Attributed social context and emotional content recruit frontal and limbic brain regions during virtual feedback processing

Sebastian Schindler^{1,2} · Onno Kruse^{3,4} · Rudolf Stark³ · Johanna Kissler^{2,5}

Published online: 9 November 2018
© Psychonomic Society, Inc. 2018

Abstract

In communication, who is communicating can be just as important as what is said. However, sender identity in virtual communication is often inferred rather than perceived. Therefore, the present research investigates the brain structures activated by sender identity attributions and evaluative feedback processing during virtual communication. In a functional magnetic resonance imaging (fMRI) study, 32 participants were told that they would receive personality feedback, either sent from another human participant or from a randomly acting computer. In reality, both conditions contained random but counterbalanced feedback, automatically delivered by approving or denying negative, neutral, or positive adjectives. Although physically identical, feedback attributed to the “human” sender activated multiple regions within a “social brain” network, including the superior frontal, medial prefrontal, and orbitofrontal cortex, anterior and posterior parts of the cingulate cortex, and the bilateral insula. Regardless of attributed sender, positive feedback increased responses in the striatum and bilateral amygdalae, while negative compared to neutral feedback elicited stronger insula and somatosensory responses. These results reveal the recruitment of an extensive mentalizing and social brain network by mere sender attributions and the activation of brain structures related to reward and punishment by verbal feedback, demonstrating its embodied processing.

Keywords Virtual communication · Social context · Social feedback · Emotion · Language · fMRI

Introduction

Communicating with others is a core social behavior. In virtual communication, however, chatbots increasingly emulate human interaction (Lee, 2016), essentially faking human social

presence. Masked as “humans,” these bots spread content, interact, and try to influence public opinion (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). The use of such guises foregrounds an important aspect of communication: People react very differently to other humans compared to how they react to

Sebastian Schindler and Onno Kruse contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13415-018-00660-5>) contains supplementary material, which is available to authorized users.

✉ Sebastian Schindler
sebastian.schindler@ukmuenster.de

Onno Kruse
onno.kruse@psychol.uni-giessen.de

Rudolf Stark
rudolf.stark@psychol.uni-giessen.de

Johanna Kissler
johanna.kissler@uni-bielefeld.de

² Department of Psychology, Bielefeld University, Postfach 100131, 33501 Bielefeld, Germany

³ Department of Psychotherapy and Systems Neuroscience, Justus Liebig University Giessen, Otto Behaghel-Str. 10H, 35394 Gießen, Germany

⁴ Department of Clinical Psychology, University of Siegen, Siegen, Germany

⁵ Center of Excellence “Cognitive Interaction Technology” (CITEC), Bielefeld University, Bielefeld, Germany

¹ Institute of Medical Psychology and Systems Neuroscience, University of Muenster, Von-Esmarch-Str. 52, D-48149 Muenster, Germany

computers (Lucas, Gratch, King, & Morency, 2014). For instance, participants show less interpersonal display and impression management when interacting with computers and mentalize less about computer interaction partners (Chaminade et al., 2012; Kircher et al., 2009; Lucas et al., 2014). In a similar vein, human gaze cues capture early neural attention resources and interfere with task performance. This is also found when avatar-gaze is believed to be controlled by a human rather than when it is believed to be fully robot-controlled (Wykowska, Wiese, Prosser, & Müller, 2014). Likewise, joint attention is initiated and responded to differently when an avatar face in a virtual reality environment is believed to be human-controlled (Caruana, Lissa, & McArthur, 2017a; Caruana, Spirou, & Brock, 2017c; Wykowska, Wiese, Prosser, & Müller, 2014).

In general, such findings can be explained within the framework of the Intentional Stance Model of Social Attention. This model specifies mechanisms by which intentionality ascription to humans, but not machines, can cause preferential orienting towards supposed human interaction partners (Wykowska et al., 2014). Humans expect interaction partners they perceive as human or human-controlled to assume an intentional stance similar to themselves and not available to non-human entities (see Caruana, McArthur, Woolgar, & Brock, 2017b; Dennett, 1989; Wykowska et al., 2014). This assumption guides perception and interactive behavior.

Verbally conveying interpersonal evaluations is an important aspect of human communication and affects social motivation, e.g., via interest in others' opinions and their integration in one's own self-view (Schlenker & Weigold, 1992; Shrauger & Schoeneman, 1979). This begs the question of how humans process interpersonal evaluation depending on its attributed source. In recent studies, event-related potential (ERP) responses to evaluative personality feedback have been examined. The feedback, which consisted of positive, negative, or neutral trait adjectives, was putatively either given by a human or a computer. "Human" feedback elicited considerably greater early and late ERP amplitudes across all experiments (Schindler & Kissler, 2016, 2018; Schindler, Wegrzyn, Steppacher, & Kissler, 2015), even when the computer was portrayed as a socially intelligent system, although the latter attenuated the effect of humanness (Schindler & Kissler, 2016). Furthermore, EEG source reconstructions estimated the "human-driven" increases to be most likely generated in the visual cortex, as well as in the superior frontal, orbitofrontal, somatosensory, and temporal brain areas (Schindler & Kissler, 2016, 2018; Schindler et al., 2015). However, activity from deeper cerebral structures, such as the cingulate cortex, the insulae, the amygdalae, or the striatum, all of which play a considerable role in social and affective processing and may therefore also respond to attributions of social presence, is typically not localized reliably by source estimation approaches.

Extant brain-imaging data, in line with the above-mentioned intentional stance model, also indicate that in complex interaction scenarios participants' brains overall respond less to actions

by computers than by humans: For unfair human behavior, stronger hemodynamic responses in the anterior cingulate cortex (ACC) and the anterior insula are found (Harlé, Chang, van 't Wout, & Sanfey, 2012; Singer, Kiebel, Winston, Dolan, & Frith, 2004), while fair human behavior elicits enhanced activity in the ventral striatum (Harlé et al., 2012; Phan, Sripada, Angstadt, & McCabe, 2010; Singer et al., 2004).

