Does biased impression updating facilitate close relationships?

A behavioral and fMRI investigation

BoKyung Park

Liane Young

Boston College

140 Commonwealth Avenue

Chestnut Hill, MA 02467

AUTHOR'S NOTE: This research was funded by a grant from the John Templeton Foundation [5107321] awarded to Liane Young. The authors would like to thank Mauricio Delgado, Dominic Fareri, Emma Alai, and Mariel Kronitz for their contributions. Correspondence should be addressed to BoKyung Park (parkanj@bc.edu).

Keywords: Impression updating; Ingroup bias; Relationship maintenance

Abstract

Does ingroup bias provide any benefit for maintaining close relationships? We examined the possible effect of biased impression updating for ingroup members (i.e., friends) for relationship maintenance, as measured by how many friends participants reported having (Studies 1 and 2). We also investigated the underlying neural mechanism, focusing on the activity in the right temporo-parietal junction (RTPJ), a region involved in moral evaluation and updating (Study 2). Specifically, we tested whether selectively discounting negative information about close others, manifesting in reduced impression updating, and indexed by reduced RTPJ activity, is linked to maintaining close relationships. In Study 1, after imagining a friend and a stranger performing different positive and negative behaviors, participants who were reluctant to update how close they felt to their friend (friend-closeness) reported having more friends in real life. In Study 2, participants were led to believe that a friend and a stranger gave money to them or took money from them, while in the scanner. Participants who engaged in less negative updating of friends versus strangers reported having a greater number of friends. Participants who engaged in less friend-closeness updating also showed reduced RTPJ activity when their friend took money; and this neural pattern was associated with reporting having more friends. Together, these findings suggest that selectively discounting close others' negative behavior may contribute to maintaining close relationships, indicated a potential social benefit of ingroup bias.

Ingroup bias, the tendency to judge ingroup members more favorably than outgroup members (Tajfel, 1982), has been associated with a wide range of negative social outcomes: distorted perception about outgroups (Xiao, Coppin, & Van Bavel, 2016), reduced intergroup cooperation (Balliet, Wu, & Dreu, 2014; Sherif, 1966), and prejudice and discrimination (Brewer, 1999), just to name a few.

A key manifestation of ingroup bias is inaccurate impression formation and updating across group boundaries. Although accurately evaluating others is essential in many ways for successful social interactions (Shin, Kim, & Han, 2014; Zaki & Ochsner, 2011), people are motivated to see ingroup members, including close others, in a more positive light (Brewer, 1999; Taylor & Brown, 1988). Ingroup bias can lead people to forget more negative information about ingroup members (Howard & Rothbart, 1980), discount ingroup members' harmful intentions in the service of blame mitigation (Monroe & Malle, 2019), and preferentially forgive ingroup members (Baumgartner, Götte, Gügler, & Fehr, 2012; Schiller, Baumgartner, & Knoch, 2014; Wohl & Branscombe, 2009). These findings reveal suboptimal processing of information about others' negative behavior.

Relatively few studies have examined the potential benefits of bias in impression formation and updating, for maintaining or protecting close relationships, for example. Extant literature on close relationships has revealed the relatively powerful impact of negative versus positive events on the quality of a relationship (for a review, see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001), given that negative behaviors are recognized better (Acitelli, Douvan, & Veroff, 1993; Alves et al., 2015; see also Kellermann, 1984) and reciprocated more (Levenson & Gottman, 1985; Rusbult, Johnson, & Morrow, 1986). Other work has shown that greater motivation to bond with partners leads people to discount their partner's unresponsive behavior by blaming their own failure to disclose their needs, a pattern that in turn predicts greater trust in partners (Lemay & Clark, 2015; Lemay & Melville, 2014). This body of work

suggests the possibility that selectively discounting close others' negative behavior may be related to relationship facilitation; we explore this possibility in the present research.

Previous research on brain regions for thinking about agents' mental states, also known as mentalizing, mindreading, or theory of mind (ToM), provides some clues to the neural mechanisms underlying impression updating and also biased impression updating. The mentalizing network has been associated with detecting social prediction error, i.e., the gap between expectation and observation (Koster-Hale & Saxe, 2013), and ultimately updating social impressions given new information (Baron, Gobbini, Eengell, & Todorov, 2011; Boorman, O'Doherty, Adolphs, & Rangel, 2013; Mende-Siedlecki, Baron, & Todorov, 2013; Mende-Siedlecki, Cai, & Todorov, 2013; Thornton & Mitchell, 2018; see also Mende-Siedlecki & Todorov, 2016). For example, in a previous study, participants exhibited reduced activity in RTPJ, a key node of the mentalizing network, when a previously fair versus previously unfair target brought about a negative outcome (Kliemann, Young, Scholz, & Saxe, 2008); these fair targets were judged as less blameworthy and less intentional. These behavioral and neural patterns indicate that failure to accurately detect or evaluate previously fair targets' negative behaviors may be related to reduced mentalizing, as indexed by activity in RTPJ.

Other work has examined this effect more directly in the context of ingroup bias. In one study, participants were presented with targets from their own group or an outgroup (Hughes, Zaki, & Ambady, 2017). Brain activity was measured when ingroup and outgroup targets behaved in morally unexpected ways, based on their prior behaviors. Failure to recruit mentalizing regions in response to ingroup members' unexpected negative behaviors was associated with reduced impression updating. These findings again suggest that the failure to recruit mentalizing regions may underlie the failure to engage in negative impression updating of ingroup members.

The present work seeks to build on and extend the prior literature on ingroup bias and impression updating. First, we aim to replicate prior research showing that people engage in

biased updating (by discounting ingroup members' bad behavior), and that this effect is accounted for by reduced recruitment of mentalizing regions such as RTPJ. Critically, we also investigate the novel possibility that biased updating is associated with some social benefits. Our primary hypothesis is that overlooking close others' negative behavior, indexed by disengagement of mentalizing regions such as RTPJ, results in maintaining more close relationships, measured by the number of friends that people report having.

In Study 1, we examined the association between biased impression updating and reported number of close relationships. Specifically, we measured the degree to which participants updated their evaluations about a close friend versus a stranger when they learned new negative information about these agents, and we tested the association between update magnitude and reported number of friends. In Study 2, we investigated the underlying neural mechanisms of this effect by presenting participants with positive and negative behaviors performed by a close friend and stranger in the scanner and then examining the association between RTPJ activity and reported number of friends, as well as the degree to which participants updated their impressions of their friend.

Study 1

In Study 1, we explored whether participants' impression updates are linked to social outcomes. In particular, we predicted that participants showing resistance to impression updating for friends in the experimental paradigm also report having more friends in real life. For this study and Study 2, we report all measures, manipulations and exclusions.

Method

Participants. One hundred and twenty-five Mturkers (37.6% female; age M = 35.34, S.D. = 10.65), divided into two groups, one week apart from each other[1], were recruited for this study.

[1] After conducting a preliminary data analysis, we decided to double the sample size to capture a small to moderate effect size.

Fourteen responses that did not fit our inclusion criteria were removed[2], leaving 111 final participants. We found similar patterns when including all participants. A sensitivity power analysis using a significance level α = .05 and power β = .80 showed that this sample size is sufficient to detect effect size f^2 = .07[3]. Sample size was based on the expectation of a small to moderate effect size.

Materials. We used eight stories describing different moral and immoral behaviors of a target person from prior work (Mende-Siedlecki, Baron, et al., 2013). The moral and immoral behaviors were matched in terms of moral relevance, arousal, valence, and perceived frequency (*ps* > .186).

