-interest (VOI) analyses. A spherical VOI (radius = 8 mm) 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 in which participants observed Player 1's decision for each task condition. Sampling was delayed by 4 s to account for the hemodynamic lag to peak (Knutson *et al.*, 2007). Outliers that exceeded 3 standard deviations from mean activity were considered likely to be caused by artifacts rather than signal and excluded from further analyses (M = 2.13 trials per each participant, ranging from 0 to 13). The same patterns remained without excluding any outliers. Additionally, one trial in which participants responded within 50 ms 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 the previous work from our group (Fareri *et al.*, 2012, 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 and 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 the mechanisms of impression updating. Specifically, we examined whether MRI participants: (i) treated all interactions with both players similarly (Baseline model); (ii) represented positive (giving money) and negative outcomes (taking money) differently [Loss Gain (LG) model]; (iii) 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 [LG-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. For all models, PE was initialized at 0 (Fareri *et al.*, 2012, 2015), as was the initial value of each player. The first trial for each player was excluded from subsequent analyses to eliminate any bias. 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 four free parameters to test the hypothesis that participants would treat all experiences in the task similarly when updating their ratings of the players. We used a simple Rescorla–Wagner PE rule to update the player value on each trial (t) after players' decisions to give or take a given amount of money γ were revealed. We implemented a single learning rate α in this model for all possible combinations of players and outcomes to reflect participants' lack of differentiating giving or taking money as a function of player (equation 1):

$$\mathbf{V}_i\,(t+1) = \mathbf{V}_i(t) + \boldsymbol{\alpha} * (\boldsymbol{\gamma}(t) - \mathbf{V}_i(t)) \tag{1}$$

where the difference between the experienced outcome γ and the value of a player i on a trial t represents PE (∂). Initial ratings and player value were then used to separately update ratings of trustworthiness (y) and closeness (j) (equation 2):

$$\hat{R}_{iyj} = \boldsymbol{b}_0 + \boldsymbol{b}_1 R_{iyj} + \boldsymbol{b}_2 V_i \tag{2}$$

LG model. We next formalized a model with five 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). We implemented two learning rate parameters here to capture differential updating based on positive or negative social outcomes (y) (equation 3):

$$V_i\,(t+1) = V_i(t) + \alpha_y * (\gamma(t) - V_i(t)) \tag{3}$$

Ratings were then predicted by the same linear model as described in equation (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 (equation 4):

$$\mathbf{V}_i\,(t+1) = \mathbf{V}_i(t) + \boldsymbol{\alpha}_i * (\boldsymbol{\gamma}(t) - \mathbf{V}_i(t)) \tag{4}$$

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 of friends and strangers. We formalized a model with seven free parameters and allowed for the updating of player value with separate learning rates for each possible combination of player (i) and outcome (y) (equation 5):

$$\mathbf{V}_i\,(t+1) = \mathbf{V}_i(t) + \boldsymbol{\alpha}_{iy} * (\boldsymbol{\gamma}(t) - \mathbf{V}_i(t)) \tag{5}$$

Dynamic LGFS model. Finally, 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 (equation 6):

$$\hat{R}_{iyj} = b_0 + b_1 R_{iyj} + b_2 V_i + b_3 V_i * R_{iyj} \quad (6)$$

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 (equation 7):

$$SSE = \left(\hat{R}_{iyj} - R_{iyj}\right)^2 \quad (7)$$

Results

Hypothesis 1: participants update less for friends *vs* 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. For this and further analyses, we conducted a linear mixed-effects regression using the R package nlme (Pinheiro *et al.*, 2019). Since including participant gender and trial-by-trial amount did not change our findings, we dropped them from further analyses (see Supplementary Sections 5 and 6 for findings with participant gender and amount in the model).