The ventral striatum is activated by various pleasant and rewarding stimuli, including sweet taste (Wyvell & Berridge, 2000), erotic pictures (Klucken et al., 2015; Sabatinelli, Bradley, Lang, Costa, & Versace, 2007), and positive monetary reward, as well as positive social feedback (Izuma, Saito, & Sadato, 2008, 2010; Kruse, Tapia León, Stark, & Klucken, 2017). In line with humans' generally strong social motives, merely interacting with other humans also seems to be rewarding, as stronger activity in the ventral striatum has been reported when participants interacted with a virtual character, whom they expected to be controlled by another human (Pfeiffer et al., 2014). These findings support the notion that this structure tracks a stimulus' general motivational salience value from low-level reinforcement all the way up to reward in complex social exchange (Báez-Mendoza & Schultz, 2013). Positive social feedback is also found to engage the ventral ACC and mPFC (Korn, Prehn, Park, Walter, & Heekeren, 2012; Somerville, Kelley, & Heatherton, 2010), underscoring these structures' role in social salience representation (Somerville et al., 2010). Interestingly, compensatory positive self-evaluations in response to social-evaluative threat have been found to recruit the mPFC and the orbitofrontal cortex, but also the amygdala and the insula (Hughes & Beer, 2012). The amygdala is thought to be involved in networks supporting both social aversion and social affiliation (Bickart, Dickerson, & Barrett, 2014). In general, amygdala responses have been observed to both positive and negative socio-emotional stimuli of various modalities and complexities (e.g., Cremers et al., 2010; Herbert et al., 2009; Kim et al., 2004; Klucken et al., 2015; Veer et al., 2011), in line with this structure's function as a salience detector (Sander, Grandjean, & Scherer, 2005). Regarding the insula, which is involved in many cognitive and emotional processes (Menon & Uddin, 2010; Nieuwenhuys, 2012), specifically its connections with the somato-sensory cortices are suggested to play a role in feeling empathy, and anterior insula responses probably code for the interoceptive component of social feedback (for reviews see Adolphs, 2009; Berridge & Kringelbach, 2015). Insula activity may even reflect embodied processing of abstract stimuli such as disgust- and pain-related words (Ponz et al., 2013; Richter, Eck, Straube, Miltner, & Weiss, 2010). Similarly, somatosensory responses, besides being activated by pain observation and empathy, are further modulated by affective significance (Gazzola et al., 2012), and are observed during emotion processing (Niedenthal, 2007) and as a correlate of the embodied processing of word meaning (i.e., the meaning of action-depicting verbs activate distinct somato-motor areas; Pulvermüller, 2005). Finally, the anterior insula seems to be

sensitive to intentionally delivered aversive stimuli (Liljeholm, Dunne, & O'Doherty, 2014).

When receiving social feedback from an unknown partner, attribution of intentional stance, self-reflection, and updating, as well as mentalizing about the sender are likely to take place, and this is also true when actual knowledge about the sender is limited (Heleven & Overwalle, 2018). Previous fMRI social neuroscience studies have revealed the medial prefrontal cortex (mPFC) and the posterior cingulate cortex (PCC) extending to the precuneus as critical nodes of the cortical midline structures in mentalizing tasks (Molenberghs, Johnson, Henry, & Mattingley, 2016; Northoff & Bermpohl, 2004; Northoff et al., 2006; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Uddin, Iacoboni, Lange, & Keenan, 2007). Self-reflection and awareness have been related to the PCC (Lieberman, 2007), while the mPFC was found to be involved in compensatory, positive self-evaluations as well as in self-updating processes (Hughes & Beer, 2012; Korn et al., 2012).

Based on the above findings, attributions of human social presence and intentional stance can be expected to affect stimulus processing in multiple ways, and may recruit an extensive network of brain regions, including mentalizing and self-reflection (mPFC and PCC), salience detection (e.g., amygdala), coding of interaction partners' intentionality, and embodied processing (e.g., insula and somatosensory cortex). Social presence may therefore imbue stimuli with motivational relevance. It can be created through instructions, where the mere notion of interacting with a “human” may change the salience of a stimulus. Such an approach allows for perfect experimental control, while still revealing core effects of social feedback processing. Importantly, it closely resembles modern virtual communication, where the identity of an interaction partner is typically inferred rather than seen (Biocca & Levy, 2013; Ferrara et al., 2016).

Using fMRI, the present study aimed to identify the cerebral correlates of evaluative feedback processing when feedback is assumed to be given by humans rather than being machine-generated, extending previous social neuroscience knowledge to the area of virtual interaction, which is increasingly relevant in everyday communication. To our knowledge, no study has addressed how attributed sender identity affects verbal feedback processing, as it might be encountered in typical exchanges of electronic text messages. Language-based feedback is arguably the most natural form of complex feedback exchanged among humans in the day and age of ubiquitous virtual interaction. Therefore, understanding its underlying mechanisms is gaining scientific and practical relevance. Moreover, its virtual nature allows for isolation of the role of sender attribution processes without physical confounds. As a starting point for fMRI studies using the present paradigm, we were most interested in the effect of attributed sender identity and virtual feedback content on the recipient's brain activity. This set-up can be described in terms of

Laswell's communication model (Laswell, 1948), which is one of the earliest and most influential models in this area and applies an analogous one-way perspective. It differentiates who (attributed sender) says what (feedback content) to whom using which channel with what effect (hemodynamic response), but does not consider explicit reciprocal communication from the receiver. It therefore provides a suitable model for the one-sided quasi-realistic communication that is the rule in current neuroimaging experiments (see also Eisenberger, Inagaki, Muscatell, Haltom, & Leary, 2011; Hughes & Beer, 2012; Korn et al., 2012), and enhances experimental control, but at the same time captures only parts of the interaction process, thereby restricting ecological validity (see also Caruana, McArthur, et al., 2017b and Wiese, Metta, & Wykowska, 2017 for design considerations in virtual interaction studies).