Procedure. After providing informed consent, participants were asked to enter their same-sex best friend's first name. Participants then completed the Inclusion of Other in Self scale (IOS; Aron, Aron, & Smollan, 1992), consisting of pairs of circles representing themselves and their friend, with increasing overlaps between circles indicating increasing closeness. Those who chose the 4th pair or more on this 7-point scale (M = 5.32, S.D. = .90) were recruited. After the IOS question, participants delivered two different ratings, trustworthiness ("To what extent is [Name] trustworthy?") and closeness ("How close are you and [Name]?"), on 8-point scales, about their friend and a gender-matched stranger ("Jacob" for male participants and "Emily" for female participants) ("pre-exposure ratings").

Next, participants read 8 different behaviors and were asked to imagine that their friend or the stranger ("target") performed these behaviors in the past. Their friend's name and the stranger's name were piped into the stories. Participants first read two stories describing two positive behaviors that one of the targets performed (e.g., "[Name] helped an elderly woman get up from her wheelchair"), followed by two stories describing two negative behaviors performed

[2] We excluded responses from repeat IP addresses (N = 2), nonsensical responses to open-ended questions (N = 8), and responses where participates rated a stranger as closer to them than their friend (N = 4).
[3] Although our main statistical analyses used ordinal regressions, we based our power analyses on multiple regression models given that G*Power does not provide an ordinal regression option.

by the same target (e.g., “[Name] spread rumors about a coworker’s work ethic”). Afterwards, participants read two stories describing two positive behaviors performed by the other target and two stories describing two negative behaviors performed by that target, consecutively. We counterbalanced the following conditions: (1) presentation order of two stories within each positive and negative block, (2) which target was presented first, and (3) which behaviors were presented as performed by friend or stranger. Because our primary interest is in how people update their impressions based on new negative behaviors, positive stories were always presented first, followed by negative stories. After reading each story, participants made ratings of trustworthiness (“How trustworthy is [Name]?”) and closeness (“How close are you and [Name]?”) for the target agent in the story on 8-point scales, based on all the information available up to that point.

Participants were then asked to evaluate both trustworthiness and closeness in general (as described above for the “pre-exposure ratings”) for the target (“post-exposure ratings”), followed by questions about the number of their friends (1 = 0-2 friends; 2 = 3-5 friends; 3 = 6-9 friends; 4 = 10-19 friends; 5 = 20-49 friends; 6 = 50-99 friends; and 7 = more than 100 friends) and how often they make new friends (1 = Almost every week; 2 = A few weeks; 3 = Every month; 4 = A few months; 5 = Every year; 6 = A few years; and 7 = About every 10 years), along with other items not explored here (See Supplementary Section 1). Finally, participants answered demographic questions and were thanked and compensated. All procedures were approved by the Institutional Review Board at Boston College.

Analyses and results

To measure the extent to which participants updated their impressions of their friends, we subtracted participants’ ratings for the last negative story from their ratings for the last positive story, generating four updating scores (updates in friend-closeness, friend-trustworthiness, stranger-closeness, and stranger-trustworthiness ratings). Approximately half of participants did not update their closeness ratings (51.4% for friend, 48.6% for stranger), so we

recoded participants into three categories; those who never updated their closeness ratings ("No update" = 0; friend-closeness N = 57; stranger-closeness N = 54), those who decreased their closeness ratings ("Negative update" = -1; friend-closeness N =53; stranger-closeness N = 55), and those who unexpectedly increased their closeness ratings ("Positive update" = +1; friend-closeness N = 1; stranger-closeness N = 1) after reading negative stories. The findings were similar when excluding people who unexpectedly engaged in positive updating (See Supplementary Section 2 for changes in evaluations between the last positive story and the last negative story).

We examined whether participants who were more resistant to updating their friend-closeness ratings reported having a greater number of friends. Ordinal regression on the number of friends that participants reported having revealed that those who never updated, and those who positively updated, reported having more friends, Estimate = .73, S.E. = .34, Wald = 4.61, Odds ratio = 2.07, p = .032, 95% CI for Estimate = [.06, 1.39], compared to those who negatively updated their friend-closeness ratings (Figure 1[4]). This effect remained consistent after we controlled for how often participants made new friends, as well as updates in friend-trustworthiness, stranger-closeness, and stranger-trustworthiness. Importantly, when entered in the same model, none of the updates in other impressions was significantly associated with the number of friends participants reported having, ps > .48, Odds ratio = .88 - 1.07[5].

As a validity check, we subtracted participants' closeness ratings in their post-exposure evaluation (i.e., after they read all stories) from those in their pre-exposure evaluation (i.e., before they read any stories), generating pre-post closeness updates. Using this new metric, we found again that participants who more negatively updated their friend-closeness ratings reported having fewer friends, Estimate = -.61, S.E. = .24, Wald = 6.40, Odds ratio = .54, p

[4] We did not plot "Positive update" group in the figure as this group was an N of 1.
[5] We used raw update values for trustworthiness ratings given that more participants varied in these ratings, compared to closeness ratings. However, even when we categorized participants into "Never update," "Negative update," and "Positive update" groups as with the closeness ratings, we found no significant effects.

= .011, 95% CI for Estimate = [-1.09, -.14]. Again, these effects did not change after we controlled for how often participants made new friends and updates in friend-trustworthiness, stranger-closeness, and stranger-trustworthiness[6] (See Supplementary Section 3 for correlations between covariates in regression models).

Study 1 discussion

In Study 1, we found that people who were more resistant to updating how close they felt to their friend after learning about their friend's negative behavior reported having more friends, suggesting that biased impression updating in favor of close others might contribute to maintaining more close relationships. Although we did not expect distinctive effects of updating for friend-closeness versus friend-trustworthiness, participants' updates in friend-trustworthiness, as well as in stranger-closeness and stranger-trustworthiness, were not associated with number of friends.

In this study, participants were asked to *imagine* that their friend and the stranger committed the behaviors described in experimental stimuli. In the next study, we asked participants to bring their own friends to the scanning session, where participants and their friends engaged in ostensibly real-time interactions while participants' brain activity was measured. Thus, Study 2 aimed to further explore the mechanisms that contribute to biased impression updating and relationship maintenance.

Study 2

Study 2 extended Study 1 in two ways. First, we had participants bring their close friends with them to the scan session, where they observed positive and negative behaviors performed by their friend. Moreover, the friends' behaviors targeted the participants rather than a third

[6] A majority of the participants did not update their ratings between before and after reading the stories especially for friends (77.3% for friend-closeness, 68.2% for friend-trustworthiness, 67.3% for stranger-closeness, and 32.7% for stranger-trustworthiness). However, we found similar effects after replacing the continuous pre-post closeness updates with categorized closeness updates (No update, Negative update, and Positive update).

party. Thus, unlike to Study 1, in which participants were asked to merely imagine that their friend engaged in certain behaviors, in Study 2, participants were led to believe that their friends were actually behaving positively and negatively toward the participants in real time.

Second, we examined the underlying neural mechanisms by which people discount their friend's negative behavior and maintain close relationships. Specifically, we predicted that disengagement of mentalizing in response to friends' negative behavior (indexed by reduced RTPJ activity) would be correlated with (1) reduced updates in friend-closeness, suggesting that decreased RTPJ activity accounts for biased impression updating for close others, and (2) a greater number of friends, supporting the argument that overlooking close others' negative behavior may contribute to maintaining more close relationships.

Method

Participants. We recruited thirty right-handed and neurologically and psychologically intact participants to bring a close same-sex friend with them to the scanner. Participants who selected equal to or greater than 4-points on the same IOS scale as in Study 1 were recruited. Six participants were excluded from further analyses[7], leaving a total of 24 participants in the final sample. Since this dataset was initially collected for a different study ([Blinded for peer review]), we did not determine the sample size in advance. All procedures were approved by the Institutional Review Boards at Boston College and the Massachusetts Institute of Technology.