First, we found the main effects of Player ($b = 1.68$, SE =0.02, $t = 71.80$, $P < 0.001$), Valence ($b = 0.57$, SE =0.02, $t = 24.22$, $P < 0.001$) and Task ($b = -0.14$, SE =0.02, $t = -6.03$, $P < 0.001$), indicating higher ratings for (i) friend ($M = 6.49$, SE =0.10) *vs* stranger ($M = 3.12$, SE =0.10), (ii) giving ($M = 5.37$, SE =0.10) *vs* taking ($M = 4.24$, SE =0.10) and (iii) trustworthiness ($M = 4.94$, SE = 0.10) *vs* closeness ($M = 4.66$, SE =0.10). A Player × Valence interaction ($b = -0.10$, SE =0.02, $t = -4.46$, $P < 0.001$) revealed that the difference in ratings between giving and taking was greater for strangers ($M = 1.34$, SE = 0.07) than friends ($M = 0.93$, SE = 0.07). Across all conditions, participants consistently rated their friend more positively than the stranger (Figure 2A). These effects were not modulated by any other factors (See Supplementary Section 7).

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 the main effects of Valence ($b = 0.04$, SE =0.02, $t = 2.11$, $P = 0.035$) and Task ($b = -0.09$, SE =0.02, $t = -5.33$, $P < 0.001$), indicating that the participants updated more when Player 1 gave money ($M = 1.12$, SE = 0.19) *vs* took money ($M = 1.05$, SE =0.19) and updated more for trustworthiness ($M = 1.18$, SE =0.19) *vs* closeness ($M = 0.99$, SE =0.19). Critically, we found a significant main effect of Player ($b = -0.20$, SE =0.02, $t = -11.48$, $P < 0.001$), such that the participants updated less for friend ($M = 0.89$, SE =0.19) *vs* stranger ($M = 1.29$, SE =0.19) overall (Figure 2B). These effects were not modulated by any other factors.

Hypothesis 2: impression updates for friends are explained by social PE

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 Dynamic LGFS model [Bayesian information criteria (BIC) = −18.56], which fit participants' data significantly better than our baseline model (BIC = 71.12; $z = 3.6$, $P < 0.0005$), as well as all other models [LG model (BIC: 73.64; $z = 3.63$, $P < 0.0005$); SP model (BIC: 75.04; $z = 3.49$, $P < 0.0005$); LGSP model (BIC = −1.91; $z = 3.54$, $P < 0.0005$) (Figure 3A)]. We also examined differences in estimated learning rates for updating partner value (Figure 3B). A non-parametric repeated measures analysis of variance (Friedman test) revealed a significant Player × Valence interaction on learning rates ($\chi^2 = 10.60$, $df = 3$, $P < 0.02$): participants similarly updated player value for strangers regardless of outcome valence ($\alpha_{pos,stranger} = 0.63$, $\alpha_{neg,stranger} = 0.59$; $P > 0.45$) but exhibited differential (lower) learning rates when experiencing positive ($\alpha_{pos,friend} = 0.39$) relative to negative ($\alpha_{neg,friend} = 0.49$) outcomes with a close friend ($P < 0.05$; Durbin–Conover pairwise comparisons). Participants also demonstrated lower learning rates when experiencing positive outcomes with their friend relative to a stranger ($P < 0.002$). No differences emerged between learning rates from negative outcomes for friends and strangers ($P > 0.59$).

We conducted a linear mixed-effects regression on participants' trial-by-trial PE values derived from Dynamic LGFS model, 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. The first trial from each player was excluded to avoid any bias from the initialized values. Significant main effects of Player ($b = -0.09$, SE =0.01, $t = -10.52$, $P < 0.001$) and Valence ($b = 0.79$, SE =0.01, $t = 96.14$, $P < 0.001$) indicated more positive PE values for (i) stranger ($M = 0.11$, SE =0.05) than for friend ($M = -0.06$, SE =0.05), and (ii) for giving ($M = 0.81$, SE =0.05) than for taking ($M = -0.76$, SE =0.05) conditions. These effects were modulated by a significant Player × Valence interaction ($b = -0.03$, SE =0.01, $t = -3.95$, $P < 0.001$). Participants showed more negative PE in response to friend-taking ($M = -0.82$, SE =0.05) than to stranger-taking ($M = -0.71$, SE =0.05) and more positive PE in response to stranger-giving ($M = 0.93$, SE =0.05) than to friend-giving ($M = 0.70$, SE =0.05) conditions, $Ps < 0.001$ (Figure 3C).