From an experimental design perspective, written language stimuli can be well controlled regarding many physical and cognitive parameters, and in the present scenario perfect counterbalancing of conditions is possible: In our study, participants were video-taped, describing themselves in a structured interview. They were further told that this video was a basis for another participant, who would evaluate them with positive, negative, or neutral trait adjectives as personality feedback. For the control condition, participants underwent the same videotaped structured interview but later were told to expect random computerized feedback similarly consisting of positive, negative, or neutral trait adjectives.

We focused on the effects of sender (human vs. computer) and content (positive, negative, neutral), and explored potential interactions. A whole-brain analysis tested for main effects and interactions between sender and emotion without a priori assumptions. A complementary region of interest (ROI) approach examined specifically the following assumptions: Regarding sender effects, for “human” feedback, we expected to find activation of the medial prefrontal cortex, orbitofrontal cortex, and the posterior cingulate cortex, reflecting enhanced mentalizing and the retrieval of self-related knowledge and self-evaluation. Further, enhanced amygdala activity was hypothesized for “human” feedback, reflecting its ability to track stimulus salience across a wide range of stimuli and tasks. Finally, for human feedback, insula activity might reflect ascribed intentionality, potentially resulting in different degrees of embodied feedback processing. Regarding main effects of emotional content, positive feedback should lead to enhanced activity in the striatum (e.g., Izuma et al., 2008, 2010), ventral ACC, and mPFC (Somerville et al., 2010), and in line with a salience detection account, both positive and negative feedback might activate the amygdala (Bickart et al., 2014; Sander et al., 2005), while negative feedback might additionally activate the insula (e.g., Harlé et al., 2012; Liljeholm et al., 2014; Singer et al., 2004).

Method

Participants

Thirty-two participants were recruited at the University of Giessen by responding to a mass mailing. They provided written informed consent according to the Declaration of Helsinki and received 15 Euros for participation. The study had been approved by the Bielefeld University Ethics Committee as part of a larger study program.

Four participants were excluded. Two were excluded due to abnormal brain anatomy, and two due to excessive movement artifacts (>2 mm or more than 5% outlying volumes). Of the remaining 28 subjects, two reported doubts about the cover story (see below) at the end of the experiment in an open questionnaire. Additionally, all participants were debriefed and interviewed orally again about the manipulation. After debriefing, ten participants reported to have been uncertain in the middle or at the end of the experiment about whether they really received feedback from the “human” sender, since they felt misjudged sometimes. All participants were included in the analysis, which can be regarded as a conservative approach, resulting in 28 participants (20 females) who were 24.19 years of age on average ($SD=2.43$), and all were right-handed, native German speakers who had normal or corrected-to normal vision. None reported a previous or current neurological or psychiatric disorder. Screening with Beck’s Depression Inventory (Beck, Steer, & Hautzinger, 2001) showed no sign of a clinically relevant depression ($M = 1.97$; $SD = 2.17$).

Stimuli

A pre-rated stimulus set used previously (e.g., see Schindler & Kissler, 2018) was reduced to 20 adjectives per condition. These adjectives had been rated by 22 students in terms of valence and arousal using the Self-Assessment Manikin (Bradley & Lang, 1994), and similarly in concreteness (how well an adjective describes a person). Raters were instructed to consider the adjectives’ valence and arousal in an interpersonal evaluative context. The selected 60 adjectives (20 negative, 20 neutral, 20 positive) were matched in their linguistic properties (see Table 1), such as word length, frequency, familiarity, regularity, and number of neighbors (i.e., words differing only in one letter, e.g., fearless and tearless).

Procedure

Participants were tested in a within-subjects design. They were told already at the first email contact that an appointment for two had to be made, as the experiment supposedly was about “formation of first impressions.” In the lab, they were told that they would be evaluated by an unknown other person

or by a randomly operating computer algorithm. Upon arrival, participants were instructed to briefly describe themselves in a structured interview in front of a camera (see the [Online Supplementary Material](#) for the questionnaires used). They were informed that the video of their self-description would be presented to the human judges to give them an impression of the participant. Finally, participants expected to judge the quality/descriptiveness of the evaluative feedback after scanning, which, however, never happened. For characterization of the sample, participants also completed a demographic questionnaire. To ensure face validity, while participants were prepared for the fMRI session, one of the investigators left the scanner room ahead of the fictitious feedback, delivering the video to the “unknown person” next door.

Stimuli were presented by software described as “Interactional Behavioral Systems” supposedly allowing instant online communication (see Fig. 1). To ensure credibility of the situation, the fictitious software desktop image that showed the “Interactional Behavioral Systems” environment changed its status during testing. The presented feedback was randomly generated in both conditions. Overall, by each putative sender (putative human outside the scanner or computer) 10 negative, 10 neutral, and 10 positive adjectives were endorsed as descriptive of the participant and 10 per condition were denied, leading to 10 affirmative negative, 10 affirmative neutral, and 10 affirmative positive decisions. Since it is questionable if the denial of a negative trait is something unambiguously positive or the denial of a positive trait clearly something negative, we only used affirmative decisions for analyses (on positive, neutral, and negative adjectives). Thus, the main effect of feedback emotional content consists of clearly positive (10 human decisions, 10 computer decisions), neutral (10 human decisions, 10 computer decisions), or negative feedback (10 human decisions, 10 computer decisions). For the sender main effect (human vs. computer) we therefore had 30 decisions for each sender, whereas for the main effect of feedback emotional content there were 20 trials per valence, and for the interaction, there were 10 trials per cell available. The desktop environment and stimulus presentation were created using Presentation (www.neurobehavioralsystems.com). In both conditions, color changes between 1,500 and 3,500 ms after adjective onset indicated the feedback by the putative interaction partner. We counterbalanced two colors (orange and purple) and two intensities (bright and dark) to present the feedback. In order to avoid confusing the participants, either purple or orange were used for the “human” and either the dark or the bright colors meant “affirmative” (or endorsed, i.e., true for the participant). An extensive demonstration (20 demo trials) of how the software worked was shown to the participants to learn color-feedback assignments. The onset of the color change was jittered randomly with an equal distribution and color changes lasted for 2,000 ms, followed by a fixation cross. Trial length varied between 10 and 12 s. After