Social judgment task. Participants were informed that they would play a game ("Social Judgment Task") with two different targets, their friend and a gender-matched stranger whom they met at the scanner. Participants were instructed as follows: In the game, there are two different roles, "Player 1" and "Player 2." On each trial, each player receives $20 for use in that trial. Player 1 decides how much to give to or take from Player 2 in $5 increments; Player 2

[7] Because of excessive head movement (three participants), a structural abnormality (one participant), a deviant expectation about their friend (one participant), or completing fewer than half of the trials (one participant). Participants who completed trials equal to or more than 50% of all trials were included in the final sample.

passively views Player 1's decision. After that, participants were told that they were randomly assigned to play as Player 2, and they would see their friend and the stranger taking turns as Player 1. In reality, all Player 1 decisions were pre-programmed.

As Player 2, participants were asked to make either closeness or trustworthiness ratings about Player 1 in each trial after seeing Player 1's decision. During the game, participants were first presented with the identity (friend or stranger) of Player 1 on each trial (2s), followed by rating type (2s), closeness or trustworthiness. After a jittered fixation (2-6s), participants observed how much Player 1 gave or took (2s; $5, $10, $15, or $20); and then after another jittered fixation (2-6s), participants were able to make their ratings on an 8-point scale (4s). Each trial was divided by a jittered fixation (2-6s) (Figure 2). Participants were told that one of the trials would be randomly selected, and they and Player 1 would receive the amount of money received in that trial in addition to their base compensation.

Procedure

Participants arrived at the scan session with their friend, and met a gender-matched confederate (stranger). Pre-scan impressions were measured by questions asking the participants and their friend to rate how trustworthy they felt their friend and the stranger to be, and how close they felt to their friend and to the stranger ["Pre-scan evaluation"]. All three people were then instructed together about the game structure; real participants were escorted to a separate scanning area. There, participants were told that they were assigned to play as Player 2 in the game and would make trustworthiness and closeness ratings about Player 1 at the end of each trial. Participants then completed 8 practice trials of the game and entered the scanner. Once inside the scanner, participants completed 192 trials of the game (16 trials in each of 12 runs, total time = 74min 24sec), while functional scans were acquired.

After completing the Social Judgment Task, participants were presented with two runs of a theory-of-mind localizer task (ToM; 10 trials in each run; total time = 9min 4sec) (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011), composed of conditions in which participants had to infer

another person's mental states ("belief" condition) or physical representations of an object ("photo" condition). Afterwards, participants exited the scanner, made the general trustworthiness and closeness ratings again for their friend and the stranger ["Post-scan evaluation"][8], and were debriefed and compensated.

To assess the number of participants' close relationships, we contacted all participants 2-7 months after the scanning session for a follow-up survey. Sixteen out of 24 total participants completed this survey, responding to questions about the number of their friends (1 = 0-2 friends; 2 = 3-5 friends; 3 = 6-9 friends; 4 = 10-19 friends; 5 = 20-49 friends; 6 = 50-99 friends; and 7 = more than 100 friends) and how often they make new friends (1 = Almost every week; 2 = A few weeks; 3 = Every month; 4 = A few months; 5 = Every year; 6 = A few years; and 7 = About every 10 years) along with other items[9]. A sensitivity power analysis using a significance level α = .05 and power β = .80 showed that this sample size was sufficient to detect correlations ≥ .64 (N = 16) or ≥ .54 (N = 24)[10].

FMRI acquisition and preprocessing

We used a 3T Siemens scanner outfitted with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. Functional scans were acquired while participants were playing the Social Judgment Task and the ToM Task. Thirty-two 3x3x3mm slices of gradient echo T2* weighted echo-planar images (EPI) provided whole brain coverage (TR = 2s, TE = 30ms, flip

[8] These included exploratory items: how negative/positive, aroused, and surprised participants felt when their friend or the stranger gave or took a certain amount; how much they trust other people in daily life; how they would explain their friend's and the stranger's positive and negative behaviors; how they would plot varying relationships (ranging from a stranger – best friend) on a scale; and demographic questions. These questions are not analyzed here.
[9] Other items included IOS, how much participants liked their friend, how many hours they spent time with their friend, participants' most pleasant and unpleasant memory with their friend, and how likely they would play an economic game with their friend over a random person. These questions were administered for exploratory purposes and were not analyzed here, except for participants' pleasant and unpleasant memory recall (See Supplementary Section 4).
[10] Since eight participants did not respond to our request to complete the follow-up survey, this was the maximum level of power we could achieve. Post-hoc power analyses revealed that our power ranged from .45 to .83. Since G*Power does not provide an option for ordinal regressions, we used correlation models to calculate power.

angle = 90°). Before the functional scans, high-resolution anatomical scans were acquired (TR = 2.53s, TE = 1.69ms) while participants were looking at a blank screen.

Brain data were analyzed using Analysis of Functional Neural Images (AFNI; AFNI_16.2.06 version) software (Cox, 1996). The first six functional scans before the task of each run were removed to compensate for magnet stabilization. All other images were de-obliqued, slice timing corrected (using the first slice as reference), motion corrected (using the third volume as a reference and Fourier interpolation), spatially smoothed (using a 3D isotropic Gaussian kernel of an 8mm full width at half maximum), and normalized to average percent signal change and high-pass filtering (omitting frequencies < .01Hz, as described in Wu, Samanez-Larkin, Katovich, & Knutson, 2014).

A spherical VOI (radius = 8mm) centered on the RTPJ coordinates [57, -58, 19] derived from the whole-brain ToM t-test map contrasting belief condition versus photo condition was constructed (See Supplementary Section 5 for ToM findings). Percent signal change (PSC) data within this VOI were averaged for each 2 (Agent: Friend, Stranger) X 2 (Task: Closeness, Trustworthiness) X 2 (Valence: Taking, Giving) X 2 (Amount: Low [$5, $10], High [$15, $20]) condition, extracted, and averaged across Amount and Task conditions. Thus, final RTPJ activity values were averaged within friend-taking, friend-giving, stranger-taking, and stranger-giving conditions. Sampling was delayed by 4s to account for the hemodynamic lag to peak (Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007).

Results

To test whether changes in participants' evaluations about friend versus stranger were associated with number of relationships, we created two indexes. First, we subtracted participants' pre-scan impression from their post-scan impression, creating general impression updating scores for friend and stranger and for closeness and trustworthiness (See Supplementary Section 6 for the pre-scan versus post-scan evaluations comparisons). Then we subtracted general impression updating scores for stranger from those for friend to control for

overall tendency to update before and after the game, respectively for closeness and trustworthiness, generating “general closeness updates” and “general trustworthiness updates”. An ordinal regression on the number of friends participants reported having with general closeness updates, controlling for how often participants made new friends, revealed that the less negatively participants updated friend-closeness ratings compared to stranger-closeness ratings, the more friends they had, Estimate = .88, S.E. = .37, Wald = 5.54, Odds ratio = 2.41, p = .019, 95% CI for Estimate = [.15, 1.61] (Figure 3A; see Supplementary Section 7 for the effect of how often participants made new friends). We ran the same analysis after substituting general closeness updates with general trustworthiness updates, and found that biased general trustworthiness updates in favor of one’s friend were also associated with more friends, Estimate = 1.36, S.E. = .54, Wald = 6.33, Odds ratio = 3.88, p = .012, 95% CI for Estimate = [.30, 2.41].