To confirm the findings from modeling analyses, we conducted a linear mixed-effects regression on participants' degrees of updating with 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. A significant main effect of Player ($b = -0.09$, SE =0.04, $t = -2.39$, $P = 0.017$) showed that participants updated more for the stranger ($M = 1.27$, SE =0.18) than for their friend ($M = 0.91$, SE =0.18) overall. A significant main effect of the absolute PE ($b = 0.46$, SE =0.04, $t = 11.86$, $P < 0.001$) indicated that greater PE

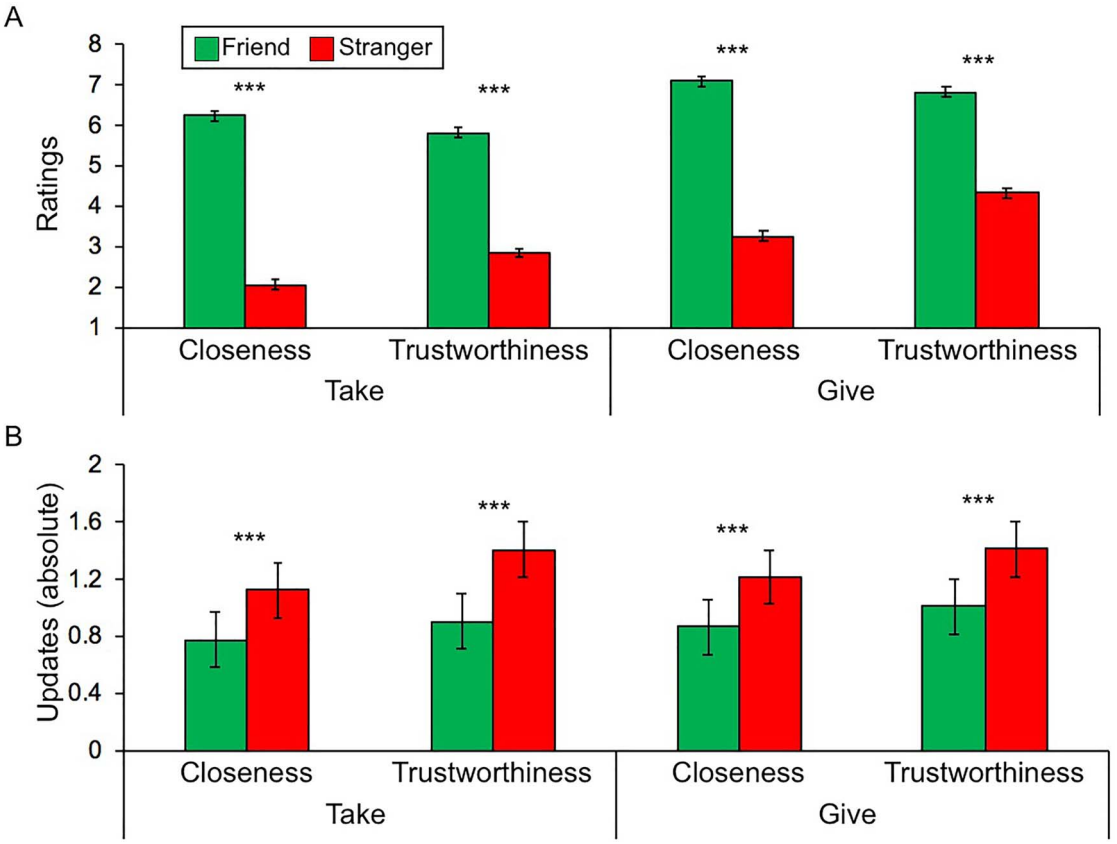

Fig. 2. (A) Differences in participants' ratings during Social Judgment Task. Participants differentiated giving and taking conditions more for a stranger than for their friend. Participants rated their friend as more positive than the stranger across all conditions. (B) Differences in participants' updates during Social Judgment Task. Participants updated less for their friend compared to the stranger across all conditions. ***$P < 0.001$.