Table 1 Comparisons of negative, neutral, and positive adjectives by one-way ANOVAs

Variable	Negative adjectives (n=20)	Neutral adjectives (n=20)	Positive adjectives (n=20)	F (2,57)
Valence	2.96 ^a (0.53)	4.94 ^b (0.30)	7.54 ^c (0.52)	495.36***
Arousal	4.59 ^a (0.55)	2.88 ^b (0.71)	4.70 ^a (0.47)	60.60***
Concreteness	3.10 ^a (0.61)	5.54 ^b (1.29)	2.59 ^a (1.04)	48.19***
Word length	9.35 (1.79)	9.65 (2.46)	9.60 (1.90)	0.12
Word frequency (per million)	531.85 (532.43)	504.15 (643.57)	492.45 (519.57)	0.03
Familiarity (absolute)	9892.65 (15586.27)	12513.55 (29954.91)	17157.15 (25833.41)	0.45
Regularity (absolute)	134.50 (247.34)	110.85 (201.11)	234.35 (339.52)	1.19
Neighbors Coltheart (absolute)	1.80 (1.47)	1.90 (1.17)	2.50 (2.44)	0.91
Neighbors Levenshtein (absolute)	3.55 (2.33)	4.50 (2.56)	5.00 (3.66)	1.28

*** $p \leq 0.001$

Standard deviations appear in parentheses below means; means in the same row sharing the same superscript letter do not differ significantly from one another at $p \leq 0.05$; means that do not share superscripts differ at $p \leq 0.05$ based on Fisher's least significant difference (LSD) test post hoc comparisons

the experiment, participants filled in a questionnaire asking them what the experiment was about. After debriefing, participants were again asked if they had any doubts and, if so, when these occurred.

fMRI recording and analyses

All MRI images were acquired using a 3 Tesla whole-body tomograph (Siemens Prisma) with a 64-channel head coil. The structural images consisted of 176 T1-weighted sagittal slices (slice thickness 0.9 mm; FOV = 240 mm; TR = 1.58 s; TE = 2.3 s). For the functional images, a total of 615 images was acquired. Images were acquired with a T2*-weighted gradient echo-planar imaging (EPI) with 36 slices covering the whole brain (voxel size = $3 \times 3 \times 3.5$ mm; gap = 0.5 mm; descending slice acquisition; TR = 2 s; TE = 30 ms; flip angle = 75; FOV = 192×192 mm; matrix size = 64×64 ; GRAPPA = 2). The field of view (FOV) was positioned automatically relative to

the AC-PC line with an orientation of -30° . Preprocessing, first-, and second-level analyses were done using Statistical Parametrical Mapping (SPM12, Wellcome Department of Cognitive Neurology, London, UK) implemented in Matlab 2014b (Mathworks Inc., Natick, MA, USA). The onsets of the color changes (i.e., the feedback decisions) were used for the analyses. For preprocessing, all EPI images were co-registered to an EPI template, realigned and unwarped, slice-time corrected, normalized to MNI standard space using parameters derived from segmentation of the structural images, and smoothed with a Gaussian Kernel at 6 mm FWHM (see the [Online Supplementary Material](#) for more detailed preprocessing information). Regressors on the first level were “positive word,” “negative word,” “neutral word,” “human positive word descriptive,” “human negative word descriptive,” “human neutral word descriptive,” “human positive word not descriptive,” “human negative word not descriptive,” “human neutral word not descriptive,” “computer

Fig. 1 Experimental design. Trial length varied between 10 and 12 s. Each adjective was presented first, while a color change indicated a yes/no decision of a putative human or computer sender

positive word descriptive,” “computer negative word descriptive,” “computer neutral word descriptive,” “computer positive word not descriptive,” “computer negative word not descriptive,” “computer neutral word not descriptive.” All regressors were stick functions modeled with 0-ms duration and convolved with the canonical hemodynamic response function. The six movement parameters were entered as covariates alongside regressors of no interest for the identified outlying volumes. As is standard for SPM analyses, a constant was also included in the first-level models. The time series was then high-pass filtered (time constant = 128 s). Betas for each feedback sender and emotion condition (six conditions in total) were calculated for each subject. On the group level, a flexible factorial design (two sender conditions $\times$ three emotion conditions) was used for the computed first level contrasts to examine main effects of feedback sender and feedback emotional content, as well as their interaction. First, voxel-wise whole brain analyses were performed (suggested to be more conservative against false-positive findings compared to clusterwise approaches, e.g., see Eklund, Nichols, & Knutsson, 2016) and are reported using a $p < .05$ family-wise error correction (FWE) based on the random field theory with an additional cluster extent threshold of $k > 5$ voxels. Secondly, region of interest (ROI) analyses were conducted using the small volume correction in SPM12 with $p < .05$ (FWE) with a cluster extent threshold of $k > 5$ voxels. Interactions between sender and feedback content did not survive FWE correction. For sender main effects, ROI masks for insula, amygdala, medial prefrontal cortex, orbitofrontal cortex, and anterior and posterior cingulate cortex were taken from the “Harvard-Oxford cortical and subcortical structural atlases” provided by the Harvard Center for Morphometric Analysis (25% probability threshold). For main effects of feedback emotional content, ROIs for the striatum, insula, amygdala, medial prefrontal cortex, and anterior cingulate cortex were used.

Results

Main effect of feedback sender

Whole brain analyses for the comparison between feedback from the “human” sender and computer sender showed broad activations in the frontal cortex and cortical midline structures (see Fig. 2, Table 2). Enhanced activations were found in the broad superior and inferior orbitofrontal and the posterior cingulate cortex. In addition, whole-brain FWE-corrected activations for the sender main effect were also observed in the bilateral insula and the right caudate.