Next, we tested whether participants’ RTPJ activity during the Social Judgment Task was associated with the degree to which they updated their impressions about friend versus stranger and how many friends they reported having. RTPJ activity from the friend-taking, friend-giving, stranger-taking, and stranger-giving conditions were entered in the model. The regression analysis on participants’ general closeness updates [friend-stranger] revealed that only activity from friend-taking condition significantly explained updates in closeness ratings, B = -9.63, S.E. = 3.57, $ß$ = -.62, t = -2.70, p = .014 (Figure 3B). RTPJ activity from the other conditions was not significantly associated with general closeness impression updates, |$ßs$| < .30, ps > .23. We ran the same model after substituting general closeness updates with general trustworthiness updates, but RTPJ activity from the friend-taking condition did not show a significant effect, $ß$ = -.15, p = .492 (See Supplementary Section 8).

Finally, to test for associations between participants’ RTPJ activity during the Social Judgment Task and maintenance of close relationships, we conducted an ordinal regression on the number of friends participants reported having with RTPJ activity from the friend-taking

condition as the predictor[11], controlling for how often participants made new friends. We found that greater RTPJ activity in response to the friend's taking behavior was associated with smaller number of friends, Estimate = -.10, S.E. = .05, Wald = 4.44, Odds ratio = .91, p = .035, 95% CI for Estimate = [-.19, -.01] (Figure 3C; see Supplementary Section 7 for the effect of how often participants made new friends). These effects were similar when controlling for RTPJ activity in the stranger-taking, friend-giving, and stranger-giving conditions, and without controlling for how often participants made new friends (See Supplementary Section 9 for correlations between variables in regression models).

Study 2 discussion

Study 2 provided a conceptual replication and extension of Study 1. Participants who engaged in less negative updating of friend-closeness compared to stranger-closeness reported having a greater number of friends. Moreover, these participants also showed reduced RTPJ activity when their friend took money from them, which was in turn associated with having more friends. These findings suggest that participants who disengaged in mentalizing in response to their friend's negative behavior during the Social Judgment Task reported maintaining more social relationships in real life. These findings suggest that selectively discounting negative information about close others, manifesting in reduced impression updating, and indexed by decreased RTPJ activity, can be linked to maintaining close relationships.

General Discussion

Does ingroup bias come with the specific benefit of maintaining close relationships, and, if so, how? In two studies, we found that selective processing of information about friends within an experimental task was associated with reporting a greater number of friendships in real life. In Study 1, participants who were reluctant to update how close they felt to their friend after

[11] To facilitate the interpretation of the odds ratio, we multiplied RTPJ PSC by 100 (converted from ranging between [-.26 – .19] to ranging between [-26 – 19]).

imagining their friend performing negative behaviors reported having more friends. This result suggests that selective discounting of negative information about close others may facilitate relationship maintenance and perhaps the broadening of one's social network. Study 2 explored a potential mechanism for this effect in examining activity in RTPJ, a region that has been implicated in encoding and integrating mental state information for moral judgment (Decety & Cacioppo, 2012; Young, Cushman, Hauser, & Saxe, 2007). We found that reduced RTPJ activity in response to a friend's negative behavior was associated with reporting a greater number of friends in real life. Moreover, reduced RTPJ activity in response to a friend's negative behavior also accounted for less negative updating in closeness ratings for one's friend versus a stranger. Together, the findings of Studies 1 and 2 suggest that neglecting the negative behavior of ingroup members may be associated with maintaining close relationships.

In both studies, we used two different dimensions of impression updating, closeness and trustworthiness, without no strong prediction that they would function differently. However, in Study 1, only updates in friend-closeness ratings were associated with the number of friends participants reported having, while Study 2 revealed associations for both impression dimensions—closeness and trustworthiness. We note that, in Study 1, participants were asked to *imagine* that their friend committed some bad behavior that also did not directly impact the participants themselves. By contrast, in Study 2, participants believed they were interacting with their friend (whom they brought with them to the scan session) in real time, and also that they were the direct target of their friend's behaviors. Given these differences, we think it is possible that the friend versus stranger contrast and the negative behaviors were more salient in Study 2, leading to more robust effects.

Although general believability was enhanced in Study 2 as described above, a separate question is whether participants perceived new information about their friend versus the stranger as similarly useful, reliable or informative. Put plainly, participants already know a lot about their friend, and any new data, especially data that are inconsistent with their strong prior

impressions, might be perceived as less useful or reliable, compared to new data about the stranger. Thus, in neglecting to update their impressions of their friend, participants might not be showing *bias* per se but rather a form of Bayesian-rational updating, discounting new evidence inconsistent with strong priors (Hahn & Harris, 2014). We do note, however, that discounting new evidence might also rely on generating an alternative explanation to account for the unexpected behavior (e.g., my friend took money from me because she will share the spoils after the experiment is over) (Gershman, 2019); elsewhere, we have proposed that generating this account would require increased mentalizing ([Blinded for peer review]). Given our finding of *reduced* RTPJ activity associated with biased impression updating, we think it is more likely that participants in our study were motivated to preserve their impressions of their friends and protect their relationships, in a biased fashion.

These findings raise a number of interesting questions for future work. For example, what kind of evidence and how much of it would require people to update their impressions of their friends and reconsider existing relationships? One possibility is that sufficient evidence of a friend's immorality may lead to a "turning point" in a relationship, at which point one can choose to leave the relationship. Indeed, previous research on partner choice has suggested the primacy of moral signals: people care about whether potential partners are cooperative (Hardy & Van Vugt, 2006; Jordan, Hoffman, Nowak, & Rand, 2016; Pleasant & Barclay, 2018; Sylwester & Roberts, 2010) and more likely to choose fair versus rich partners (Raihani & Barclay, 2016). Other work on essentialism reveals that people perceive moral traits to be especially essential to identity (Strohminger & Nichols, 2015) and also base their impressions of others more on morality than any other characteristics, such as warmth or competence (Brambilla, Carraro, Castelli, & Sacchi, 2019; Goodwin, 2015). Future work can consider whether people are less likely to show biased impression updating for non-moral traits. Finally, examining whether biased impression updating is associated with one's centrality within one's

social network (Weaverdyck & Parkinson, 2018) would extend the scope of work on the effect of biased impression updating on real social relationships.

Accurately forming and revising impressions about others is critical for effectively navigating the social world. However, resisting impression updates may also have the effect of facilitating close relationships. The current research advances our understanding of the potential benefits of ingroup bias, at a mechanistic level, such as consolidating social bonding within one's network. Future work may apply these findings to enhancing intergroup cooperation and negotiation.

References

Acitelli, L. K., Douvan, E., & Veroff, j. (1993). Perceptions of conflict in the first year of marriage: How important are similarity and understanding? *Journal of Social and Personal Relationships, 10*, 5-19.

Alves, H., Unkelbach, C., Burghardt, J., Koch, A. S., Krüger, T., & Becker, V. D. (2015). A density explanation of valence asymmetries in recognition memory. *Memory and Cognition, 43*, 896-909.

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*(4), 596-612.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556-1581.

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience, 6*(5), 572-581.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

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

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron, 80*, 1558-1571.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64-73.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues, 55*(3), 429-444.

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

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

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

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin and Review*, doi: https://doi.org/10.3758/s13423-018-1488-8

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38-44.

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

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32*(10), 1402–1413. https://doi.org/10.1177/0146167206291006

Howard, J. W. & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology, 38*(2), 301-310.

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.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences, 113*(31), 8658-8663. https://doi.org/10.1017/CBO9781107415324.004

Kellermann, K. (1984). The negativity effect and its implications for initial interaction.

Communication Monographs, 51, 37-55.

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

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

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

Lemay, E. P. & Clark, M. S. (2015). Motivated cognition in relationships. *Current Opinion in Psychology, 1*, 72-75.

Lemay, E. P. & Melville, M. C. (2014). Diminishing self-disclosure to maintain security in partner's care. *Journal of Personality and Social Psychology, 106*(1), 37-57.

Levenson, R. W. & Gottman, J. M. (1985). Physiological and affective predictors of change in relationship satisfaction. *Journal of Personality and Social Psychology, 49*, 85-94.