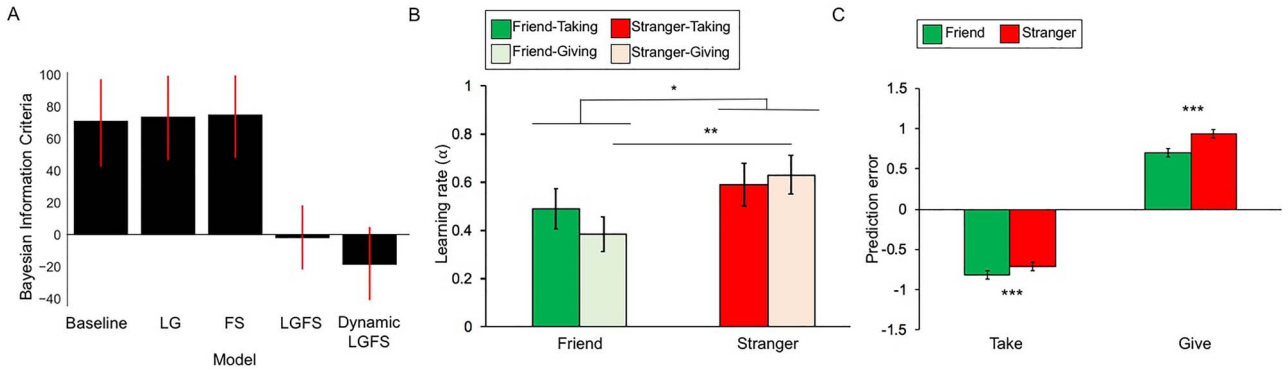

Fig. 3. (A) BIC values for models. From left, BIC values of Baseline model, LG model, FS model, LGFS model and Dynamic LGFS model. (B) Learning rate (α) values divided by conditions. Participants showed lower learning rates in the friend condition than in the stranger condition overall. Pairwise comparisons revealed that it was driven by participants' lower learning rates in the friend-giving condition than in the stranger-giving condition. (C) Differences in PE values between conditions. Participants showed more negative PE values in response to friend-taking condition than in stranger-taking condition, and more positive PE values in response to stranger-giving condition than in friend-giving condition. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

was associated with greater updating. These main effects were modulated by a significant Player × absolute PE interaction effect ($b = -0.11$, SE =0.04, $t = -3.08$, $P = 0.002$), revealing that PE was more closely related to updating for the stranger ($b = 0.57$, SE =0.05, $t = 10.63$, $P < 0.001$, 95% confidence interval [CI] = [0.47, 0.68]) than for friend ($b = 0.34$, SE =0.05, $t = 6.45$, $P < 0.001$, 95% CI = [0.24, 0.45]) (See Supplementary Section 8 for a directional modulation effect of Valence).

Hypothesis 3: greater PE was associated with increased rTPJ activity

To test whether PE signal derived from the Dynamic LGSP model accounted for trial-by-trial rTPJ activation during the Social Judgment Task, we first conducted exploratory whole-brain analyses with parametric regressors of model-derived PE values. Whole-brain analyses revealed increased subgenual anterior cingulate cortex (sgACC) activity associated with greater PE under a lenient

threshold, $P < 0.01$, cluster >166 continuous voxels, corrected $P < 0.05$ (Supplementary Section 4). Although we did not find rTPJ activity in response to the PE regressors at the whole-brain level, given our preregistered predictions, we proceeded with further analyses focused on the PSC data extracted from an rTPJ VOI.