Within the ROIs (Table 3 reports only significant clusters that did not already appear in the whole brain FWE-corrected analysis), enhanced activity was found in additional parts of

the PCC, in the medial prefrontal cortex, and the bilateral insular cortex (see Fig. 2, Table 3), while no differences in amygdala activation were observed between the two senders.

Main effect of feedback emotional content

For the main effect of feedback emotional content, FWE-corrected whole brain analyses showed enhanced activity for positive compared to both negative and neutral feedback in parts of the striatum, namely the anterior right caudate (see Fig. 3, Table 2). ROI analyses in the striatum additionally showed that bilateral caudate and nuclei accumbens as well as the bilateral putamen responded more strongly to positive feedback (see Fig. 3; Table 4 reports only clusters that are not already included in the whole brain analyses). Further, these ROI analyses showed that positive feedback led to stronger amygdala responses compared to negative (left amygdala) and neutral feedback (bilateral amygdalae). Additionally, compared to neutral feedback, positive feedback led to stronger activations in some parts of the mPFC. Negative feedback elicited stronger activations in FWE-corrected whole brain analyses in pre- and post-central regions compared to neutral but not to positive feedback (see Fig. 3, Table 2). ROI analyses showed that negative compared to neutral feedback elicited more activations in the left insular cortex and also in small areas of ACC.

Interactions of feedback sender and emotional content

Finally, whole brain FWE-corrected analyses did not reveal significant interactions between sender and feedback emotional content.

Discussion

This study investigated how the supposed presence of another person as a virtual interaction partner giving unidirectional personality feedback and positive, negative, and neutral feedback content affect hemodynamic correlates of verbal evaluative feedback processing in a design mimicking digital virtual interaction. Although on average identical random feedback was shown, supposedly “human-generated” feedback elicited stronger hemodynamic responses than “computer-generated” feedback in the superior frontal gyrus, the medial prefrontal cortex (mPFC), the inferior orbitofrontal cortex, the anterior and posterior cingulate cortex (ACC and PCC), the bilateral paracingulate gyri, and the bilateral insula. Moreover, whereas positive feedback activated the dorsal striatum, negative feedback activated somatosensory areas and the insula as evident from the whole brain and ROI analyses, respectively. In general, whole-brain and ROI analyses showed good convergence, with ROIs typically being identifiable in the whole brain

Fig. 2 Whole brain analyses for sender main effects. **(a)** Upper row: Enhanced FWE-corrected activity for the “human” sender is found in the bilateral inferior orbital cortex and insula cortex, the right caudate, and superior frontal regions, highlighted by red circles. The middle row highlights enhanced activations in the precuneus and the right caudate,

the lower row shows enhanced prefrontal cortex activations. **(b)** Mean beta weights for whole-brain FWE-corrected main effects of sender. Error bars represent ± 1 standard error of the mean. Laterality L/R, L left, R right, G gyrus, paracing. paracingulate

analysis at a threshold of $p < .001$, while some ROIs (particularly striatal structures) were also visible in the FWE analysis.

Mentalizing and integration of social feedback

The results reveal how by mere attribution of higher social relevance to a physically identical stimulus, an extensive cerebral network is recruited for stimulus evaluation and subsequent cognitive processing. Activity in the ACC is frequently reported during social exclusion and has been related to social pain (Eisenberger et al., 2011; Masten et al., 2009; Masten, Morelli, & Eisenberger, 2011; Rotge et al., 2014) and also found in response to pain-related negative words (Richter et al., 2010). However, since vACC and mPFC also respond to positive social feedback (Korn et al., 2012; Somerville et al., 2010), this might more likely reflect the enhanced representation of social salience for the “human” sender.

The found mPFC and PCC activations can be related to attribution of intentional stance (Wykowska et al., 2014) and mentalizing (Northoff & Bermpohl, 2004; Uddin et al., 2007), which in the present study suggests stronger reasoning about responses from the “human” sender, who is supposedly the more

relevant interaction partner (cf. Chaminade et al., 2012; Kircher et al., 2009). Changes in mPFC, ACC, and PCC activity due to positive, negative, or neutral verbal feedback have recently been interpreted as reflecting different degrees of self-referential processing (van Schie et al., 2018). Such reflective processes (Lieberman, 2007) may lead to an adaptation of self-ratings to peer evaluation, which, however, seem more likely for positive peer evaluation (Korn et al., 2012). Interestingly, Hughes and Beer (2012) report a similar pattern of activations during compensatory positive self-evaluations after negative social feedback, including activities in the mPFC, the orbitofrontal cortex, and the insula. Based on these findings, we speculate that participants may have not only reflected on and integrated the feedback with their self-concept, but also tried to maintain a positive self-view (for experiments on positive feedback expectation and self-enhancement, see Hepper, Hart, Gregg, & Sedikides, 2011). This may at least partly account for the overall stronger cerebral impact of positive feedback. What is more, extensive prefrontal activity has recently been reported when participants prepared to speak in a social context, suggesting that social context engages multiple frontal areas during both language comprehension and speech preparation (Kuhlen, Bogler, Brennan, & Haynes, 2017).