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

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

Mende-Siedlecki & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience, 11*(9), 1489-1500.

Monroe, A. E. & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, DOI: 10.1037/pspa0000137

Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological Science, 29*(6), 868-876.

Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society Open Science, 3*(11), 160510. https://doi.org/10.1098/rsos.160510

Rusbult, C. E., Johnson, D. J., & Morrow, G. D. (1986). Impact of couple patterns of problem solving on distress and nondistress in dating relationships. *Journal of Personality and Social Psychology, 50*, 744-753.

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior, 35*, 169-175.

Sherif, M. (1966). *In common predicament.* Boston: Houghton Mifflin.

Shin, Y. S., Kim, H. Y., & Han, S. (2014). Neural correlates of social perception on response bias. *Brain and Cognition, 88*, 55-64.

Strohminger, N. & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26*(9), 1469-1479.

Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters, 6*(5), 659–662. https://doi.org/10.1098/rsbl.2010.0209

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology, 33*, 1-39.

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

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

Weaverdyck, M. E. & Parkinson, C. (2018). The neural representation of social networks. *Current Opinion in Psychology, 24*, 58-66.

Wohl, M. J. A. & Branscombe, N. R. (2009). Group threat, collective angst, and ingroup

forgiveness for the war in Iraq. *Political Psychology, 30*(2), 193-217.

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.

Xiao, Y. J., Coppin, G., & Van Bavel, J. (2016). Clarifying the role of perception in intergroup relations: Origins of bias, components of perception, and practical implications. *Psychological Inquiry, 27*(4), 358-366, DOI: 10.1080/1047840X.2016.1237822

Young, L., Cushman, F., Hauser, M, & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of United States of America. 104*(20), 8235-8240.

Zaki, J. & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry, 22*(3), 159-182.

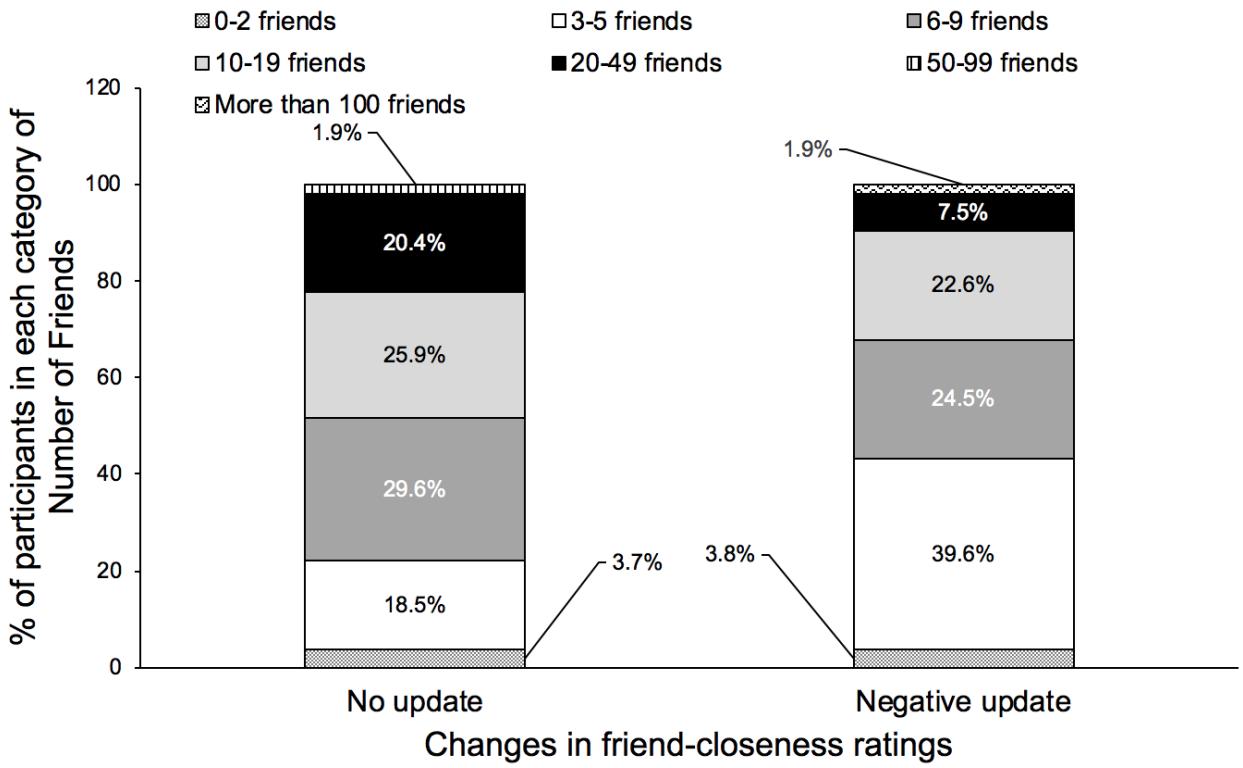

Figure 1. *Distribution of participants' reported number of friends by updates in friend-closeness ratings.* Participants who did not update friend-closeness ratings reported having a greater number of friends. For visualization, we depicted the % of participants who chose each category.

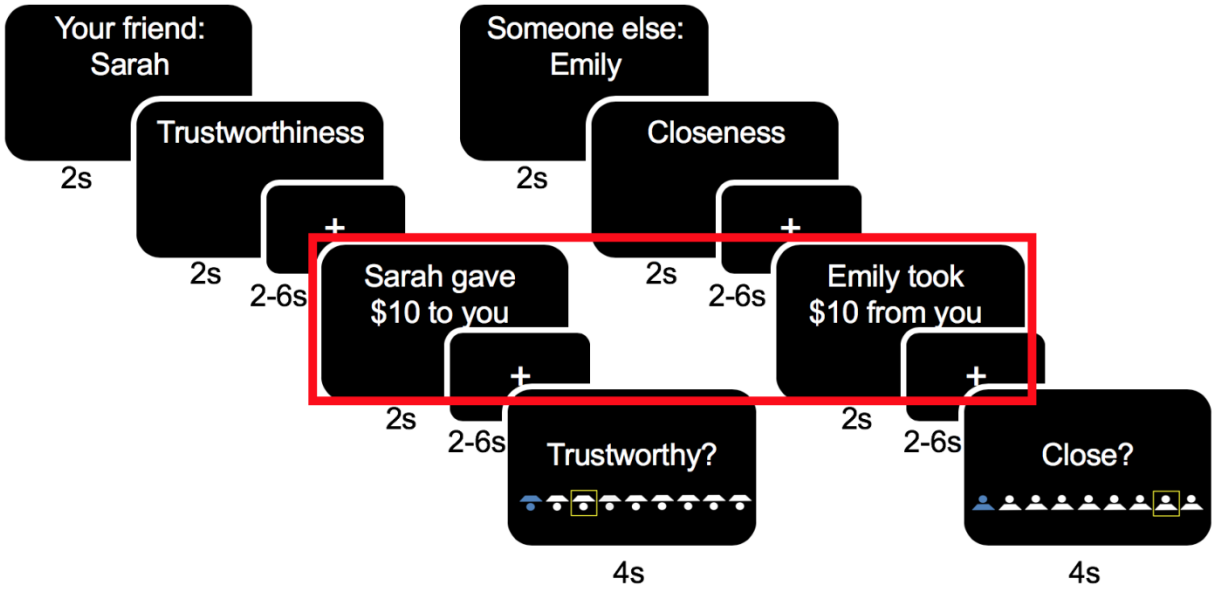

Figure 2. *Representative trials of Social Judgment Task.* Participants viewed the name of Player 1 (2s); the rating that they would make in the trial (2s); a jittered fixation cross (2-6s); the decision of the Player 1 (2s); a jittered fixation cross (2-6s); and made their ratings (4s). Each trial was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