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 PE values, and interactions between these factors as fixed effects, while individual participants were included as random effects. A main effect of Player ($b = -0.04$, SE =0.01, $t = -2.59$, $P = 0.010$) revealed that participants showed overall lower rTPJ activity in response to their friend ($M = -0.03$, SE =0.02) than to a stranger ($M = 0.05$, SE =0.02), although directionally this effect was more prominent in the taking condition than in the giving condition (Figure 4A). More importantly, there was a significant Player × Valence × PE interaction ($b = 0.03$, SE =0.01, $t = 2.08$, $P = 0.037$), indicating that more negative PE in the friend-taking condition was associated with increased rTPJ activity ($b = -0.07$, SE =0.03, $t = -2.24$, $P = 0.025$, 95% CI = [−0.13, −0.01]), while PE signal from other conditions did not significantly track rTPJ activity (friend-giving: $b = -0.01$, SE =0.03, $t = -0.23$, $P = 0.819$, 95% CI = [−0.06, 0.05]; stranger-taking: $b = 0.04$, SE =0.03, $t = 1.47$, $P = 0.141$, 95% CI = [−0.01, 0.09]; stranger-giving: $b = -0.01$, SE =0.03, $t = -0.49$, $P = 0.625$, 95% CI = [−0.07, 0.04]) (Figure 4B). This effect was not modulated by any other factors.

Hypothesis 4: 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 associated with more positive ratings ($b = -0.09$, SE =0.05, $t = -1.95$, $P = 0.051$), this effect was qualified by a Player × Valence × rTPJ interaction ($b = 0.09$, SE =0.05, $t = 1.89$, $P = 0.059$). 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 ($b = -0.22$, SE =0.10, $t = -2.29$, $P = 0.022$, 95% CI = [−0.42, −0.03]) (Figure 4C). RTPJ activity was not associated with ratings in any other conditions (friend-giving: $b = -0.07$, SE =0.10, $t = -0.70$, $P = 0.487$, 95% CI = [−0.26, 0.12]; stranger-taking: $b = 0.06$, SE = 0.10, $t = 0.63$, $P = 0.526$, 95% CI = [−0.13, 0.25]; stranger-giving: $b = -0.15$, SE =0.10, $t = -1.55$, $P = 0.122$, 95% CI = [−0.34, 0.04]). These effects were not modulated by other factors (see Supplementary Section 9 for other main effects and interactions).

Discussion

The present research sought to examine impression updating for close friends and strangers, paying special attention to the contributions of motivated cognition and PE. 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: (i) participants engaged in less updating overall for friends *vs* strangers, reflecting participants' motivation to protect their positive impressions of their friends; (ii) 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 PE update rule; (iii) rTPJ activity was associated with PE evoked by a friend's negative behavior—in the condition in which a friend took money from the participant, greater PE corresponded to greater rTPJ activity and (iv) reduced updating for participants' friends 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.

Together, first, the findings suggest that people's failure to negatively update their impressions of a close friend is accounted for by positive prior impressions of friends, which subsequently drive lower learning rates. 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 morality (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 *vs* the stranger shaped their subsequent evaluations about them over the course of the experiment, contributing to reduced updating for friends *vs* strangers.

Second, 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 PE. These patterns indicate that participants' failure to represent unexpected negative information about a friend, as opposed to rationally explaining away the negative information, contributed to participants' maintenance of their positive impressions of the friend (Kim *et al.*, 2020; Park *et al.*, in press). Moreover, participants' prior experiences with their friend, measured by how long participants reported knowing their friend and how many hours per week participants reported spending with their friend, were not associated with participants' impression updating for their friend in separate analyses (Park and Young, 2020). Of course, participants' motivation to protect positive impressions of their friends would need to be directly assessed in future research. Nevertheless, the present results provide initial support for a motivational account of participants' resistance to updating impressions of their friends, above and beyond the influence of their strong priors. Notably, when participants overcame this bias with successful recruitment of rTPJ, they were more likely to negatively rate their friend, reflecting the impact of PE.

