Behavioral/Systems/Cognitive

Positively Biased Processing of Self-Relevant Social Feedback

Christoph W. Korn,[1,2,3] **Kristin Prehn,**[3,4] **Soyoung Q. Park,**[1,2,3] **Henrik Walter,**[2,5] **and Hauke R. Heekeren**[1,2,3,4]

[1]Department of Education and Psychology, Freie Universität Berlin, 14195 Berlin, Germany, [2]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10117 Berlin, Germany, [3]Dahlem Institute for Neuroimaging of Emotion, Freie Universität Berlin, 14195 Berlin, Germany, [4]Cluster of Excellence "Languages of Emotion," Freie Universität Berlin, 14195 Berlin, Germany, and [5]Department of Psychiatry, Division of Mind and Brain Research, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany

Receiving social feedback such as praise or blame for one's character traits is a key component of everyday human interactions. It has been proposed that humans are positively biased when integrating social feedback into their self-concept. However, a mechanistic description of how humans process self-relevant feedback is lacking. Here, participants received feedback from peers after a real-life interaction. Participants processed feedback in a positively biased way, i.e., they changed their self-evaluations more toward desirable than toward undesirable feedback. Using functional magnetic resonance imaging we investigated two feedback components. First, the reward-related component correlated with activity in ventral striatum and in anterior cingulate cortex/medial prefrontal cortex (ACC/MPFC). Second, the comparison-related component correlated with activity in the mentalizing network, including the MPFC, the temporoparietal junction, the superior temporal sulcus, the temporal pole, and the inferior frontal gyrus. This comparison-related activity within the mentalizing system has a parsimonious interpretation, i.e., activity correlated with the differences between participants' own evaluation and feedback. Importantly, activity within the MPFC that integrated reward-related and comparison-related components predicted the self-related positive updating bias across participants offering a mechanistic account of positively biased feedback processing. Thus, theories on both reward and mentalizing are important for a better understanding of how social information is integrated into the human self-concept.

Introduction

Humans are often confronted with social feedback about their character when interacting with other people and have to integrate this feedback into their self-concept. For example, if somebody tells you that you are polite you weigh this feedback and integrate it into how polite you see yourself. Importantly, people tend to see themselves in a positive light (Leary, 2007) and expect to receive more positive than negative feedback (Hepper et al., 2011). It has been proposed that humans can achieve and maintain a positive self-concept because cognitive processing mechanisms distort incoming information in a positive direction (Taylor and Brown, 1988). Studying positively biased self-views bears far-reaching implications for psychiatry, health psychology, and policy making, since positivity biases have often been linked to mental health, personal well being, and professional success (Leary, 2007). The goal of the present study was to determine the information processing mechanisms at play when people receive feedback relevant for their self-concept.

Received June 21, 2012; revised Aug. 30, 2012; accepted Sept. 23, 2012.

Author contributions: C.W.K., K.P., H.W., and H.R.H. designed research; C.W.K. performed research; C.W.K., S.Q.P., and H.R.H. analyzed data; C.W.K., K.P., S.Q.P., H.W., and H.R.H. wrote the paper.

This work was supported by the Excellence Initiative of the German Federal Ministry of Education and Research, Deutsche Forschungsgemeinschaft, Grants GSC86/1-2009 and EXC 302.

The authors declare no competing financial interests.

Correspondence should be addressed to Christoph W. Korn, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. E-mail: christoph.w.korn@gmail.com.

DOI:10.1523/JNEUROSCI.3016-12.2012

 0270-6474/12/3216832-13$15.00/0

The idea that processing mechanisms distort incoming information in a positive direction suggests that reward should play a central role in social feedback processing. Neuroscientific studies have shown that nonsocial rewards (e.g., money) and social rewards (e.g., positive feedback on character traits) are processed within shared brain regions, notably the ventral striatum and a region at the border of the pregenual anterior cingulate cortex (ACC), the ventral medial prefrontal cortex (MPFC), and the medial orbitofrontal cortex (OFC; Fehr and Camerer, 2007; Fliessbach et al., 2007; Izuma et al., 2008; Beckmann et al., 2009; Rushworth et al., 2011). However, neural activity related to social reward has not been linked to positively biased self-views.

When receiving social feedback about character traits, people compare their own view to the view of others. Self-referential processing, such as judging one's own personality traits, has been linked to the frontal midline, especially ventral MPFC (Amodio and Frith, 2006; Moran et al., 2006; Northoff et al., 2006; Lieberman, 2007, Wagner et al., 2012). Moreover, inferring the mental states of other persons—known as mentalizing or theory-of-mind—has been reliably associated with a network comprising dorsal MPFC, temporoparietal junction (TPJ), superior temporal sulcus (STS), temporal poles (TPs), and orbital inferior frontal gyrus (IFG) (Amodio and Frith, 2006; Gilbert et al., 2006; Saxe, 2006; Van Overwalle, 2009; Bahnemann et al., 2010; Mar, 2011). Activity within the mentalizing network has been observed across a variety of tasks, such as reading stories about false beliefs (Saxe and Powell, 2006), viewing diagrams or videos of social interactions (Walter et al., 2004; Wolf et al., 2010), and engaging in strategic interactions (Behrens et al., 2008; Hampton et al., 2008;

Yoshida et al., 2010). Social feedback processing arguably involves inferring other persons' mental states to integrate their views into one's self-concept. However, it has not been tested whether regions associated with mentalizing are implicated in social feedback processing.

Here, we mainly investigated how humans process feedback about their own character traits but we were additionally interested in comparing self-related versus other-related feedback. We hypothesized that humans process social feedback in a positively biased way and expected feedback processing to include two components. First, we expected a reward-related component to be linked to activity in the ventral striatum and ACC/MPFC. Second, we hypothesized that the comparison between participants' own views and the feedback ratings they received would be reflected in regions previously associated with mentalizing. We expected activity in the MPFC in particular, since distinctive subregions of the MPFC have been linked to processes that we expected to be relevant for social feedback processing. First, a region at the border of pregenual ACC, ventral MPFC, and OFC shows involvement in reward and value processing (Beckmann et al., 2009; Rushworth et al., 2011). Second, meta-analytic evidence suggests a spatial gradient within the MPFC—with more ventral subregions involved in self-referential processing and more dorsal subregions involved in other-referential processing including mentalizing (Denny et al., 2012).

Materials and Methods

Participants

In total, 30 right-handed subjects participated. Three participants had to be excluded (one did not tolerate the scanner environment, another showed excessive head movement (>8 mm), and data from another subject could not be used due to technical problems) leaving 27 subjects for analyses (14 female, mean age = 24.3 years, SD = 2.46). All subjects gave written informed consent.

Experiment

The experimental procedure is outlined in Figure 1. We wanted participants to believe that they would get realistic feedback on their personality traits from peers with whom they had interacted in real life. We tested how much this feedback changed participants' self-concept by asking them to rate their own personality before and after receiving social feedback. Additionally, each participant rated one other person before and after receiving social feedback for this person. Participants came into the laboratory on 2 consecutive days. The purpose of the first day was to create a real-life interaction among peers so that the social feedback would be meaningful for participants. The purpose of the second day was to assess participants' self-concept change after receiving social feedback.

Day 1—social interaction and rating of three players. On the first day (Fig. 1*A*), participants came into the laboratory in groups of five people of the same sex and got to know each other by playing a table-top version of the popular board game Monopoly (Hasbro) for 1 h and 15 min. We made sure that participants did not know each other before the experiment. We chose the board game Monopoly for the social interaction because it is highly engaging, quite well known, and allows players to show a variety of cooperative and competitive behaviors. Furthermore, within 1 h 15 min nobody was eliminated from the game. The rules of the game were explained to all participants. The study was introduced as a study about the neural correlates about how people get to know each other. Participants knew before they started to play the game that they were going to be rated by the other players of their group and they believed that their own ratings were going to be shown to the other players in an anonymous fashion. During the game participants were free to talk about whatever topics they wanted. Participants wore name tags and we made sure that participants knew the names of all players after the game. After 1 h 15 min we assessed the ranking of the participants in the game, i.e., assigned the first rank to the winner and so on. After the game, each participant rated three of the four other participants on 80 trait adjectives (Table 1; see below, Stimuli) on a Likert scale from 1 (this trait does not apply the person at all) to 8 (this trait does apply the person very much) on a PC using the MATLAB toolbox Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php). Each of the three persons was rated in a separate block. On each trial participants saw one of the 80 adjectives with the first name of the person to rate and had up to 10 s to respond. At the end of day 1 each participant had rated three other participants and in turn each participant had been rated by three other participants. Participants had not yet rated themselves (Fig. 1*A*, yellow) and had not yet rated one other player (depicted in green).