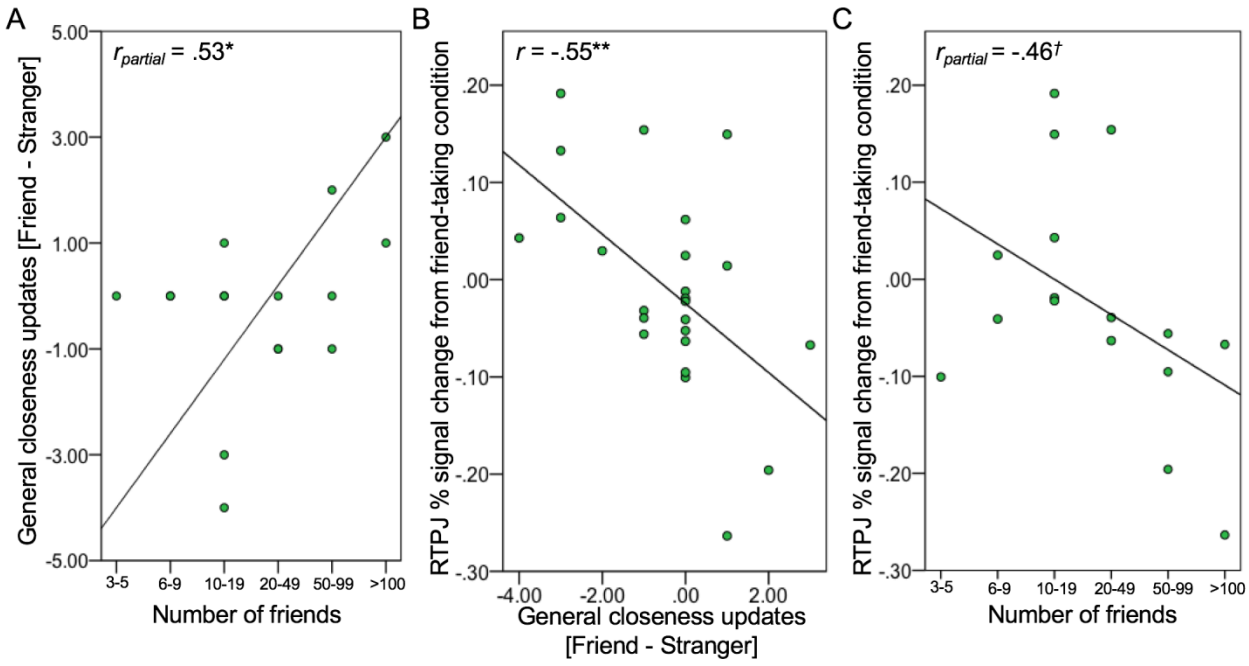

Figure 3. *Associations between general closeness updates, RTPJ activity, and the number of friends in Study 2*. (A) Participants who updated friend-closeness ratings less negatively compared to stranger-closeness ratings reported having a greater number of friends. (B) Participants who showed decreased RTPJ activity when their friend took money from them engaged in less negative updating for friend-closeness ratings compared to stranger-closeness ratings. (C) Participants who showed decreased RTPJ activity when their friend took money from them reported having a greater number of friends. For the visualization purpose, the zero-order correlations are depicted. $r_{partial}$ = Correlation coefficients after controlling for how often participants make new friends. *†p < .10, *p < .05, **p < .01.*

Supplementary Materials for

Does biased impression updating facilitate close relationships?

A behavioral and fMRI investigation

Park, B. and Young, L.

1. Descriptions and findings for exploratory survey items in Study 1
2. Comparisons between evaluations at the last positive story and those at the last negative story in Study 1
3. Correlations between covariates in regression models in Study 1
4. Associations between participants' impression updating and their use of emotion words in memory recall
5. Activation in response to Theory of Mind task, belief versus photo in Study 2
6. Pre-scan versus post-scan evaluations in Study 2
7. The effect of how often participants made new friends on how many friends participants reported having in Study 2
8. Association between participants' RTPJ activity and general trustworthiness updates in Study 2
9. Correlations between covariates in regression models in Study 2

Supplementary Section 1. Descriptions and findings with exploratory survey items in Study 1.

For exploratory purposes, in Study 1, after participants read all stories about their friend and a stranger, they also answered a series of questionnaires intended to capture different aspects of their social interactions. These questionnaires included the positive relations with others subscale (e.g., "I enjoy personal and mutual conversations with family members or friends") from the Psychological Well-being Scale (Ryff, 1989; Ryff & Keyes, 1995), the individual loyalty subscale (e.g., "If I make a promise to a friend, I will keep it") from the Individual and Group Loyalty Scale (Beer & Watson, 2009), Relational-Interdependent Self-Construal Scale (e.g., "When I think of myself, I often think of my close friends or family also") (Cross, Bacon, & Morris, 2000), and the Unidimensional Relationship Closeness Scale with the friend whose name they submitted to the survey (e.g., "My relationship with [friend's name] is close.") (Dibble, Levine, & Park, 2011), along with other items[1].

In exploratory analyses, we found that the more negatively participants updated their trustworthiness ratings for a stranger after reading negative stories about the stranger, the greater relational satisfaction (Pearson's $r = .28$, $p = .004$), greater loyalty ($r = .45$, $p < .001$), greater relational self-construal ($r = .28$, $p = .003$), and closer relationships with their friend ($r = .29$, $p = .003$) they had. These effects remained similar after controlling for updates in friend-closeness, friend-trustworthiness, and stranger-closeness ratings. These findings suggest participants who are more satisfied with their current relationships in real life might be more likely to rate strangers more negatively within the experimental paradigm, perhaps resulting in a favorable comparison with their friends. Also possible is that participants who are more sensitive

[1] Other items included participants' most pleasant and unpleasant memory with their friend, how long they have known their friend, how long is their oldest friendship, loyalty subscale from Moral Foundations Questionnaire (Graham, Haidt, & Nosek, 2008), extraversion and emotional stability subscales from Ten Item Personality Measure (Gosling, Rentfrow, & Swann, 2003), how much participants liked their friend, and how many hours they spend or communicate with their friend. Our findings remained similar after we statistically controlled for these variables. Additionally, we also measured how many followers they have on their social media account(s), but 24.3% - 60.4% of participants did not have the account(s) that we asked about (facebook, instagram, and twitter). Thus, to secure enough statistical power, these variables were not included in the analyses any further.

to others' negative behavior might be more selective in choosing friends, enhancing the quality of their close relationships and increasing commitment to these relationships.

Supplementary Section 2. Comparisons between evaluations at the last positive story and those at the last negative story in Study 1.

To examine how much people updated after considering target agents' (hypothetical) negative behaviors, we ran a 2 (Agent: Friend, Stranger) X 2 (Task: Closeness, Trustworthiness) repeated ANOVA on participants' updates (ratings for the last positive story – ratings for the last negative story). We found a significant main effect of Agent, $F(1,109) = 14.56$, $p < .001$, partial eta-squared = .12, indicating that participants updated their evaluation about a stranger (M = 2.48, S.E. = .13) more than their evaluation about their friend (M = 1.92, S.E. = .16). A significant main effect of Task, $F(1,109) = 253.54$, $p < .001$, partial eta-squared = .70, showed that participants updated less for closeness ratings (M = 1.09, S.E. = .12) than for trustworthiness ratings (M = 3.31, S.E. = .16). However, these effects were modified by a significant Agent X Task interaction, $F(1,109) = 31.55$, $p < .001$, partial eta-squared = .22. While participants updated their closeness ratings for friend and stranger similarly (Friend M = 1.11, S.E. = .15; Stranger M = 1.07, S.E. = .14, $p = .827$, 95% CI = [-.29, .37]), they updated trustworthiness ratings for the stranger (M = 3.88, S.E. = .18) more than for their friend (M = 2.73, S.E. = .19), $p < .001$, 95% CI = [-1.54, -.77] (Figure S1). Thus, participants were particularly less reluctant to update their ratings for stranger's trustworthiness.