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 TPJ is associated with favorable evaluation of ingroup members (Baumgartner *et al.*, 2012), other studies have found that decreased TPJ activity is associated with more positive

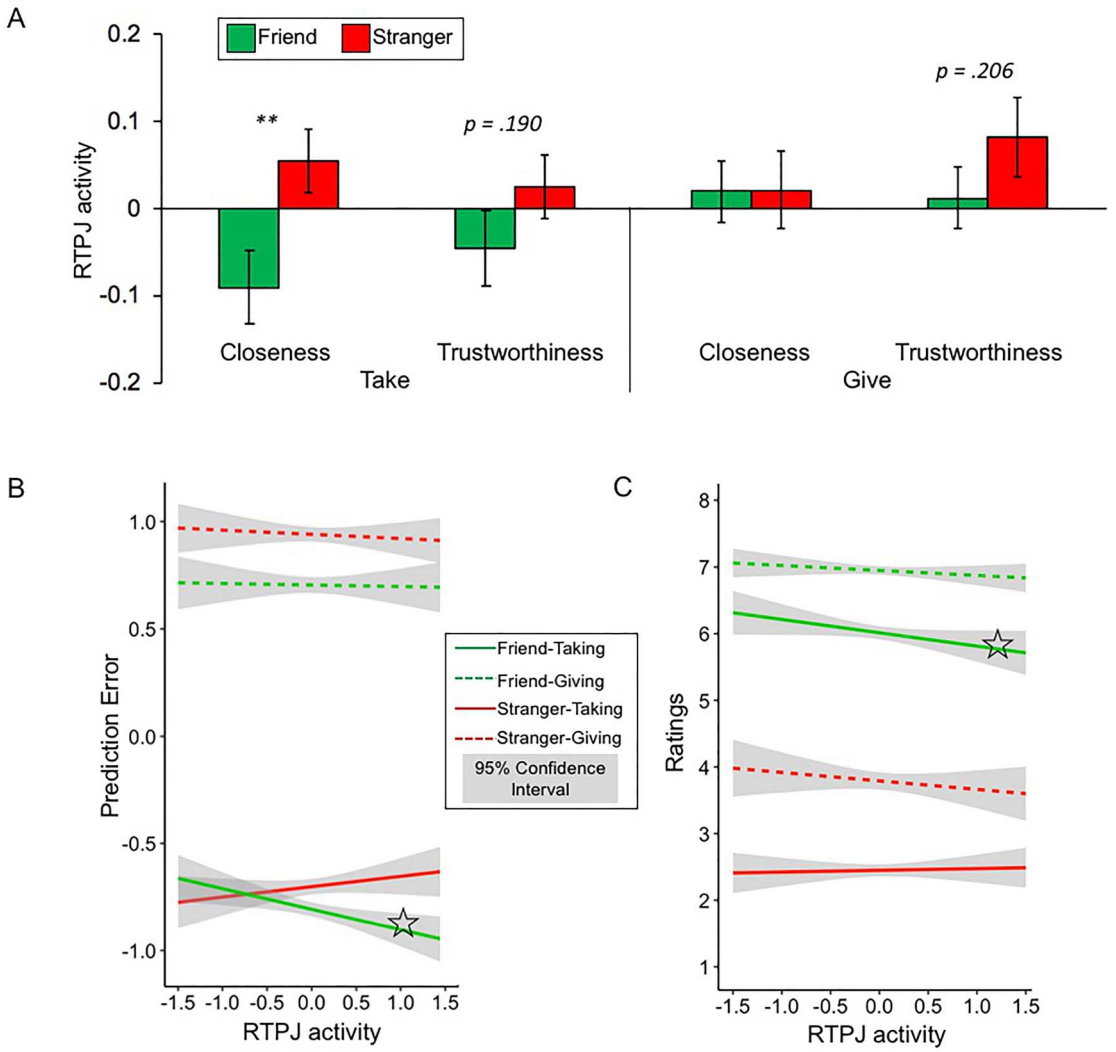

Fig. 4. (A) Differences in rTPJ activity between conditions. Participants showed reduced rTPJ activity in response to friend than to stranger. These effects were more prominent in the taking condition. (B) Association between rTPJ activity and prediction error. Participants' trial-by-trial rTPJ activity was significantly associated with their prediction error signal evoked when their friend took money from them. The more negative the PE participants experienced in response to their friend's taking behavior, the greater rTPJ activity they showed. (C) 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. **$P < 0.001$; *95% CI did not contain 0.

ingroup perception (Kliemann *et al.*, 2008; Hughes *et al.*, 2017; Park *et al.*, 2017). We suggest that TPJ may function as the hub for coordinating a response based either on motivated cognition or PE-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 and 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.*, 2020; Park *et al.*, in press).

While the present work was not motivated by hypotheses regarding differential associations between rTPJ activity and PE signals evoked by friends *vs* strangers, we did find that rTPJ activity was uniquely associated with PE for the condition in which a friend took money. Participants may have devoted more mentalizing effort to resolving PE in this condition, in attempting to understand their friends' intent behind their taking behavior (Zaki *et al.*, 2016). Another possibility is that the psychological experience of surprise occurs primarily when participants' strong priors are violated, which occurs uniquely in the friend-taking condition. Directly testing this possibility would be useful in an exploration of whether the degree of required mentalizing and the subjective importance of a specific condition can modulate the link between PE and rTPJ activity. Interestingly, in spite of the relationship between rTPJ and PE 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 PE for updating. We note that our modeling approach was closely informed by previous work from our group (Fareri *et al.*, 2015), but that alternative approaches may also be informative. For example, future work may further interrogate motivated impression updating via implementation of alternative classes of models that may capture the hidden mechanistic biases at play when failing to update impressions of close others after instances of negative behavior (e.g. Dorfman *et al.*, 2019)