Table 2 FWE-corrected whole brain activations within main effects of feedback sender and feedback emotional content

Structure	K	MNI coordinates			$Z_{\max}$	P -FWE corrected
		x (mm)	y (mm)	z (mm)		
Harvard-Oxford Atlas	voxel					
Human feedback > computer feedback						
Sup Frontal G	443	0	44	44	6.32	<.001
L Paracingulate G	14	-6	40	-8	5.52	.002
R Caudate	42	8	10	-4	5.49	.003
L OFC	20	-30	16	-16	5.43	.003
L Precuneus	30	-6	-62	32	5.41	.004
R OFC	74	36	26	-14	5.34	.005
R Paracingulate G	15	2	52	18	5.20	.010
R Paracingulate G	14	6	48	10	5.20	.011
L OFC	16	-40	24	-14	5.13	.015
L PCC	6	-14	-46	36	5.08	.018
Positive feedback > negative feedback						
R Caudate	15	6	22	0	5.44	.003
Positive feedback > neutral feedback						
R Caudate	34	8	22	0	5.65	.001
R Caudate	15	24	-26	22	5.63	.001
Negative feedback > neutral feedback						
L Postcentral G	131	-20	-36	64	5.96	<.001
R Precentral G	56	8	-18	66	5.50	.002
R Precentral G	7	26	-28	54	5.18	.011

For each five-corrected significant peak, respective coordinates (x, y, and z) are displayed in MNI space

Laterality L/R, *L* left, *R* right, *Sup* superior, *G* gyrus

No significant activations were found for the computer > human, the negative > positive, the neutral > positive, and the neutral > negative comparisons

Ascribed intentionality of social feedback

We also observed anterior insula activations in response to putative human feedback, which are found in a variety of studies on social feedback processing (Eisenberger et al., 2011; Hughes & Beer, 2012; Kross, Berman, Mischel, Smith, & Wager, 2011; Liljeholm et al., 2014; Somerville et al., 2010). Intentionality of actions from an interactive partner is suggested to be decoded prior to mentalizing processes about the partner (Wykowska et al., 2014). This seems to affect brain responses as well, where human-controlled

avatars (i.e. supposedly another human participant controls the avatar's eyes) elicit differential ERP responses, but also lead to different behavioral responses to gaze cues (Caruana, Lissa, et al., 2017a; Caruana, Spirou, et al., 2017c; Wykowska et al., 2014). A study by Singer et al. (2004) showed that perceived intentionality activated the bilateral insula. More specifically, the anterior insula seems to code the interaction partner's intentionality when this partner can intentionally choose to deliver either aversive (salty tea) or non-aversive stimuli (Liljeholm et al., 2014). In the current study, sender effects in the anterior insula therefore might code the

Table 3 Region of interest (ROI) activations for main effect of sender, “human” feedback–“computer” feedback

Structure	K	MNI coordinates			$Z_{\max}$	P small volume corrected five
		x (mm)	y (mm)	z (mm)		
Harvard-Oxford Atlas	voxel					
mPFC	157	0	52	-20	4.17	.005
L insular cortex	78	-30	16	-14	5.08	<.001
R insular cortex	76	30	20	-8	5.03	<.001

For each FWE-corrected significant peak, respective coordinates (x, y, and z) are displayed in MNI space. An ROI can exhibit more than one cluster. In each cluster, only the largest peak is reported

Laterality L/R, *L* left, *R* right, *ACC* anterior cingulate cortex, *PCC* posterior cingulate cortex, *OFC* orbitofrontal cortex, *mPFC* medial prefrontal cortex

Fig. 3 Whole brain analyses for feedback main effects. **(a)** Positive compared to negative feedback led to enhanced activations in right caudate, the left putamen and the left amygdala, highlighted by red circles. **(b)** Positive compared to neutral feedback led to enhanced activations in bilateral striatum and bilateral amygdalae, as well as in post-central regions and the mPFC. **(c)** Negative feedback led to larger

pre- and post-central activations compared to neutral feedback. **(d)** Mean beta weights for whole-brain FWE-corrected main effects of emotional content in pre- and post-central areas and the right caudate, as well as for the whole striatum and the amygdala. Error bars represent ± 1 standard error of the mean. Laterality L/R, L left, R right

perception of intentionality, i.e., the deliberate affirmation of feedback by the human, which would suggest that adjacent regions in the anterior insula respond to intentionally delivered negative physical and to negative psychological outcomes, although, in principle, the insula contains functionally distinct networks, including ones separating physical and social pain (Cacioppo et al., 2013). In our study, we observe strong anterior insula activations for general “human” feedback, in line with the notion that only the human interaction partner can act intentionally. According to this interpretation, the anterior insula activation might be related to the interoceptive component of the feedback (see also below).

Reward and salience activations for positive feedback

Positive feedback elicited large activity in several structures, including the striatum (Izuma et al., 2008, 2010). Specifically, anterior parts of the caudate and the nucleus accumbens, which are part of the neuronal reward system (Berridge & Kringelbach,

2015; Klucken et al., 2015), were strongly activated. These findings are in line with those on positive or fair “human” behavior (Harlé et al., 2012; Phan et al., 2010; Singer et al., 2004), but also similar to those by Sabatinelli and colleagues, who reported strong responsiveness for pleasure but not for salience in the nucleus accumbens and medial prefrontal cortex (Sabatinelli et al., 2007). Accordingly, we also observe more activity in the medial prefrontal cortex for positive compared to neutral feedback, similar to reported activations by rewarding peer evaluation (Kom et al., 2012). Previous ERP studies also showed the most pronounced long latency emotion effects for positive feedback (Schindler & Kissler, 2016; Schindler et al., 2015), possibly reflecting a prolonged activation of multiple brain regions involved in the integration of positive feedback, including the amygdala, but also striatal and mPFC areas.