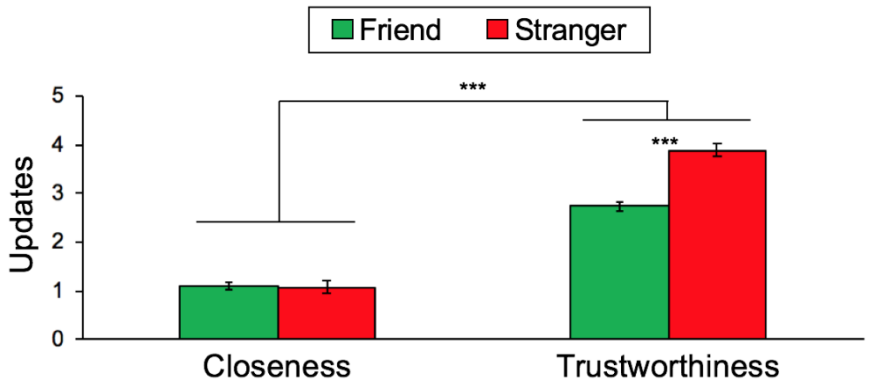

Figure S1. *Updates in closeness and trustworthiness ratings in Study 1.* Participants were more likely to update their trustworthiness ratings than their closeness ratings in general, and were

more likely to update stranger-trustworthiness, after reading two consecutive negative versus positive behaviors. ***$p < .001$.

Supplementary Section 3. Correlations between covariates in regression models in Study 1.

Updates while reading the stories	Updates in friend-closeness	Updates in friend-trustworthiness	Updates in stranger-closeness	Updates in stranger-trustworthiness	How often people make new friends
Updates in friend-closeness	-				
Updates in friend-trustworthiness	-.55***	-			
Updates in stranger-closeness	.19†	-.25*	-		
Updates in stranger-trustworthiness	-.25*	.45***	-.16†	-	
How often people make new friends	-.11	.03	.18†	.13	-

Note. Correlations between variables controlled in regression models (Study 1). Updates in friend-closeness and stranger-closeness are categorized (0 = Never update, -1 = Negative update, +1 = Positive update). Other updating scores are computed by subtracting participants' ratings for the last negative story from their ratings for the last positive story. *†p < .10, *p < .05, ***p < .001*.

Updates before and after the stories	Updates in friend-closeness	Updates in friend-trustworthiness	Updates in stranger-closeness	Updates in stranger-trustworthiness	How often people make new friends
Updates in friend-closeness	-				
Updates in friend-trustworthiness	.53***	-			
Updates in stranger-closeness	.01	.04	-		
Updates in stranger-trustworthiness	.04	.04	.34***	-	
How often people make new friends	.04	-.004	.06	.17†	-

Note. Correlations between variables controlled in regression models (Study 1). Updating scores are computed by subtracting participants' ratings in their post-exposure evaluation (i.e., after they read all stories) from those in their pre-exposure evaluation (i.e., before they read any stories). *†p < .10, *p < .05, ***p < .001*.

Supplementary Section 4. Associations between participants' impression updating and their use of emotion words in memory recall

In Studies 1 and 2, for exploratory purposes, participants were asked to describe their most pleasant and unpleasant moments with their friend in the time they have known each other. We used the Linguistic Inquiry and Word Count, 2007 program (LIWC; Pennebaker, Francis, & Booth, 2001) to quantify word use. Specifically, after removing the clauses that the participants repeated from the question (e.g., "The most pleasant moment with my friend was…" "The most unpleasant moment with my friend was…"), we analyzed the percentage of mutually exclusive positive emotion (e.g., love, nice, sweet; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) and negative emotion (e.g., hurt, ugly, nasty; Pennebaker et al., 2007) word use, to assess participants' quality of interaction with their friend.

Although the degrees to which participants updated their closeness and trustworthiness judgments about their friend and stranger were not associated with the use of emotional words in their memory recall in Study 1, participants' RTPJ activity during the Social Judgment Task was significantly associated with their use of emotional words in Study 2. Specifically, we found that participants who showed greater RTPJ activity in response to their friend's taking behavior during the Social Judgment Task used more negative emotion words when they recalled the most unpleasant memory with their friend, Pearson's r = .59, p = .034, 95% CI = [.07, .90]. These findings suggest that participants who downplayed their friend's taking behavior less, thus those who might be more sensitive to their friend's negative behavior, experienced or recalled greater negative emotion when they recalled a social event. In contrast, participants' RTPJ response to their friend's giving behavior was not associated with their use of positive emotion words when they recalled the most pleasant memory with their friend, Pearson's r = -.15, p = .615, 95% CI = [-.67, .41]. RTPJ activity in response to stranger's giving or taking was not significantly associated with emotional word use, ps > .135.

Supplementary Section 5. Activation in response to Theory of Mind task, belief versus photo in Study 2

Region	x	y	z	Peak Z	Voxels
LTPJ	-46	-64	22	6.95	1434
R Superior temporal gyrus/ RTPJ	57	-58	19	5.82	1199
L Superior frontal gyrus/ L dmPFC	-7	47	37	5.64	918
L Precuneus	-4	-58	28	6.28	572
L Middle frontal gyrus	-37	35	24	-5.13	292
R Middle frontal gyrus	44	35	24	-4.95	111
L Inferior temporal gyrus	-56	-50	-18	-4.84	87
R Medial frontal gyrus	5	52	-2	4.38	67
R Precuneus	29	-65	41	-4.20	51
L Precuneus	-22	-68	35	-4.23	41

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

Supplementary Section 6. Pre-scan versus post-scan evaluations in Study 2.

To examine participants' general impression updates before and after the Social Judgment Task in Study 2, we compared their pre-scan and post-scan impression evaluations of their friend and the stranger. We conducted a 2 (Agent: Friend, Stranger) X 2 (Task: Closeness, Trustworthiness) X 2 (Time: Pre-scan, Post-scan) repeated-measures ANOVA on participants' ratings outside the scanner.

There were significant main effects of Agent (F[1, 23] = 351.29, p < .001, partial eta-squared = .94) and Task (F[1, 23] = 149.37, p < .001, partial eta-squared = .87), indicating that participants gave more positive evaluations to their friend (M = 7.03, S.E. = .14) than to the stranger (M = 2.90, S.E. = .15), and for the trustworthiness ratings (M = 5.69, S.E. = .14) than for the closeness ratings (M = 4.24, S.E. = .08)[2]. These effects were modified by a significant interaction between Agent X Task, F(1, 23) = 93.11, p < .001, partial eta-squared = .80, indicating that participants rated closeness for stranger (M = 1.44, S.E. = .11) lower than trustworthiness for stranger (M = 4.35, S.E. = .25), p < .001, while these ratings did not differ for friend (closeness M = 7.04, S.E. = .16; trustworthiness M = 7.02, S.E. = .15), p = .873.