In line with recent proposals, the current findings suggest that a primary function of TPJ is to support the processing of social PE and updating. For example, some studies have suggested that TPJ regions are associated with impression updating and PE encoding in contexts that do not involve mentalizing (Wittmann *et al.*, 2016; Lockwood *et al.*, 2018). In the current study, rTPJ activity was uniquely associated with social PE and ratings in the case of a friend's bad behavior.

In line with other literature implicating components of the reward circuit (e.g. medial prefrontal cortex, striatum) in encoding PE and social learning (Garrison *et al.*, 2013; Fareri *et al.*, in press), we observed correlates of a social reward PE signal that did not differentiate friends and strangers in sgACC and anterior caudate nucleus. 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), learning about the generosity of others (Hackel *et al.*, 2015) and representing social learning signals (for a review, see Lockwood and Wittman, 2018). Thus, it is possible that reward-related regions in this study encode a more generic social PE signal. 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 PEs evoked by friends and strangers are processed in dissociable networks from monetary outcomes (c.f. Behrens *et al.*, 2008; Stanley, 2016). Moreover, exploring functional connectivity between sgACC and rTPJ would be an interesting next step in investigating how social PE signals supported by these regions may be integrated and reflected in impression updating.

Forming, maintaining and revising impressions of other agents are 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.

Supplementary data

Supplementary data are available at *SCAN* online.

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.

Funding

This work was supported by a grant from the John Templeton Foundation [5107321] awarded to L.Y. and NIH Shared instrumentation grant [S10OD021569] awarded to McGovern Institute for Brain Research at MIT.

Conflict of interest

The authors declare that there is no conflict of interest.

References

Akaike, H. (1974). A new look at the statistical model identification. In: Parzen, E., Tanabe, K., Kitagawa, G., editors. *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics)*, New York, NY: Springer.