Day 2—functional magnetic resonance imaging task and post-functional magnetic resonance imaging ratings. On the second day (Fig. 1*B*), participants performed the following functional magnetic resonance imaging (fMRI) experiment, which was presented using the MATLAB toolbox Cogent 2000. On each trial, participants first saw a cue (1 s) indicating

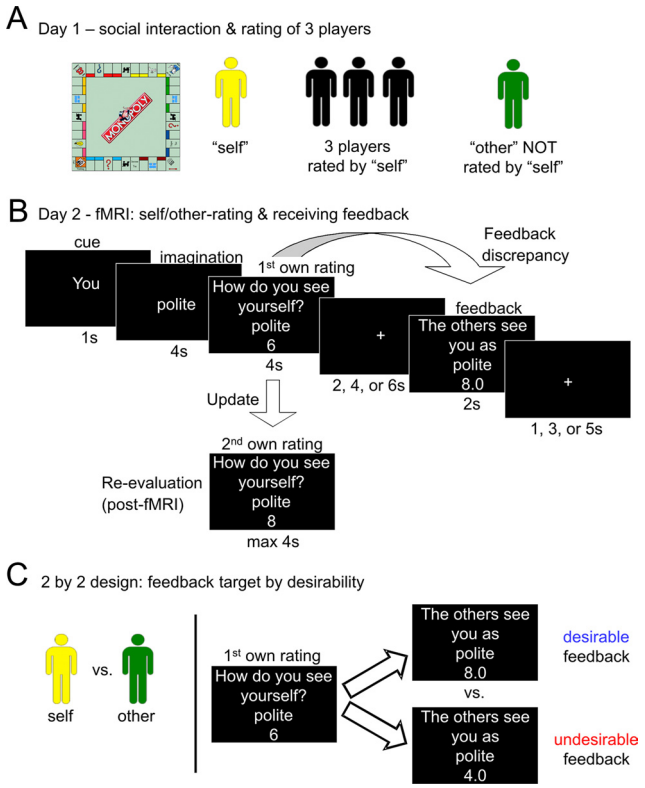

Figure 1. Task design, receiving social feedback from peers after a real-life interaction. ***A***, Participants came to the laboratory in groups of five on 2 consecutive days. On the first day they got to know each other by playing the board game Monopoly for 1 h 15 min. Afterward, each person rated three of the other players on 40 positive and 40 negative trait adjectives on a Likert scale from 1 (this trait does not apply to the person at all) to 8 (this trait applies to the person very much). On the first day participants did not rate themselves (yellow) and did not rate one of the other players (green). See Table 1 for a list of the trait adjectives. ***B***, On the second day participants performed the following task in the fMRI scanner. They first saw a cue indicating whether the following trial was about themselves or about the other person they had not rated on the previous day. They then saw one of 40 positive or 40 negative trait adjectives and had to imagine how much the trait applied to themselves or to the other person. They first gave their own rating and then saw the feedback in the form of the mean rating they believed three other participants had given on the previous day. The absolute difference between participants' own ratings and the feedback ratings they received was conceptualized as feedback discrepancies and manipulated during the experiment. Outside the scanner participants rated themselves and the other player a second time so that we could assess how much they updated their ratings after receiving feedback. ***C***, For the main behavioral analyses we used a 2 by 2 design with the factors feedback target (self/other) and feedback desirability (desirable/undesirable). Desirable feedback was defined as feedback ratings that were higher than participants' own first ratings (e.g., own first rating for polite was 6 and feedback rating was 8.0). Conversely, undesirable feedback was defined as feedback ratings lower than participants' first ratings (e.g., own first rating for polite was 6 and feedback rating was 4.0). All ratings for negative trait adjectives were reverse coded. Thus, feedback desirability was independent of the valence of the trait adjective.

Table 1. List of trait adjectives

German original	English translation
Positive trait adjectives	
Aufrichtig	Honest
Bescheiden	Modest
Diszipliniert	Organized
Effizient	Efficient
Einfühlsam	Empathetic
Enthusiastisch	Enthusiastic
Fleißig	Hard-working
Freundlich	Friendly
Geistesgegenwärtig	Quick-witted
Gelassen	Composed
Geschickt	Skilled
Gesellig	Sociable
Großzügig	Generous
Hilfsbereit	Helpful
Höflich	Polite
Kompetent	Competent
Kooperativ	Cooperative
Kreativ	Creative
Lebenslustig	Fun-loving
Locker	Easy-going
Loyal	Loyal
Offen	Open-minded
Ordentlich	Tidy
Respektvoll	Respectful
Scharfsinnig	Astute
Schlagfertig	Articulate
Selbstständig	Self-reliant
Sorgfältig	Diligent
Souverän	Confident
Spontan	Spontaneous
Tatkräftig	Dynamic
Tolerant	Tolerant
Vernünftig	Level-headed
Verständnisvoll	Understanding
Vertrauenswürdig	Trustworthy
Vielseitig	Versatile
Weitsichtig	Perspicacious
Wissbegierig	Inquisitive
Zielstrebig	Goal-oriented
Zuverlässig	Reliable
Negative trait adjectives	
Aggressive	Aggressive
Ängstlich	Anxious
Arrogant	Arrogant
Bieder	Overly conservative
Chaotisch	Chaotic
Egoistisch	Selfish
Eitel	Conceited
Engstirnig	Narrow-minded
Feige	Cowardly
Gehässig	Spiteful
Großmäulig	Loud-mouthed
Heuchlerisch	Two-faced
Hinterhältig	Conniving
Humorlos	Humorless
Inkonsequent	Inconsistent
Kalt	Cold-hearted
Launisch	Moody
Leichtsinnig	Foolhardy
Nachtragend	Unforgiving
Naive	Naive
Oberflächlich	Superficial
Opportunistisch	Opportunistic
Pedantisch	Pedantic
Rücksichtslos	Inconsiderate
Scheu	Unassertive
Stur	Stubborn
Träge	Lazy
Unentschlossen	Indecisive
Ungeduldig	Impatient
Unnahbar	Inapproachable
Unpünktlich	Tardy
Unsicher	Insecure
Unsympathisch	Unpleasant
Verschwenderisch	Wasteful
Voreilig	Rash
Voreingenommen	Biased
Wehleidig	Whiny
Zickig	Catty
Zwanghaft	Obsessive
Zynisch	Cynical
Adjectives used during the training session	
Intelligent	Intelligent
Unsportlich	Unathletic

(*Table continues.*)

whether the trial was about themselves (self-condition) or about the fourth participant (other-condition) they had not rated on the first day. Then, they saw 1 of 80 trait adjectives and had to think about how much that trait applied to themselves or to the other person (imagination phase, 4 s). When the words "How much does this trait apply to you/to this person?" appeared participants had to indicate their rating on an 8 point Likert scale via two button boxes with four buttons each (rating phase, 6 s). After a jittered fixation cross (2, 4, or 6 s) participants saw what they believed to be the mean rating of three other participants from the previous day (feedback phase, 2 s). This mean rating, which served as the feedback rating, was a number with one decimal, ranging from 1.0 to 8.0 in steps of 0.3. The feedback rating was determined by the program during the experiment to reliably create a sufficient number of trials in which participants received desirable and undesirable feedback (see below, Task conditions and behavioral analyses for a detailed description). After a second jittered fixation cross (1, 3, or 5 s) a new trial began. Participants performed four training trials before scanning. The experiment was split up into four blocks with the same 10 positive and the same 10 negative trait adjectives for self and other trials within one block. Trials for self and other were randomly intermixed. Adjectives were randomly assigned to the four blocks for each person.

Immediately after the scanning session participants performed a second rating outside the fMRI scanner on a PC to measure how much participants changed their self-ratings and other-ratings after having received social feedback in the scanner. Specifically, they rated themselves and the other person again on all 80 trait adjectives in two separate blocks (one for themselves and one for the other person). These blocks were randomized for order. For each trait adjective participants had up to 6 s to respond.

Day 2—additional behavioral tasks: memory and individual difference scores. After rating themselves and the other person a second time, participants were assessed for their memory of the feedback they had received in the scanner. For all 80 trait adjectives participants had to recollect the feedback they had seen in the scanning sessions and had to type in that number, i.e., a number between 1 and 8 with one decimal such as 1.0, 1.3, or 1.7. Participants had to recollect the feedback in two separate blocks (one for themselves and one for the other person), which were randomized for order. They had up to 12 s to respond.

Participants rated how similar they thought the other person was to them on a Likert scale from 1 (not similar at all) to 8 (very similar) and completed the Rosenberg self-esteem scale (Rosenberg, 1965).

Stimuli

Adjectives were selected on the basis of a comprehensive list of trait adjectives (Anderson, 1968), which had been previously used to create stimuli for social neuroscience experiments (Fossati et al., 2003; Izuma et al., 2008), and on the basis of the Berlin Affective Word List (Võ et al., 2006). We selected 40 positive adjectives describing socially desirable traits and 40 negative adjectives describing socially undesirable traits. To assess whether participants really perceived the trait words as positive and negative in the way we had predefined them, participants rated all 80 trait adjectives on social positivity on a scale from 1 (not positive at all) to 8 (very positive) at the very end of the experiment. Mean ratings for positive and negative trait words differed significantly from each other and from the midpoint of the scale (mean rating: positive words = 6.6, SD = 0.49; negative words = 2.4, SD = 0.44; paired t test comparing ratings for positive with those for negative words $t_{(26)} = 29.3$, $p < 0.001$; one-sample t tests comparing ratings to the mid-point of the scale: for positive words $t_{(26)} = 22.7$, $p < 0.001$; for negative words $t_{(26)} = -25.7$, $p < 0.001$).

We used adjectives describing different trait concepts and avoided synonyms or antonyms. Word frequency per million words ranged from 0.09 ("touchy") to 61.32 ("open-minded") with a median frequency of 1.23 ("respectful") as assessed by the lexical database DLEX (Heister et al., 2011; www.dlexdb.de/). See Table 1 for a list of trait adjectives.

Task conditions and behavioral analyses

Task conditions. The main behavioral analyses used a 2 by 2 design with the within-subject factors feedback target (self/other) and feedback desirability (desirable/undesirable; Fig. 1*C*).

First, feedback was either targeted to the self, i.e., participants rated themselves before and after receiving feedback for themselves, or targeted to one other person, i.e., participants rated one of the other persons he or she had met on the first day before and after receiving feedback for that person.