ROI analyses further revealed enhanced activity of the left amygdala for positive compared to negative feedback and bilaterally for positive compared to neutral feedback. Although often related to negative content (e.g., Eisenberger et al., 2011;

Table 4 ROI activations for positive, neutral, and negative feedback

Structure	K	MNI coordinates			Z _{max}	P small volume corrected five
		x (mm)	y (mm)	z (mm)		
Harvard-Oxford Atlas	voxel					
Positive feedback > negative feedback						
L Putamen	33	-20	16	-4	3.62	.030
R Caudate	38	10	22	0	4.59	.001
L Amygdala	6	-22	-2	-26	3.51	.019
Positive feedback > neutral feedback						
L Nuc Accumbens	64	-10	16	-8	4.32	<.001
L Caudate	7	-10	20	-2	3.67	.019
R Nuc Accumbens	45	10	20	-4	4.95	<.001
R Putamen	8	32	-14	-2	3.49	.046
R Caudate	60	10	22	0	5.35	<.001
R Caudate	15	18	10	20	3.89	.009
mPFC	94	2	54	-8	3.78	.022
L Amygdala	24	-20	-8	-16	3.48	.021
R Amygdala	55	22	0	-14	4.16	.002
Negative feedback > neutral feedback						
ACC	19	6	-14	46	4.01	.024
L Insular Cortex	31	-30	-26	16	4.16	.008

For each FWE-corrected significant peak, respective coordinates (x, y, and z) are displayed in MNI space

Laterality L/R, L left, R right, G gyrus, Nuc nucleus, mPFC medial prefrontal cortex

Kross et al., 2011; Somerville et al., 2010), the amygdala is thought to also be involved in networks supporting social affiliation (Bickart et al., 2014). Specifically, the left amygdala is known to also respond to positive word stimuli (e.g. Herbert et al., 2009), self-reference induction (Herbert, Herbert, & Pauli, 2011), cooperative behavior (Singer et al., 2004), and social affiliation (Bickart et al., 2014). In the present scenario, the relevance of positive feedback might have been higher, as participants tend to have optimistic expectations regarding receiving positive feedback (Hepper et al., 2011), as well as a general tendency to view themselves in a positive light (Watson, Dritschel, Obonsawin, & Jentsch, 2007).

Insula and somatosensory activations for negative feedback

Negative feedback activated the left somatosensory cortex and the left insula ROI more strongly than neutral feedback did. This is in agreement with a number of studies using different paradigms (e.g., Harlé et al., 2012; Kross et al., 2011; Masten et al., 2009) as well as previous source reconstructions revealing somatosensory activations in an EEG version of the present paradigm (Schindler & Kissler, 2016). Projections from the somatosensory cortex to the insula seem to be activated by embodied socio-emotional processing, e.g., feelings of caress by a touch (for reviews see Adolphs, 2009; Berridge & Kringelbach, 2015). The concurrent somatosensory and insula

activations for negative feedback might point to such an embodied processing of negative feedback. Previous fMRI (Richter et al., 2010) and intracranial (Ponz et al., 2013) studies report embodied processing of negative emotional words in the anterior insula. In particular, regions in the anterior insula might be necessary for emotional awareness, integrating stimulus-driven and top-down information (Gu, Hof, Friston, & Fan, 2013). Given the above-mentioned considerations that connections from the insula to the somato-sensory cortices play a role in the interoceptive component of social feedback (for reviews see Adolphs, 2009; Berridge & Kringelbach, 2015), and the concomitant activation of the insula and somatosensory cortex for negative feedback, we speculate that this might code an interoceptive aspect of negative feedback. Interoception and emotional awareness across various modalities might be a higher order function sub-served by the anterior insula. Still, it should be noted that a recent meta-analysis shows functionally distinct networks in this structure, including ones separating physical and social pain (Cacioppo et al., 2013).

Interactions between sender and feedback emotional content

Intuitively, statistical interactions between feedback source and feedback emotional content could have been expected. It would appear plausible for emotionally salient feedback to induce stronger BOLD responses when perceived as coming

from another intentionally acting human than when coming from a randomly acting computer. In fact, previous EEG studies had revealed interaction effects in that largest EPN (Schindler & Kissler, 2016, 2016) or P3 (Schindler & Kissler, 2018) amplitudes were found for "human" emotional (positive and negative) feedback. At a conservative threshold, this was not observed on a whole brain level in the present study, although descriptively emotional content effects appeared larger within the human sender. The limited number of trials realized in the presently used slow fMRI design, while limiting acquisition time, allowing for an optimized stimulus set, might have contributed to insufficient power to detect such theoretically plausible interactions. Further, tiredness might have lowered data quality, and the use of two colors and intensities might have increased difficulty compared to some of the previous feedback studies. Moreover, specific sender effects in such a randomized within-subjects design may be partially obscured by inertia effects in that participants might have trouble repeatedly and randomly switching mental contexts between senders while processing the feedback content. While the main effects may still be detectable, this added noise may obscure the detection of an interaction effect. In the present study, in line with Laswell's early communication model (Laswell, 1948), we placed emphasis on the effect of sender identity on the recipient in a within-participant design. Thereby, we took the risk of losing interaction effects. Interaction effects could be the focus of future studies, where a between-participants design might commend itself to realize adequate power in a reasonable acquisition time. Unfortunately, between-participant designs come with problems of their own.

Natural interaction versus quasi-interaction

As most other extant fMRI studies in this area, this "quasi-realistic" communicative scenario is one-sided, thus it is likely that socio-cognitive processes are more complex in "real" interactions. Although, in line with Laswell's communication model, our scenario invokes second-person engagement in that the participant has the impression of being directly addressed and is not a mere third-person observer, it currently lacks reciprocity, which has been suggested to be crucial in understanding natural social cognition (Schilbach et al., 2013). Clearly, reciprocal social interactions are more interesting, but also even more complex and difficult to perceptually control during neuroimaging (Caruana, McArthur, et al., 2017b; Hasson & Frith, 2016). In the future, virtual reality may commend itself as a way to facilitate more natural interactions in controllable environments (see also Caruana et al., 2017b; Wiese et al., 2017 for extensive discussions). Regarding the neural effects of one-way communication conceptually similar to the present one, synchronized brain activities among participants were found for watching the same

movie (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004), but note that participants were not scanned simultaneously. In most current "hyperscanning" studies, fMRI recordings have time-lags (Hasson & Frith, 2016; Hasson, Yang, Vallines, Heeger, & Rubin, 2008), and use video recordings of one interaction partner (Anders, Heinzle, Weiskopf, Ethofer, & Haynes, 2011) as real-time simultaneous interaction is very difficult to achieve and for fMRI would require the simultaneous availability of two scanners. Saito and colleagues (Saito et al., 2010), were in the lucky position to realize such simultaneous recording in a joint attention and gaze synchronization task. For real communication, the timescale also likely differs due to the coupling of various levels (e.g., perception, decoding, speech preparation) of abstract representations (Dumas, Guzman, Tognoli, & Kelso, 2014; Hasson & Frith, 2016). Although the understanding of real social interactions is our aim on the long run, here using a cover story to induce interactive expectations allowed perfect experimental control, whilst strongly affecting hemodynamic responses. Future studies should aim to record from two participants exchanging messages in restricted communicative settings to improve ecological validity. Currently, neuroscientific studies of simultaneous on-line communication are easier to implement in EEG or fNIRS, but at the cost of information about subcortical activities.