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.

Asch, S.E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, **41**(3), 258–90.

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**, 1452–69.

Behrens, T.K.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, **456**, 245–9.

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.

Crocker, J., Hannah, D.B., Weber, R. (1983). Person memory and causal attributions. *Journal of Personality and Social Psychology*, **44**(1), 55–66.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage*, **55**, 705–12.

Dorfman, M.H., Bhui, R., Hughes, B.L., Gershman, S.J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, **30**(4), 516–25.

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), 8170–80.

Fareri, D.S., Chang, L.J., Delgado, M.R. (2020). Neural mechanisms of social learning. In: Gazzaniga, M.S., Mangun, G.R., Poeppel, D., editors. *The Cognitive Neurosciences, 6th Edition*. Cambridge, MA: MIT Press.

FeldmanHall, O., Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, **3**, 426–35.

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), 3602–11.

Garrison, J., Erdeniz, B., Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, **37**(7), 1297–310.

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**, 1233–5. doi: 10.1038/nn.4080.

Hahn, U., Harris, A.J.L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In: Brian, H.R., editor. *Psychology of Learning and Motivation*, San Diego, CA: Academic Press.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic

interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(18), 6741–6.
Hill, C.A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J.P., Ruff, C.C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, **20**(8), 1142–9.
Hughes, B.L., Beer, J.S. (2012). Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. *Cerebral Cortex*, **22**, 1372–81.
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. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Science*, **24**(2), 101–11.
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**, 2949–57.
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**, 836–48.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, **108**(3), 480–198.
Lockwood, P.L., Wittmann, M. (2018). Ventral anterior cingulate cortex and social decision-making. *Neuroscience and Behavioral Reviews*, **92**, 187–91.
Lockwood, P.L., Wittmann, M.K., Apps, M.A.J., *et al.* (2018). Neural mechanisms for learning self and other ownership. *Nature Communications*, **9**, 4747.
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), 19406–15.
Mende-Siedlecki, P., Cai, Y., Todorov, A. (2013b). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, **8**(6), 623–31.
Murray, S.L. (1999). The quest for conviction: motivated cognition in romantic relationships. *Psychological Inquiry*, **10**(1), 23–34.
Park, B., Young, L. (2020). An association between biased impression updating and relationship facilitation: a behavioral and fMRI investigation. *Journal of Experimental Social Psychology*, **87**, 103916.
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), 1083–96.
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., editors. *The Neural Basis of Mentalizing*, New York: Springer.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2019). nlme: Linear and nonlinear mixed effects models. R package version 3.1-143. https://CRAN.R-project.org/package=nlme
Rescorla, R., Wagner, A. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcment and non-reinforcement. In: Black, A.H., Prokasy, W.F., editors. *Classical Conditioning II*, New York: Appleton-Century-Crofts, pp. 64–99.
Siegel, J.Z., Mathys, C., Rutledge, R.B., Crockett, M.J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, **2**, 750–6.
Stanley, D.A. (2016). Getting to know you: general and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience*, **11**(4), 525–36. 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), 403–8.
Thornton, M.A., Mitchell, J.P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, **28**, 3505–20.
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*, **19**, 985–97. doi: 10.3758/s13415-019-00697-0.
Vonk, R. (1994). Trait inferences, impression formation, and person memory: strategies in processing inconsistent information about persons. *European Review of Social Psychology*, **5**(1), 111–49. doi: 10.1080/14792779543000039.
Wittmann, M.K., Kolling, N., Faber, N.S., Scholl, J., Nelissen, N., Rushworth, M.F.S. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron*, **91**, 482–93.
Wright, J.C., Mischel, W. (1988). Conditional hedges and the intuitive psychology of traits. *Journal of Personality and Social Psychology*, **55**(3), 454–69.
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), 1270–82. doi: 10.1162/jocn_a_00978.