Second, for each participant trials were classified according to whether feedback was desirable or undesirable. Desirable feedback was defined as feedback ratings that were more "positive" than participants' own initial ratings. For a positive trait adjective, desirable feedback indicated that the feedback rating was numerically higher than the initial rating (e.g., a participant's initial rating for polite was 6 and the feedback rating was 8). For a negative trait adjective, desirable feedback indicated that the original feedback rating was numerically lower than the original initial rating (e.g., a participant's initial rating for "aggressive" was 3 and the feedback rating was 1). Conversely, undesirable feedback was defined as feedback ratings that were more "negative" than participants' own initial ratings. For a positive trait adjective, undesirable feedback indicated that the feedback rating was numerically lower than the initial rating (e.g., a participant's initial rating for polite was 6 and the feedback rating was 4). For a negative trait adjective, undesirable feedback indicated that the original feedback rating was numerically *higher* than the original initial rating (e.g., a participant's initial rating for "aggressive" was 3 and the feedback rating was 5).

Reverse coding. Importantly, by the above definition feedback desirability was independent of the valence of the trait word. For all analyses we reverse-coded ratings for negative trait adjectives. Specifically, all ratings were on an 8 point Likert scale ranging from 1 (this trait does not apply the person at all) to 8 (this trait does apply the person very much). Ratings for negative traits were subtracted from 9. For example, if the original rating for a negative trait adjective (e.g., unpleasant) was 1 this number was transformed into 8 for the analyses, i.e., into the rating of the corresponding positive trait adjective (e.g., pleasant).

Feedback discrepancy. For each trial (i.e., for each trait adjective; separately for self-conditions and other-conditions) we calculated a "feedback discrepancy" term as the absolute difference between first own ratings and feedback ratings:

$$\text{feedback discrepancy} = \text{abs}(\text{feedback rating} - \text{first own rating}).$$

This feedback discrepancy term indicated the social comparison component of receiving social feedback (i.e., the comparison between own ratings and feedback ratings depended on the absolute magnitude of their difference). Since feedback discrepancies were an independent variable of our task we manipulated their magnitude using a random number generator.

Random number generator for feedback discrepancy. Feedback discrepancies were determined by a random number generator during the fMRI task to reliably create a similar range of feedback discrepancies across participants and to create a sufficient number of trials with desirable and undesirable feedback. Specifically, on each trial the number of previous trials of the same target condition (self or other) within the same scanning session was determined. These previous trials were classified as either desirable or undesirable according to the definition given above (see above, Task conditions). If the numbers of previous trials with desirable and undesirable feedback differed by more than two trials, the feedback type that had been used less was chosen (e.g., if there had been seven trials with desirable feedback and four trials with undesirable feedback the feedback of the current trial would by undesirable). Otherwise feedback desirability was chosen randomly.

Once feedback desirability was determined, a random number generator was used to create a feedback discrepancy so that the feedback rating lay between the first own rating on the endpoints of the scale. (For example, a participant had rated herself 6 on polite and the feedback should be desirable. In that case the feedback rating had to lie between 6.0 and 8.0. The random number generator determined a feedback rating within that range, i.e., a number between 6 and 8 with one decimal, in steps of 0.3).

To ensure believability of the feedback rating, feedback discrepancies could be zero. These trials were excluded from behavioral analyses (see below, Behavioral analyses—ANOVA).

Updates. To assess how much participants changed their self-concept after receiving social feedback, we calculated an update term quantifying how much participants changed their own ratings:

$$\text{update} = \text{second own rating} - \text{first own rating}.$$

We expected participants to change their ratings on average toward the feedback ratings. That is, for desirable feedback (i.e., feedback ratings higher than own first rating) participants should increase their ratings (i.e., updates should be positive). For undesirable feedback (i.e., feedback ratings lower than own first rating) participants should decrease their ratings (i.e., updates should be negative).

However, the critical test for positively biased updating is that the change toward desirable feedback (i.e., the increase) is larger than the change toward undesirable feedback (i.e., the decrease). Therefore, trials were split into trials with desirable feedback and trials with undesirable feedback for each participant and both target conditions (self-desirable, self-undesirable, other-desirable, other-undesirable). We first calculated the mean of all signed updates for each participant within each condition and then calculated absolute mean updates. We then scaled absolute mean updates across conditions and participants by the respective mean feedback discrepancies. That is, we obtained relative absolute mean updates for each participant and condition by dividing absolute mean updates by the respective mean feedback discrepancies:

$$\text{relative absolute mean update} = \text{absolute mean update}/\text{mean feedback discrepancy}.$$

Relative updates can be interpreted in a straightforward way; e.g., a relative update of 0.3 indicates that the change in ratings was on average 30% of the difference between initial own ratings and feedback ratings.

Behavioral analyses—ANOVA. For our main behavioral analysis, we performed a 2 (target: self/other) by 2 (desirability: desirable/undesirable) repeated-measures ANOVA on relative absolute mean updates. Trials with adjectives for which participants failed to respond in time for the first or second rating were excluded from all analyses (self: mean = 1.7 trials, SD = 1.9; other: mean = 2.2 trials, SD = 2.0). Furthermore, trials with a feedback discrepancy of zero were excluded from behavioral analyses since these trials could not be clearly assigned to either receiving desirable or receiving undesirable feedback (self: mean = 5.5 trials, SD = 2.3; other: mean = 6.4 trials, SD = 2.5).

Absolute memory errors. To assess how well participants remembered the feedback presented we asked them to recall all feedback ratings in a separate session. Memory errors were calculated as the absolute differences between the recollected number and the actual feedback rating:

absolute memory error = abs(feedback rating – recollection of feedback rating).

Mean absolute memory errors were compared in a 2 (target: self/other) by 2 (desirability: desirable/undesirable) repeated-measures ANOVA.

fMRI data acquisition

MRI data were acquired on a 3 T scanner (Trio; Siemens) using a 12-channel head coil. Functional images were acquired with a gradient echo T2*-weighted echo-planar sequence (TR = 2000 ms, TE = 30 ms, flip angle = 70, 64 × 64 matrix, field of view = 192 mm, voxel size = 3 × 3 × 3 mm^3). A total of 37 axial slices (3 mm thick, no gap) were sampled for whole-brain coverage. Imaging data were acquired in four separate 349 volume runs of 11 min 38 s each. The first five volumes of each run were discarded to allow for T1 equilibration. A high-resolution T1-weighted anatomical scan of the whole brain was acquired (256 × 256 matrix, voxel size = 1 × 1 × 1 mm^3).

fMRI data analysis

Preprocessing. Image analysis was performed using SPM8 (www.fil.ion.ucl.ac.uk/spm). Echoplanar imaging images were realigned, unwarped, and coregistered to the respective participant's T1 scan; normalized to a standard T1 template based on the Montreal Neurological Institute (MNI) reference brain; resampled to 3 mm isotropic voxels; and spatially smoothed with an isotropic 8 mm full-width at half-maximum Gaussian kernel.

Modeling of fMRI data—overview. fMRI time series were regressed onto a general linear model (GLM) containing regressors representing the time periods of the task (Fig. 1*B*): cue (1 s), imagination phase separately for self and other (4 s), rating phase (4 s), feedback phase separately for self and other (2 s), and two motor regressors for button presses with the left and the right hands (0 s). This resulted in eight regressors per session. The imagination phase regressors for self and other were parametrically modulated by the respective first own ratings. The feedback phase regressors for self and other were modulated by the respective feedback ratings and the respective feedback discrepancies (see below, Modeling of fMRI data—parametric modulators for more details). This model included trials with feedback discrepancies of zero. The six motion-correction parameters estimated from the realignment procedure were entered as covariates of no interest. All regressors and modulators were entered independently into the design matrix, i.e., without the serial orthogonalization used as default in SPM (for a similar approach see Gläscher et al., 2010; Wunderlich et al., 2011). This ensured that only the additional variance that could not be explained by any other regressor was assigned to the respective effect and thus prevented spurious confounds between regressors. Regressors were convolved with the canonical hemodynamic response functions and low-frequency drifts were excluded using a highpass filter with a 128 s cutoff.

Modeling of fMRI data—parametric modulators. For the behavioral analyses we split trials into four categories according to feedback target (self/other) and feedback desirability (desirable/undesirable). In the functional analyses we wanted to investigate trial-by-trial fluctuations in brain activity during the feedback phase, which correlated with two different components of social feedback: reward-related and comparison-related components. In our main functional model we therefore split trials according to feedback target (self/other) for each participant and used parametric modulators of feedback ratings and feedback discrepancies to detect activity related to social reward and social comparison, respectively. Thus, we used the full parametric range of feedback ratings and feedback discrepancies across all trials (i.e., across trials with desirable and undesirable feedback).

First, the activity related to the rewarding component of social feedback should correlate positively with the feedback ratings for self. Note that feedback ratings for negative traits were reverse coded. That is, a high feedback rating indicated high self-relevant social reward (i.e., feedback that a positive trait applied to the self or that a negative trait did not apply to the self) and a low feedback rating indicated low self-relevant social reward (i.e., feedback that a positive trait did not apply to the self or that a negative trait did apply to the self). To make sure that activity related to the rewarding component of social feedback was truly self-specific, we subtracted activity that correlated with the feedback ratings for other.

Second, the activity related to the social comparison component of social feedback should correlate positively with feedback discrepancies defined as the absolute differences between first own ratings and feedback ratings. We defined feedback discrepancies as absolute differences; i.e., feedback discrepancies captured how close feedback ratings were to participants' own ratings, regardless of the direction of the differences.