More importantly, as predicted, these effects were modified by a significant 3-way interaction between Agent X Task X Time, F(1, 23) = 4.53, p = .044, partial eta-squared = .16 (Figure S2). Although participants perceived their friend as less trustworthy in the post-scan evaluation (Pre-scan M = 7.33, S.E. = .14; Post-scan M = 6.71, S.E. = .24), p = .025, 95% CI for difference = [.09, 1.17], partial eta-squared = .20, they did *not* significantly change their closeness ratings for their friend (Pre-scan M = 7.00, S.E. = .15; Post-scan M = 7.08, S.E. = .20), p = .575, 95% CI for difference = [-.39, .22], partial eta-squared = .01. Participants

[2] The Task main effect was qualified by Time, Task X Time F(1, 23) = 19.67, p < .001, partial eta-squared = .46, showing that participants rated closeness higher in the post-scan evaluation (pre-scan M = 4.08, S.E. = .08; post-scan M = 4.40, S.E. = .12), p = .025, while they rated perceived trustworthiness lower in general in the post-scan evaluation (pre-scan M = 6.04, S.E. = .15; post-scan M = 5.33, S.E. = .21), p = .006. This effect seems to be driven by their increased closeness ratings in the post-scan evaluation for the stranger.

perceived the stranger as less trustworthy in the post-scan evaluation (Pre-scan M = 4.75, S.E. = .27; Post-scan M = 3.96, S.E. = .30), p = .007, 95% CI for difference = [.24, 1.35], partial eta-squared = .27. They rated the stranger as marginally closer in the post-scan evaluation (Pre-scan M = 1.17, S.E. = .13; Post-scan M = 1.71, S.E. = .20), p = .050, 95% CI for difference = [-1.08, .001], partial eta-squared = .16[3]. These findings suggest that participants were more protective toward their perception of closeness with their friend than their perception of friends' trustworthiness, and their evaluations for the stranger.

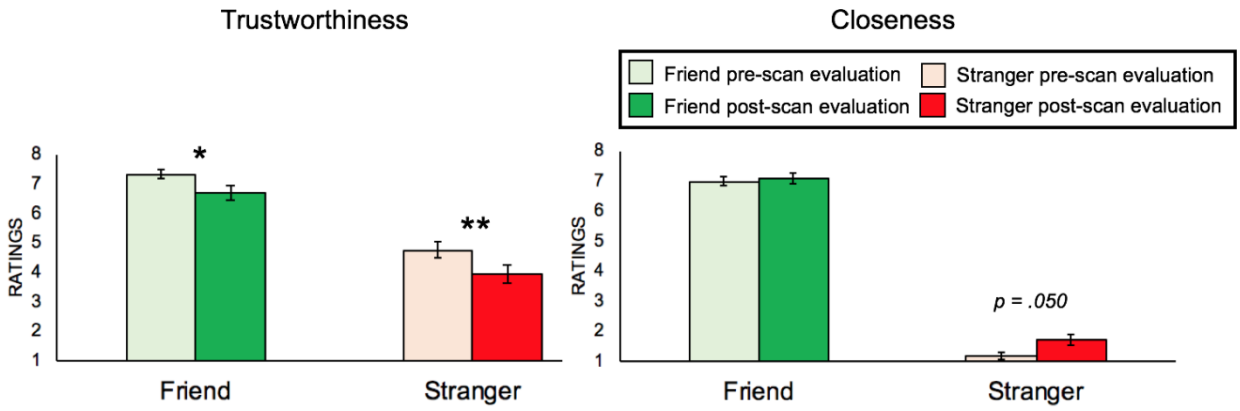

Figure S2. *General impression updates in Study 2:* Participants' trustworthiness ratings were decreased after the Social Judgment Task for both friend and stranger (Left). Participants rated the stranger marginally closer to them after the game. Closeness ratings for the friend remained the same (Right). Error bars represent standard errors (S.E.). ** p < .05, ** p < .01.*

[3] We interpret this pattern in the context of the floor effect of participants' pre-scan closeness ratings for stranger. Participants first made closeness ratings upon barely meeting the confederate (stranger), and 91.7% of the participants rated the stranger a 1 on the closeness scale, which corresponded to "A total stranger." The game, involving ostensible interactions with the stranger, may have contributed to the slight increase in the closeness ratings.

Supplementary Section 7. The effect of how often participants made new friends on how many friends participants reported having in Study 2.

In the ordinal regression on the number of friends participants reported having with general closeness updates, controlling for how often participants made new friends, we found that those who make new friends less often had a smaller number of friends, Estimate = -1.09, S.E. = .43, Wald = 6.50, p = .011, 95% CI for Estimate = [-1.93, -.25]. However, we found similar association between participants' general closeness updates and the number of friends they had without controlling for how often they made new friends.

Additionally, another ordinal regression on the number of friends participants reported having with their RTPJ activity from friend-taking condition, controlling for how often participants made new friends, also revealed that participants who make new friends less often had significantly lower number of friends, Estimate = -.74, S.E. = .36, Wald = 4.19, p = .041, 95% CI for Estimate = [-1.45, -.03]. But the association between participants' RTPJ activity and the number of friends remained similar without controlling for how often they make new friends.

Supplementary Section 8. Association between participants' RTPJ activity and general trustworthiness updates in Study 2.

For exploratory purposes, we ran regression analyses on participants' general trustworthiness updates [friend-stranger] with RTPJ activity averaged within friend versus stranger conditions, for giving versus taking trials separately. We found that general changes in trustworthiness were tracked by RTPJ activity in the friend-giving condition, B = -13.21, S.E. = 3.80, *ß* = -.69, *t* = -3.48, *p* = .003. The lower RTPJ activity participants showed in response to their friend's giving behavior, the more positively participants changed their trustworthiness ratings for their friend than for the stranger. Given that decreased efforts for mentalizing is often associated with more positive evaluations (Hughes, Zaki, & Ambady, 2017; Kliemann, Young, Scholz, & Saxe, 2008; Park, Blevins, Knutson, & Tsai, 2017), this pattern suggests that participants who did not experience the need for mentalizing when their friend gave money might evaluate their friend more positively than the stranger after the game. RTPJ activity from other conditions was not significantly related with changes in trustworthiness ratings, *ps* > .074.

Supplementary Section 9. Correlations between covariates in regression models in Study 2.

	Closeness updates [Friend – Stranger]	Trustworthiness updates [Friend – Stranger]	How often people make new friends
Closeness updates [Friend – Stranger]	-		
Trustworthiness updates [Friend – Stranger]	.53**	-	
How often people make new friends	.16	.50†	-

Note. Correlations between behavioral covariates in Study 2. *†p < .10, **p < .01*

	RTPJ activity [friend-taking]	RTPJ activity [friend-giving]	RTPJ activity [stranger-taking]	RTPJ activity [stranger-giving]	How often people make new friends
RTPJ activity [friend-taking]	-				
RTPJ activity [friend-giving]	.46*	-			
RTPJ activity [stranger-taking]	.36†	.42*	-		
RTPJ activity	.26	.06	-.35†	-	

[stranger-giving]					
How often people make new friends	.08	-.42	.03	-.12	-

Note. Correlations between neural covariates and how often people make new friends in Study 2. *†p < .10, *p < .05*

References

Beer, A. & Watson, D. (2009). The individual and group loyalty scales (IGLS): Construction and preliminary validation. *Journal of Personality Assessment, 91*(3), 277-287, DOI: 10.1080/00223890902794341

Cross, S. E., Bacon, P. L., & Morris, M. L. (2000). The relational-interdependent self-construal and relationships. *Journal of Personality and Social Psychology, 78*(4), 791-808.

Dibble, J. L., Levine, T. R., & Park, H. S. (2011). The unidimensional relationship closeness scale (URCS): Reliability and validity evidence for a new measure of relationship closeness. *Psychological Assessment, 24*(3), 565-572.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

Graham, J., Haidt, J., & Nosek, B. A. (2008). *Moral Foundations Questionnaire.* Retrieved from http://www.moralfoundations.org/questionnaires

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.

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

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

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001* (p. 71). Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Chung, C., K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007.* www.LIWC.Net.

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57*, 1069–1081.

Ryff, C. D. & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69*, 719–727.