Follow-up analyses. To visualize the correlations between neural activity and the parametric modulators (i.e., the βs of the parametric modulators for feedback ratings and the βs of the parametric modulators for feedback discrepancies) we performed follow-up functional region of interest (ROI) analyses. We extracted parameter estimates of the parametric modulators for self and other within the functional ROIs identified in the contrasts using the MarsBaR toolbox for SPM (marsbar.sourceforge.net/).

Additionally, to analyze activity for desirable and undesirable trials separately in follow-up analyses, we estimated a second GLM to analyze onset activity within functional ROIs defined by the main model described above (see above, Modeling of fMRI data—overview and Modeling of fMRI data—parametric modulators). Specifically, we estimated a GLM in which regressors for the feedback time period were split up into four conditions in the same fashion as for the main behavioral analysis (self-desirable, self-undesirable, other-desirable, other-undesirable). This follow-up GLM included no parametric modulators.

Conjunction and statistical inference. We tested the conjunction null hypothesis using the minimum *T* statistic as implemented within SPM8 (Nichols et al., 2005).

All reported activations survived a threshold of $p < 0.05$ after cluster-wise familywise error correction for multiple comparisons over the entire brain at a cluster-defining threshold of $p < 0.0001$, uncorrected.

All coordinates are reported in MNI space. Activations are displayed on the standard MNI reference brain. Brodmann areas were manually labeled using the SPM toolbox WFU PickAtlas (fmri.wfubmc.edu/software/PickAtlas).

Results

Behavioral results—positively biased updating

Participants rated how much 40 positive and 40 negative trait adjectives applied to themselves and to one other person before and after receiving feedback ratings. Importantly, all ratings for negative trait adjectives were reverse-coded so that higher ratings always indicated more positive ratings.

In an initial analysis, we performed a 2 by 2 ANOVA comparing ratings before versus after receiving feedback and ratings targeted to the self versus the other person. Participants rated themselves on average more positively than the other person (main effect: self/other; $F_{(1,26)} = 6.7$, $p < 0.05$, $\eta_p^2 = 0.21$; Fig. 2*A*), indicating a positivity bias toward the self. They also rated themselves and the other person more positively after receiving feedback (main effect: before/after; $F_{(1,26)} = 9.6$, $p < 0.005$, $\eta_p^2 = 0.27$). The interaction was not significant ($p > 0.6$).

In our main behavioral analyses, we tested whether participants showed positively biased processing of social feedback. Specifically, we assessed how participants updated their ratings depending on feedback target (self/other) and feedback desirability (desirable/undesirable; Fig. 1*C*). Desirable feedback was defined as feedback ratings that were higher than participants' first ratings. Conversely, undesirable feedback was defined as feedback ratings lower than participants' first ratings. Participants changed their ratings on average toward the feedback. They increased their ratings for desirable feedback (indicated by positive mean updates significantly different from zero) and decreased their ratings for undesirable feedback (indicated by negative mean updates significantly different from zero; mean update self-desirable = 0.5, SD = 0.35; one-sample *t* tests against zero $t_{(26)}$ =

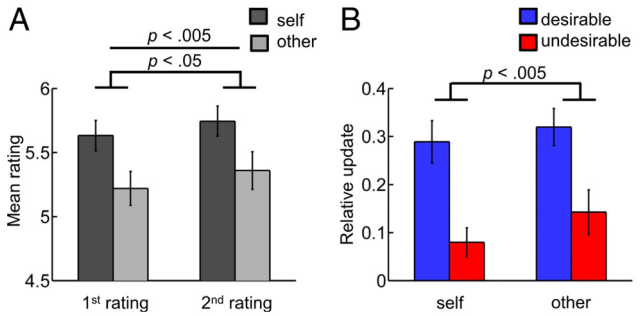

Figure 2. Positively biased updating. ***A***, Mean first and second ratings for self were significantly higher than for other. Second ratings were significantly higher than first ratings. ***B***, Participants changed their ratings more after receiving desirable than after receiving undesirable feedback both for self-related and other-related feedback. Trials were split into four conditions (self-desirable, self-undesirable, other-desirable, other-undesirable). For each condition we calculated the mean update (i.e., the mean difference between second and first ratings). Mean updates were positive for desirable feedback (indicating an increase in ratings) and negative for undesirable feedback (indicating a decrease in ratings). To test whether participants updated their ratings more toward desirable than toward undesirable feedback we calculated absolute mean updates (i.e., we compared the magnitude of the increase for desirable feedback with the magnitude of the decrease for undesirable feedback). Additionally, we scaled absolute mean updates by the respective mean feedback discrepancies for each condition and participant. The resulting relative updates indicate by how much participants changed their ratings with respect to the difference between initial own ratings and feedback ratings. Error bars indicate SEM.

7.4, $p < 0.001$; mean update self-undesirable = −0.2, SD = 0.32; $t_{(26)} = -2.7$, $p < 0.05$; mean update other-desirable = 0.6, SD = 0.33; $t_{(26)} = 8.8$, $p < 0.001$; mean update other-undesirable = −0.3, SD = 0.42; $t_{(26)} = -3.0$, $p < 0.01$).

Importantly, the critical test for positively biased updating is that the changes toward desirable feedback are larger than the changes toward undesirable feedback (i.e., that absolute mean updates are larger for desirable than for undesirable feedback). Additionally, we scaled absolute mean updates by the respective mean feedback discrepancies (i.e., the differences between first own ratings and feedback ratings) to account for possible differences in feedback discrepancies across conditions and participants (relative absolute mean updates: self-desirable = 0.3, SD = 0.23; self-undesirable = 0.1, SD = 0.16; other-desirable = 0.3, SD = 0.20; other-undesirable = 0.1, SD = 0.24). Performing a 2 by 2 ANOVA on relative absolute mean updates comparing self-directed versus other-directed feedback and desirable versus undesirable feedback, we found that participants showed positively biased processing of social feedback. After receiving desirable feedback participants updated their self-ratings and other-ratings more toward the positive than they updated their ratings toward the negative after receiving undesirable feedback (main effect: desirable/undesirable: $F_{(1,26)} = 12.9$, $p < 0.005$, $\eta_p^2 = 0.33$; Fig. 2*B*). Positively biased feedback processing did not differ between self-directed and other-directed feedback (main effect: self/other: $p > 0.1$; interaction: $p > 0.6$). In a follow-up analysis we confirmed that similar results were observed, when comparing absolute mean updates that were not scaled by the respective mean feedback discrepancies (main effect: desirable/undesirable: $F_{(1,26)} = 15.0$, $p < 0.001$, $\eta_p^2 = 0.37$; main effect: self/other: $p > 0.1$; interaction: $p > 0.8$). The magnitude of mean feedback discrepancies was equal across conditions ($p > 0.1$).

Additionally, we investigated possible influences on the positive updating bias. When participants gave the highest rating possible, they could not receive a feedback rating higher than their own rating and thus feedback could not be desirable. The reverse was true when participants gave the lowest rating possible. To exclude that this relationship between first ratings and feedback compromised our results we tested for positively biased updating only for trials with first ratings in the middle range of the scale (4, 5, and 6). Updating for desirable versus undesirable feedback was still higher when including only trials with first ratings in the middle range of the scale (main effect: desirable/undesirable: $F_{(1,26)} = 14.4$, $p < 0.001$, $\eta_p^2 = 0.36$; main effect: self/other: $p > 0.8$; interaction: $p > 0.5$). This analysis excluded the possibility that positively biased updating was driven by trials in which participants initially rated themselves or the other person on the highest or lowest points of the scale.

Furthermore, we tested whether the valence of the trait adjectives had an effect on updating. We split update scores according to the valence of the trait words and performed a 2 (trait valence: positive/negative) by 2 (feedback target: self/other) by 2 (desirability: desirable/undesirable) ANOVA on absolute mean updates divided by absolute mean feedback discrepancies. Only the main effect of desirability reached significance ($F_{(1,26)} = 13.0$, $p < 0.005$, $\eta_p^2 = 0.33$). Specifically, the interaction between the factors trait valence and desirability did not reach significance ($p > 0.9$), excluding the possibility that trait valence had an effect on positively biased updating in our paradigm.

In sum, our behavioral results establish that humans take desirable feedback more into account than undesirable feedback.

Behavioral results—control analyses and individual differences

For an additional control analysis, participants recollected outside the scanner the feedback rating they had seen inside the scanner. Mean absolute memory errors were smaller for self-related than for other-related feedback ($F_{(1,26)} = 25.4$, $p < 0.0001$, $\eta_p^2 = 0.49$) but did not differ between desirable and undesirable feedback ($p > 0.1$). Furthermore, we conducted two separate ANCOVAs, one for self and one for other, testing whether the difference between desirable and undesirable updates remained significant when entering additional scores as covariates. These scores were the differences between trials with desirable and undesirable feedback for first ratings, participants' social desirability ratings of the trait adjectives, memory errors, or reaction times on the first or second ratings. The difference between desirable and undesirable updates remained significant when controlling for these scores (self: $F_{(1,21)} = 15.8$, $p < 0.001$; other: $F_{(1,21)} = 8.6$, $p < 0.01$). Moreover, winning or losing in the board game that participants played to get to know each other before receiving feedback, did not have any influence on behavior during the task. Specifically, participants' rank order in the game did not correlate with mean ratings or any update measure using Spearman's correlation coefficient (all $p > 0.1$). Thus, positively biased updating could not be explained by differential memory, first ratings, social desirability ratings of the trait adjectives, valence of the trait words, reaction times, or performance in the board game.

Next, we aimed to establish links between performance in our task and individual differences in trait self-esteem and perceived similarity between self and other. As expected, mean first ratings for self correlated significantly with scores on the Rosenberg self-esteem scale across participants (Pearson's $r = 0.59$, $p < 0.005$); the higher a participant's trait self-esteem the more positive is his or her mean rating across all trait adjectives. Mean first ratings for the other person correlated with perceived similarity to the other person, which was assessed on a Likert scale from 1 (not similar at all) to 8 (very similar; Pearson's $r = 0.51$, $p < 0.01$). Thus, mean ratings in our task were related to intersubject differences in trait

Table 2. Significant activations in feedback onsets

	Side	Brodmann area	Peak voxel MNI coordinates (mm)			Cluster size (voxels at $p < 0.0001$)	p (cluster FWE corrected)	Peak z score
			x	y	z			
Feedback onset: self > other								
MPFC	L/R	10/9/8/6/32/24	−3	59	28	1602	<0.001	7.60
IFG (orbital part)/anterior insula	L	47/11/13/45/38	−33	17	−17	399	<0.001	7.25
IFG (orbital part)/anterior insula	R	47/11/13/38	30	20	−17	335	<0.001	7.18
Cerebellum	R	—	30	−82	−35	161	<0.001	5.98
Cerebellum	L	—	−30	−85	−38	77	<0.001	5.41
Midbrain	L/R	—	−12	−13	−14	381	<0.001	5.38
Cerebellum	L/R	—	3	−55	−35	30	0.017	4.71
Caudate body	L	—	−9	8	16	58	0.002	4.45
Feedback onset: other > self								
Precuneus/postcentral gyrus/superior temporal gyrus/supramarginal gyrus	L/R	7/6/4/1/2/3/5/18/22/40	12	−46	52	7102	<0.001	6.89
Middle temporal gyrus	R	38	51	−64	10	82	<0.001	4.86
Precentral gyrus	R	4	39	−10	58	172	<0.001	4.67
Middle frontal gyrus	R	9	27	29	40	39	0.007	4.42
Middle frontal gyrus	L	9	−30	35	25	27	0.023	4.19
Middle frontal gyrus	L	10	−36	50	13	20	0.047	4.17

All reported clusters are familywise error (FWE) corrected for multiple comparisons at $p < 0.05$; cluster-defining threshold of $p < 0.0001$. IFG, inferior frontal gyrus; MPFC, medial prefrontal cortex.

self-esteem and perceived similarity of the other person. In a next step, we explored how first ratings were related to the update bias across participants. Mean first self ratings did not correlate significantly with the updating bias for self (Pearson's $r = 0.04$, $p > 0.8$). However, mean ratings for the other person did correlate significantly with the magnitude of the update bias for this other person, i.e., the absolute relative mean update for desirable minus undesirable feedback (Pearson's $r = 0.51$, $p < 0.01$). This suggests that the higher the other person was rated on average the more pronounced was the positively biased updating pattern. Positively biased updating for self seemed to be unrelated to mean self ratings in our sample.

Behaviorally, participants showed a positively biased updating pattern after receiving feedback. Therefore, we turned to our fMRI data to establish a link between biased updating and blood oxygenation level-dependent (BOLD) signals related to feedback processing. Specifically, we examined reward-related and comparison-related components of feedback processing.

BOLD signals for self-related versus other-related feedback

In an initial step, we examined brain activity during the feedback period to find regions in which activation differed between the processing of self-related and other-related feedback. We expected regions previously implicated in self-referential and other-referential processing, notably the MPFC (Amodio and Frith, 2006). Contrasting the time point when participants received self-related versus when they received other-related feedback (self > other), we found activity in the medial prefrontal wall (peak voxel in MNI coordinates x, y, z: −3, 59, 28; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$) as well as bilaterally in the orbital part of the IFG extending into the anterior insula (left: −33, 17, −17; right: 30, 20, −17; see Table 2 for a comprehensive list of activations). The reverse contrast (other > self) revealed among other regions activity in bilateral precuneus (12, −46, 52; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; Table 2).

During the imagination phase, participants in our task rated themselves and another person in a similar way as shown in many previous studies (Northoff et al., 2006; Denny et al., 2012; Fig. 1*B*). Therefore, we wanted to explore possible differences in activity between feedback phase and imagination phase in a follow-up ROI analysis. We concentrated this analysis to the MPFC since this region has consistently been implicated in self-related processing. We extracted parameter estimates during both time points within an ROI that was independently defined based on a recent meta-analysis of self-referential processing (Denny et al., 2012; sphere with a radius of 15 mm centered at the MNI coordinates −6, 50, 4). Parameter estimates were compared in a 2 (imagination/feedback phase) by 2 (self/other) ANOVA. As expected there was a significant main effect of activity for self being higher than for other ($F_{(1,26)} = 126.1$, $p < 0.0001$, $\eta_p^2 = 0.83$). There was also a significant main effect of phase with higher activity during the feedback phase than during the imagination phase ($F_{(1,26)} = 9.2$, $p < 0.01$, $\eta_p^2 = 0.26$). The interaction was not significant ($p > 0.1$).

These results show that self-relevant feedback implicates the MPFC as has been reliably shown for self-referential processing in general (Amodio and Frith, 2006; Northoff et al., 2006; Denny et al., 2012).

BOLD signals related to the rewarding component of social feedback

The behavioral analyses showed that participants processed desirable feedback more than undesirable feedback. We hypothesized that neural activity during the feedback phase should mirror two aspects of social feedback processing: a reward-related aspect (operationalized by feedback ratings) and a comparison-related aspect (operationalized by feedback discrepancies). To identify neural activity related to these two components, we used the full parametric range of feedback ratings and feedback discrepancies. Our model included separate onset regressors for self-related and other-related feedback, which were parametrically modulated by the respective feedback ratings and feedback discrepancies. This model allowed us to search for regions in which these parameters correlated with BOLD signals in a trial-by-trial fashion.

To test for activity correlating with the rewarding component of feedback at the time point of feedback, we performed a contrast between the two parametric modulators for feedback ratings (feedback ratings for self and feedback ratings for other). First, activity related to reward should correlate positively with feedback ratings for self. That is, the higher the

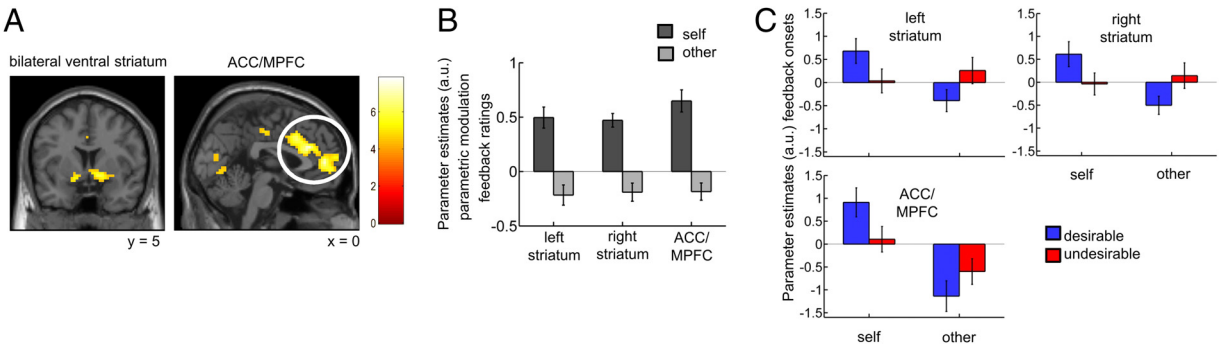

Figure 3. BOLD signals related to the rewarding component of social feedback. ***A***, BOLD signal changes in bilateral ventral striatum and ACC/MPFC correlated with the rewarding component of feedback on a trial-by-trial basis at the time point of feedback (all clusters are significant at $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$). Reward-related activity fulfilled two requirements. First, activity correlated positively with feedback ratings for self since higher feedback ratings for self indicated more rewarding feedback (e.g., a feedback rating of 8.0 on polite is more rewarding than a feedback rating of 7.0; feedback ratings for negative trait adjectives were reverse coded). Second, activity correlated more with the feedback ratings for self than for other since we searched for regions in which reward-related activity was self-specific. ***B***, For illustration purposes, we plotted parameter estimates of the parametric modulators for feedback ratings for self and other within functional ROIs. ***C***, To explore differences in onset activity, we plotted parameter estimates of the onset regressors within the functional ROIs in a second model that included separate regressors for feedback target and feedback desirability. Error bars indicate SEM.

Table 3. BOLD signals related to the rewarding component of social feedback: parametric analysis—feedback ratings

			Peak voxel MNI coordinates (mm)					
	Side	Brodmann area	*x*	*y*	*z*	Cluster size (voxels at $p < 0.0001$)	*p* (cluster FWE corrected)	Peak *z* score
Feedback rating (trial-by-trial correlation): self > other								
ACC/mid-cingulate cortex/MPFC	L/R	32/24/9/10	3	32	25	414	< 0.001	5.56
Ventral striatum (caudate head and putamen)	R	—	12	5	−8	71	< 0.001	4.73
Thalamus	R	—	21	−13	22	20	0.032	4.60
Ventral striatum (caudate head and putamen)	L	—	−15	2	−11	25	0.017	4.50
Cerebellum	L	—	−33	−73	−23	57	0.001	4.48
Cerebellum	R	—	39	−58	−26	39	0.004	4.35
Cerebellum	R	—	12	−61	−17	47	0.002	4.33
Lingual gyrus	L	18	−3	−73	−5	26	0.015	4.31
Calcarine fissure	L/R	18	3	−82	13	29	0.011	4.13
Mid-cingulate cortex	L/R	24	0	−19	43	25	0.017	4.06

All reported clusters are familywise error (FWE) corrected for multiple comparisons at $p < 0.05$; cluster-defining threshold of $p < 0.0001$. ACC, anterior cingulate cortex; MPFC, medial prefrontal cortex.

feedback rating the more rewarding was the social feedback (e.g., receiving a self-related feedback rating of 8 is more rewarding than a feedback rating of 7). Note that feedback ratings for negative trait adjectives were reverse coded so that a higher feedback rating always indicated a more positive feedback. Second, reward-related activity should be self-specific. That is, the trial-by-trial correlation of BOLD signal changes with the feedback ratings for self should be greater than those for other. Contrasting the parametric modulators for the feedback ratings for self versus other revealed activity in bilateral ventral striatum (left: −15, 2, 11, right: 12, 5, −8) and in a region encompassing ACC and MPFC (3, 32, 25; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; Fig. 3*A*; see Table 3 for a comprehensive list of activations).

To better illustrate the correlations between feedback ratings and neural activity we performed a follow-up analysis. We extracted parameter estimates of the parametric modulators for self and other within the functional ROIs identified in the above contrast (Fig. 3*B*). Parametric modulators indicate the correlation (i.e., the slope) between BOLD signals and feedback ratings but give no information about mean onset activity (i.e., the intercept). To additionally illustrate mean onset activity, we estimated a follow-up GLM. In this follow-up model, trials were separated into four categories according to feedback target (self/other) and feedback desirability (desirable/undesirable) in the same way as in the main behavioral analysis. We extracted parameter estimates of the onset regressors for the four categories within three of the functional ROIs defined by the main model (MPFC and left and right striatum). Plotting these onset regressors illustrates the interaction of feedback target and desirability as defined by the contrast in the main model (Fig. 3*C*). Additionally, mean onset activity showed a significant main effect for self versus other in the MPFC ($F_{(1,26)} = 29.9, p < 0.0001$; since we performed an ANOVA within each of the three ROIs, *p* values were adjusted using a Bonferroni correction for the number of ROIs). In the right striatum the same pattern was observed at trend level ($F_{(1,26)} = 6.1, p = 0.06$).

Additionally, we performed the reverse contrast to the one performed above, i.e., we searched for regions that correlated with other-related feedback ratings more than with self-related feedback ratings. This contrast revealed no significant voxels at a threshold of $p < 0.05$ cluster-corrected at a cluster-defining threshold of $p < 0.0001$.

Together, the rewarding component of social feedback correlated with activity in ventral striatum and ACC/MPFC, regions previously implicated in processing social and nonsocial rewards (Izuma et al., 2008; Beckmann et al., 2009).

BOLD signals related to the comparison component of social feedback

Having identified activity correlating with the rewarding aspect of feedback, we next tested for BOLD signal changes correlating

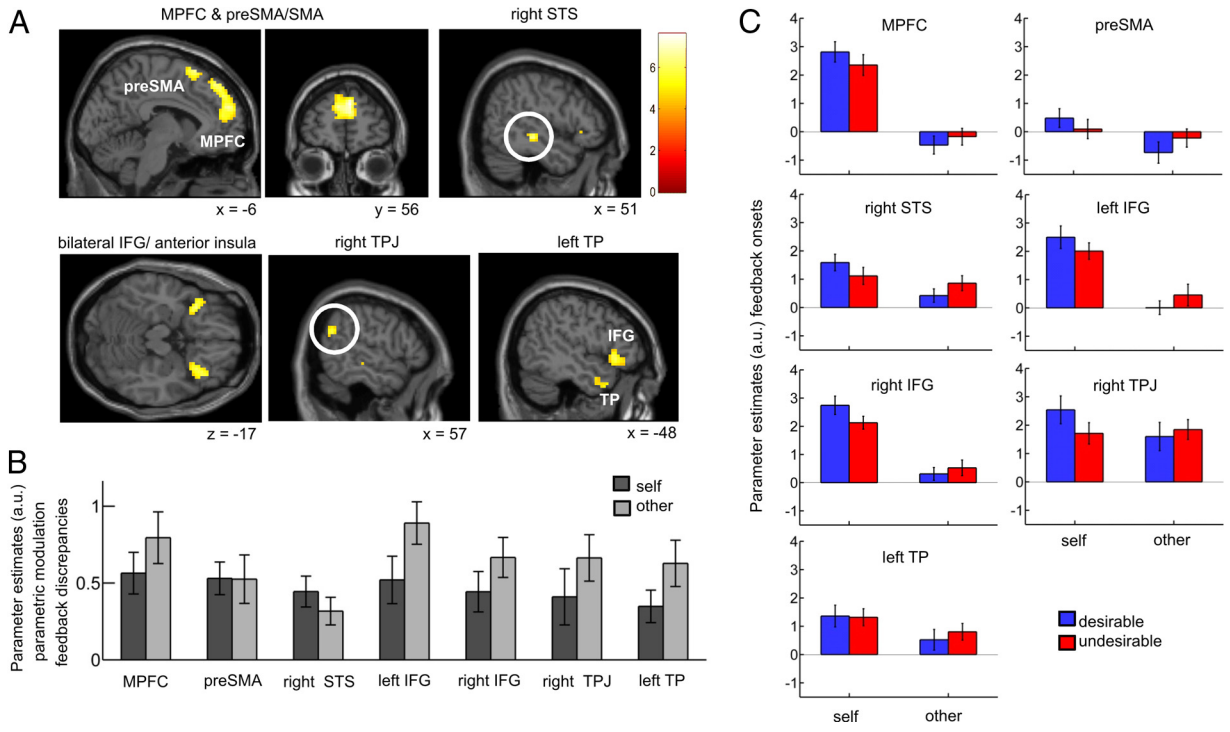

Figure 4. BOLD signals related to the comparison component of social feedback. ***A***, BOLD signal changes in the following regions correlated with the comparison-related component of feedback on a trial-by-trial basis at the time point of feedback: MPFC, pre-SMA/SMA, bilateral IFG (orbital part) extending into anterior insula, right STS, right TPJ, and left TP (all clusters are significant at $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$). Comparison-related activity correlated positively with the feedback discrepancies for both self and other, i.e., with the absolute difference between participants' own views and the feedback they received. ***B***, For illustration purposes, we plotted parameter estimates of the parametric modulators for feedback discrepancies for self and other within functional ROIs. ***C***, To explore differences in onset activity, we plotted parameter estimates of the onset regressors within the functional ROIs in a second model that included separate regressors for feedback target and feedback desirability. Error bars indicate SEM.

with the comparison-related aspect on a trial-by-trial basis at the time point of feedback—both for self-related and other-related feedback. Comparison-related activity was operationalized as activity that showed a positive correlation with feedback discrepancies, i.e., with the absolute differences between participants' own ratings and the feedback ratings they received. That is, a larger feedback discrepancy (e.g., a difference between own rating and feedback rating of 2) indicated more "need" for a comparison process than a smaller feedback discrepancy (e.g., 1) regardless of the direction of the difference.

Feedback discrepancies for both self and other correlated positively with activity in MPFC (6, 56, 28), pre-supplementary motor area/supplementary motor area (preSMA/SMA (9, 17, 64), right STS (51, −25, −8), bilateral IFG (orbital part) extending into anterior insula (left: −36, 20, −23, right: 33, 20, −17), right TPJ (57, −58, 25), left TP (−48, 11, −35), and left cerebellum (−24, −82, −35; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; Fig. 4*A*; Table 4).

As described above for BOLD signal changes related to social reward, we plotted parameter estimates of the parametric modulators for the feedback discrepancies for self and other within the functional ROIs to better illustrate the correlations of feedback discrepancies and BOLD signals (Fig. 4*B*). These parametric modulators indicate the positive correlation (i.e., the slope) between BOLD signals and feedback discrepancies but give no information about the mean onset activity (i.e., the intercept). To extract mean onset activity within seven of the functional ROIs defined by the first model (MPFC, pre-SMA/SMA, right STS, left and right IFG, right TPJ, and left TP), we conducted a follow-up analysis using a follow-up GLM, in which trials were separated into four categories according to feedback target and feedback desirability. Mean onset activity showed a significant main effect for self versus other in MPFC ($F_{(1,26)} = 116.7$, $p < 0.0001$), right STS ($F_{(1,26)} = 9.8$, $p < 0.05$), left IFG ($F_{(1,26)} = 53.6$, $p < 0.0001$), right IFG ($F_{(1,26)} = 77.8$, $p < 0.0001$), and left TP ($F_{(1,26)} = 8.7$, $p < 0.05$; since we performed an ANOVA within each of the seven ROIs, p values were adjusted using a Bonferroni correction for the number of ROIs; Fig. 4*C*). Thus, while the relation of feedback discrepancies to BOLD signal changes was the same for self-related and other-related feedback (as determined in the first model with the parametric modulators), mean activity was higher for self-related versus other-related feedback within these regions (as determined in the second model in which trials were split up into categories for desirable and undesirable feedback).

We also searched for regions in which BOLD signals correlated negatively with the feedback discrepancies for self and other (Table 4). BOLD signal changes in no region correlated differentially for self-related versus other-related feedback discrepancies, i.e., self > other or other > self, at a threshold of $p < 0.05$ cluster-corrected at a cluster-defining threshold of $p < 0.0001$.

In sum, both in the self-condition and in the other-condition the difference between participants' own views and the feedback they received, i.e., the comparison-related component, correlated with activity in regions previously implicated in mentalizing (Mar, 2011).

Updating bias for self and activity integrating reward-components and comparison-components

Having identified activity that correlated with the rewarding aspect of feedback and activity that correlated with the

Table 4. BOLD signals related to the comparison component of social feedback: parametric analyses—feedback discrepancies and conjunction

	Side	Brodmann area	Peak voxel MNI coordinates (mm)			Cluster size (voxels at $p < 0.0001$)	p (cluster FWE corrected)	Peak z score
			x	y	z			
Feedback discrepancies (positive trial-by-trial correlation): self and other								
MPFC	L/R	10/9/8/6	6	56	28	383	<0.001	5.47
Pre-SMA/SMA	L/R	8/6	9	17	64	104	<0.001	4.98
Superior/middle temporal gyrus (STS)	R	21	51	−25	−8	26	<0.001	4.93
IFG (orbital part)/ anterior insula	L	47/45/13	−36	20	−23	181	<0.001	4.88
IFG (orbital part)/anterior insula	R	47/13/11	33	20	−17	117	<0.001	4.69
Angular gyrus, TPJ	R	39/40	57	−58	25	19	0.045	4.56
TP	L	21/38	−48	11	−35	22	0.032	4.25
Cerebellum	L	—	−24	−82	−35	20	0.040	4.19
Feedback discrepancies (negative trial-by-trial correlation): self and other								
Inferior parietal lobule	L	40	−54	−37	46	154	<0.001	5.07
Middle temporal gyrus	R	21/37	60	−49	−8	53	0.002	4.54
Superior parietal gyrus	L	7	−21	−49	64	30	0.013	4.50
Superior temporal gyrus	L	22/6	−54	−10	1	37	0.007	4.40
Inferior parietal lobule	R	40	51	−37	46	48	0.002	4.34
Precentral gyrus/superior temporal gyrus	R	6/22	54	5	13	30	0.013	4.19
Conjunction of feedback rating (trial-by-trial correlation): self > other with feedback discrepancies (positive trial-by-trial correlation): self and other								
MPFC/ACC	L/R	10	3	56	19	25	0.023	5.01

All reported clusters are familywise error (FWE) corrected for multiple comparisons at $p < 0.05$; cluster-defining threshold of $p < 0.0001$. IFG, inferior frontal gyrus; MPFC, medial prefrontal cortex; SMA, supplementary motor area; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporoparietal junction.

comparison-related aspect of feedback, we next examined how neural activity was linked to the behavioral update bias for self. We postulated that neural activity mediating the update bias for self should fulfill two requirements. First, candidate regions should integrate activity related to both reward and comparison. Second, activity within this region should correlate with the behavioral update bias for self across participants, i.e., the difference between updates for desirable and undesirable feedback. To address the first requirement, we performed a conjunction analysis testing the conjunction null hypothesis to search for regions that were activated by both reward-related and comparison-related components. The conjunction revealed a region at the border of the MPFC and the ACC (3, 56, 19; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; Fig. 5*A*; Table 4). To address the second requirement, we extracted parameter estimates of self-related absolute feedback discrepancies within this region and tested for a correlation with the behavioral update bias for self. Parameter estimates of self-related absolute feedback discrepancies within the functional ROI defined by the conjunction analysis predicted the behavioral update bias for self (Pearson's $r = 0.42$, $p < 0.05$, 95% confidence interval [0.05–0.69]; Fig. 5*B*). Additionally, we extracted parameter estimates of other-related feedback discrepancies errors within the same functional ROI and found no correlation with the behavioral update bias for other (Pearson's $r = 0.07$, $p > 0.7$). However, we note that the two Pearson's correlation coefficients did not differ significantly ($z = 1.3$, $p > 0.1$, using the z_2* statistic described by Steiger, 1980).

Thus, BOLD signals within the MPFC that integrated reward-related and comparison-related components of social feedback predicted individual differences in self-related positive updating.

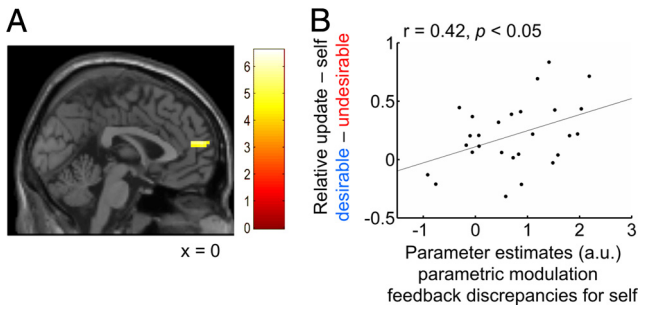

Figure 5. Updating bias for self and activity integrating reward-components and comparison-components. ***A***, Conjunction analysis of activity correlating with the rewarding aspect of feedback, i.e., feedback ratings (Fig. 3) and of activity correlating with the comparison-related aspect of feedback, i.e., feedback discrepancies (Fig. 4; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$). ***B***, Across participants, parameter estimates of the parametric modulators for self-related absolute feedback discrepancies within this region predicted the behavioral update bias, i.e., the relative mean update for self-related desirable minus undesirable feedback. Each dot represents one participant and the line is the regression slope.

Discussion

After interacting with peers in a real-life setting and then receiving social feedback from them, participants showed positively biased updating of their self-evaluations and other-evaluations. Specifically, participants updated their evaluation of themselves and of another peer more toward desirable feedback than toward undesirable feedback. Our fMRI data suggest that neural activity reflects two different components of social feedback. First, activity in the bilateral ventral striatum and in a region encompassing parts of the ACC/MPFC tracked the rewarding component. Second, parts of the mentalizing network tracked the comparison-related component. Changes in activity within the MPFC that integrated reward-related and comparison-related aspects of feedback predicted the self-related positive updating bias across participants. Our results suggest that a combination of neural signals related to social reward and to the comparison between own views and feedback mediate positively biased feedback processing.

So far only a few studies in social neuroscience have investigated social feedback processing (Somerville et al., 2006,;

2010; Izuma et al., 2008; Davey et al., 2010; Eisenberger et al., 2011; Jones et al., 2011). Three crucial aspects of our design allowed us to considerably add to these studies. First, participants in our task engaged in a real-life interaction of more than an hour whereas in previous studies participants received feedback that was based on photographs and/or questionnaires. Second, we parametrically modulated the desirability of the feedback and even more importantly we assessed the difference between participants' self-views and the feedback they received. In previous studies feedback was mostly binary and participants did not indicate their own view, e.g., participants just got to know whether they were liked or not or whether a certain trait word applied to them. Third, by assessing how feedback changed self-views we demonstrate a positivity bias in feedback processing.

Positivity biases have been documented across many domains in social cognition (Leary, 2007) and it has been proposed that they arise because cognitive processing mechanisms distort incoming information in a positive direction (Taylor and Brown, 1988). Here, we provide evidence for this idea by showing a striking asymmetry in how humans process self-relevant information about their character traits. A similar approach has been used in the domain of optimism (Sharot et al., 2011). Participants estimated their likelihood of experiencing various negative events in the future. They updated their beliefs more toward the actual statistical likelihood when it was desirable than when it was undesirable. Thus, our results suggest that positivity biases in general may arise due to asymmetric information processing.

Some recent studies investigating social conformity have used designs similar to the present study (Klucharev et al., 2009; Campbell-Meiklejohn et al., 2010; Zaki et al., 2011). In these studies participants make a first evaluation of an object (e.g., an unknown face or song) and then receive feedback from others about this object. Conformity can then be measured as the degree to which participants change their evaluation toward the others' opinion similar to the update measure in our study. However, in these conformity studies participants are unbiased (e.g., they are influenced to the same degree when they see that others judge an unknown song to be better or worse than they do). In contrast, participants in our study processed social information in a positively biased way since the "object of conformity" consisted of participants' own character traits and the character traits of peers.

Behavioral studies have commonly discussed positively biased self-views with relation to theories about the self but not with relation to theories about reward. Here, we specify neural activity related to the rewarding component of positively biased feedback processing. Using a parametric design we show that the ventral striatum and the ACC/MPFC process the self-related reward associated with social feedback. Our results thus replicate the findings by Izuma et al. (2008) and extend them to negative character traits. The striatum and the ACC/MPFC, especially its middle and more ventral parts, are connected and both structures have been linked to reward in social and nonsocial contexts (Beckmann et al., 2009). Interestingly, activity in the ventral part of the MPFC plays a role in the representation of the value of objects (Rangel et al., 2008; Rushworth et al., 2011) and this activity can be modulated by social influences (Plassmann et al., 2008; Zaki et al., 2011). In sum, our results corroborate that social reward processing can be linked to structures involved in nonsocial reward processing.

Critically, in addition to the rewarding aspect of social feedback our task also modulated the distance between participants' own views and the feedback they received. This comparison between own views and feedback correlated among others with activity in the MPFC, right STS, bilateral IFG, right TPJ, and left TP. All of these regions are part of the mentalizing network (Amodio and Frith, 2006; Bahnemann et al., 2010). Especially, the MPFC and the TPJ have been most consistently linked to various mentalizing tasks (Van Overwalle, 2009; Mar, 2011). In our study, MPFC activity showed stronger activation than TPJ activity, which is consistent with a recent meta-analysis (Mar, 2011) showing that the MPFC is particularly involved in tasks that are not based on explicit false belief stories as was the case in our task. Furthermore, such nonstory-based tasks often implicate the orbital IFG (Mar, 2011) and we therefore interpret orbital IFG activity in relation to its plausible role in mentalizing associated processes.

It is important to note that changes in neural activity in the mentalizing network have a very parsimonious interpretation in our task. Activity in the mentalizing network tracked the numerical difference between participants' own evaluations and the feedback they received both for their own character and for the character of another person. Recently, some studies have begun to investigate neural activity related to social cognition by using computational parameters derived from modified versions of reinforcement models or other types of computational models (Behrens et al., 2008; Hampton et al., 2008; Coricelli and Nagel, 2009; Yoshida et al., 2010; Biele et al., 2011). These studies provide first steps toward conceptualizing the precise computations underlying activity in the mentalizing network or parts of it (Behrens et al., 2009). In line with the results of these studies, our results provide a mechanistic account of activity in the mentalizing network for the processing of social feedback.

Our results show that the behavioral updating bias for self is associated with both reward-related and comparison-related components of social feedback. Activity within a region in the MPFC that integrated the two components predicted the amount of positively biased updating for self-related feedback across participants. This was not the case for other-related feedback. Behavioral accounts (Taylor and Brown, 1988) have argued for a filtering mechanism that distorts incoming social information toward the positive. Our results suggest that MPFC activity reflects this filtering mechanism in our task.

The implication of the MPFC in social cognition in general and in self-related processing in particular has been reliably shown by many studies (Mitchell, 2009; Denny et al., 2012; Wagner et al., 2012). Importantly, Moran et al. (2006) have shown that MPFC activity was higher when participants made trait ratings that were self-descriptive compared with when they made self-ratings that were not self-descriptive—independent of trait valence. The MPFC region that integrated reward-related and comparison-related components in our task was within the region described by Moran et al. (2006). This suggests that neural processes related to thinking about trait self-descriptiveness and neural processes related to receiving feedback on trait self-descriptiveness might be instantiated in a common MPFC region. The relation of the MPFC to self-related positively biased updating is also concordant with a previous study in which participants received information that they were either liked or not liked by other persons (Somerville et al., 2010). In this study, trait self-esteem corre-

lated with the differential activity toward positive versus negative feedback in a similar region of the MPFC as identified in our task. Importantly, our results are also in line with literature linking different subregions of the MPFC to reward processing, self-referential thinking, and mentalizing. Specifically, reward and value processing have been consistently linked to a ventral MPFC region at the border of pregenual ACC and medial OFC (Rangel et al., 2008; Beckmann et al., 2009; Rushworth et al., 2011). Self-referential thinking most consistently involves a ventral part of the MPFC whereas mentalizing involves a more dorsal part (Northoff et al., 2006; Mar, 2011; Denny et al., 2012). The MPFC regions that integrated reward-related and comparison-related components of social feedback in our task lay at a border position in which there might be some overlap between reward-, self-, and mentalizing-related activity. Our results suggest that this MPFC region seems to be ideally suited for positively biased integration of social information into one's self-concept and that it might be interesting to investigate this region's involvement in reward-, self-, and mentalizing-related processes more closely.

Conclusions

Many studies have tried to weigh the benefits (e.g., reduced anxiety) and costs (e.g., overly risky decision making) of positivity biases against each other (Taylor and Brown, 1988; Leary, 2007). Positivity biases seem to be generally adaptive but can be detrimental if they are too extreme. To further specify their costs and benefits, it is fundamental to understand the underlying mechanisms. Our results show that positively biased social feedback processing is related to an integration of activity linked to reward and mentalizing. This underscores the importance of integrating theories on reward and mentalizing. By directing the focus toward the interplay between reward processing and mentalizing, we provide an essential step toward a better understanding of how social information is integrated into the human self-concept.

References

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277. CrossRef Medline

Anderson NH (1968) Likableness ratings of 555 personality-trait words. J Pers Soc Psychol 9:272–279. CrossRef Medline

Bahnemann M, Dziobek I, Prehn K, Wolf I, Heekeren HR (2010) Sociotopy in the temporoparietal cortex: common versus distinct processes. Soc Cogn Affect Neurosci 5:48–58. CrossRef

Beckmann M, Johansen-Berg H, Rushworth MF (2009) Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J Neurosci 29:1175–1190. CrossRef Medline

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. Nature 456:245–249. CrossRef Medline

Behrens TE, Hunt LT, Rushworth MF (2009) The computation of social behavior. Science 324:1160–1164. CrossRef Medline

Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. PLoS Biol 9:e1001089. CrossRef Medline

Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD (2010) How the opinion of others affects our valuation of objects. Curr Biol 20:1165–1170. CrossRef Medline

Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc Natl Acad Sci U S A 106:9163–9168. CrossRef Medline

Davey CG, Allen NB, Harrison BJ, Dwyer DB, Yücel M (2010) Being liked activates primary reward and midline self-related brain regions. Hum Brain Mapp 31:660–668. Medline

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24:1742–1752. CrossRef Medline

Eisenberger NI, Inagaki TK, Muscatell KA, Byrne Haltom KE, Leary MR (2011) The neural sociometer: brain mechanisms underlying state self-esteem. J Cogn Neurosci 23:3448–3455. CrossRef Medline

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419–427. CrossRef Medline

Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A (2007) Social comparison affects reward-related brain activity in the human ventral striatum. Science 318:1305–1308. CrossRef Medline

Fossati P, Hevenor SJ, Graham SJ, Grady C, Keightley ML, Craik F, Mayberg H (2003) In search of the emotional self: an fMRI study using positive and negative emotional words. Am J Psychiatry 160:1938–1945. CrossRef Medline

Gilbert SJ, Spengler S, Simons JS, Steele JD, Lawrie SM, Frith CD, Burgess PW (2006) Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. J Cogn Neurosci 18:932–948. CrossRef Medline

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585–595. CrossRef Medline

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A 105:6741–6746. CrossRef Medline

Heister J, Würzner K-M, Bubenzer J, Pohl E, Hanneforth T, Geyken A, Kliegl R (2011) dlexDB—eine lexikalische Datenbank für die psychologische und linguistische Forschung. Psychologische Rundschau 62:10–20. CrossRef

Hepper EG, Hart CM, Gregg AP, Sedikides C (2011) Motivated expectations of positive feedback in social interactions. J Soc Psychol 151:455–477. CrossRef Medline

Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. Neuron 58:284–294. CrossRef Medline

Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ (2011) Behavioral and neural properties of social reinforcement learning. J Neurosci 31:13039–13045. CrossRef Medline

Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G (2009) Reinforcement learning signal predicts social conformity. Neuron 61:140–151. CrossRef Medline

Leary MR (2007) Motivational and emotional aspects of the self. Annu Rev Psychol 58:317–344. CrossRef Medline

Lieberman MD (2007) Social cognitive neuroscience: a review of core processes. Annu Rev Psychol 58:259–289. CrossRef Medline

Mar RA (2011) The neural bases of social cognition and story comprehension. Annu Rev Psychol 62:103–134. CrossRef Medline

Mitchell JP (2009) Social psychology as a natural kind. Trends Cogn Sci 13:246–251. CrossRef Medline

Moran JM, Macrae CN, Heatherton TF, Wyland CL, Kelley WM (2006) Neuroanatomical evidence for distinct cognitive and affective components of self. J Cogn Neurosci 18:1586–1594. CrossRef Medline

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660. CrossRef Medline

Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain – a meta-analysis of imaging studies on the self. Neuroimage 31:440–457. CrossRef Medline

Plassmann H, O'Doherty J, Shiv B, Rangel A (2008) Marketing actions can modulate neural representations of experienced pleasantness. Proc Natl Acad Sci U S A 105:1050–1054. CrossRef Medline

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556. CrossRef Medline

Rosenberg M (1965) Society and the adolescent self-image. Princeton, NJ: Princeton UP.

Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054–1069. CrossRef Medline

Saxe R (2006) Uniquely human social cognition. Curr Opin Neurobiol 16:235–239. CrossRef Medline

Saxe R, Powell LJ (2006) It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17:692–699. CrossRef Medline

Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. Nat Neurosci 14:1475–1479.

Somerville LH, Heatherton TF, Kelley WM (2006) Anterior cingulate cortex responds differentially to expectancy violation and social rejection. Nat Neurosci 9:1007–1008. CrossRef Medline

Somerville LH, Kelley WM, Heatherton TF (2010) Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. Cereb Cortex 20:3005–3013. CrossRef Medline

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251. CrossRef

Taylor SE, Brown JD (1988) Illusion and well-being: a social psychological perspective on mental health. Psychol Bull 103:193–210. CrossRef Medline

Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829–858. CrossRef Medline

Võ ML, Jacobs AM, Conrad M (2006) Cross-validating the Berlin affective word list (BAWL). Behav Res Methods 38:606–609. CrossRef Medline

Wagner DD, Haxby JV, Heatherton TF (2012) The representation of self and person knowledge in the medial prefrontal cortex. Wiley Interdiscip Rev Cogn Sci 3:451–470. CrossRef

Walter H, Adenzato M, Ciaramidaro A, Enrici I, Pia L, Bara BG (2004) Understanding intentions in social interaction: the role of the anterior paracingulate cortex. J Cogn Neurosci 16:1854–1863. CrossRef Medline

Wolf I, Dziobek I, Heekeren HR (2010) Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. Neuroimage 49:894–904. CrossRef Medline

Wunderlich K, Symmonds M, Bossaerts P, Dolan RJ (2011) Hedging your bets by learning reward correlations in the human brain. Neuron 71:1141–1152. CrossRef Medline

Yoshida W, Seymour B, Friston KJ, Dolan RJ (2010) Neural mechanisms of belief inference during cooperative games. J Neurosci 30:10744–10751. CrossRef Medline

Zaki J, Schirmer J, Mitchell JP (2011) Social influence modulates the neural computation of value. Psychol Sci 22:894–900. CrossRef Medline