Role of individual differences

Some of the reported activities might be malleable by individual differences, such as have been found in the pattern of amygdala responses (Cremers et al., 2010; Laeger et al., 2012), and in the mPFC, ACC, PCC, and insula (van Schie et al., 2018). For instance, a stronger coupling between the amygdala and prefrontal regions was observed for participants with higher levels of neuroticism when presented with negative facial expressions (Cremers et al., 2010), as well as for participants with higher levels of trait anxiety when presented with negative words (Laeger et al., 2012). Moreover, in a recent study, participants, who ranged from having low to high self-esteem, received positive and negative feedback (van Schie et al., 2018). Enhanced activations for negative feedback in the PCC for participants with high self-esteem were observed. However, participants exhibiting low self-esteem showed decreased activations in the mPFC, PCC, ACC, and insula for positive feedback. This was interpreted by the authors to reflect a less pronounced self-referential processing.

In addition, Peterburs and colleagues (Peterburs, Sandrock, Miltner, & Straube, 2016), investigated hemodynamic responses to computer- or human-generated positive and negative performance feedback in preselected high- and low-anxious participants. Here, in the high social anxious group, hyperactivation in vmPFC/ventral ACC and insula was found for social relative to computer feedback, and in

mPFC/pregenual ACC for positive relative to negative feedback. Moreover, in the right rostral ACC high social anxious participants exhibited decreased activation for positive relative to negative feedback from a human. This indicates that a similar network of regions is activated in socially-relevant feedback processing across a variety of tasks, and that the specific activity pattern is, unsurprisingly, modulated by personality traits such as anxiety. In this regard, the present design may commend itself for future investigations of individual differences and clinical research.

Hemodynamic responses versus EEG source localization

Variants of this paradigm have already been employed in high-density EEG studies where source reconstructions have also been performed (Schindler & Kissler, 2016, 2018; Schindler et al., 2015). Although for more superficial cortical structures, there are similarities between previous source estimations and the current fMRI findings, previous source reconstructions and present fMRI findings also differ in some aspects. Unsurprisingly, fMRI revealed theoretically expected subcortical activities primarily in the striatum, deeper parts of the cingulate, and the amygdalae, which had not been captured by EEG source reconstruction. The superior frontal involvement could be validated, while previous source estimations also suggested strong fusiform/visual engagement, which would also be in line with Wykowska et al.'s (2014) Intentional Stance Model of social attention, and has been by us similarly interpreted as an index of socially motivated attention (Schindler & Kissler, 2016, 2018; Schindler et al., 2015). The faster presentation speed in the ERP studies might have contributed to a stronger engagement of the visual system. The differences might also be related to the rather transient ERP versus more tonic fMRI measures.

Conclusions

Summarizing the findings, attributed sender identity led to recruitment of an extensive brain network in response to evaluative feedback. Generally, a high context-induced social relevance strongly increased activities in brain regions consistent with mentalizing about the self and the sender and embodied processing. In particular, while validating some of the previous cortical source estimations, this study revealed the activation of deep brain structures, related to reward (striatum), intentionality, and emotional awareness/interoceptive feedback component (insula) and relevance (amygdala) processing in virtual interaction. Since virtual communication is becoming a reality for a majority of people and affects the ways we communicate with others (Biocca & Levy, 2013), our scenario was designed to share many characteristics with virtual communication channels widely used. It shows how easily attributions

can be implemented and how strong the resulting impact on the processing of physically identical stimuli is. Finally, it highlights that, when chatbots emulate human interactive behavior and adopt a particular human “sender identity” (Lee, 2016; Michael, 2016), we might not spot the difference. Thus, when chatbots are programmed to spread certain opinions under the guise of a “real personality,” our brains will boost these messages, increasing the likelihood that we will eventually be influenced by them.

Acknowledgements The authors declare that they have no conflicts of interest with respect to their authorship or the publication of this article. We thank all participants who contributed to this study. This work was supported by the Bielefelder Nachwuchsfond (Bielefeld University) and a Research Training Fellowship of the Society of Psychophysiological Research (SPR) awarded to SS.

References

- Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual Review of Psychology*, 60, 693.
- Anders, S., Heinze, J., Weiskopf, N., Ethofer, T., & Haynes, J.-D. (2011). Flow of affective information between communicating brains. *NeuroImage*, 54, 439–446.
- Báez-Mendoza, R., & Schultz, W. (2013). The role of the striatum in social behavior. *Frontiers in Neuroscience*, 7, 233.
- Beck, A. T., Steer, R. A., & Hautzinger, M. (2001). Beck-Depressions-Inventar (BDI): Testhandbuch (2., überarb. Aufl., 1. Nachdr). Bern: Huber.
- Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, 86(3), 646–664.
- Bickart, K. C., Dickerson, B. C., & Barrett, L. F. (2014). The amygdala as a hub in brain networks that support social life. *Neuropsychologia*, 63, 235–248.
- Biocca, F., & Levy, M. R. (2013). *Communication in the age of virtual reality*. Routledge.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *J Behav Ther Exp Psychiatry*, 25, 49–59.
- Cacioppo, S., Frum, C., Asp, E., Weiss, R. M., Lewis, J. W., & Cacioppo, J. T. (2013). A Quantitative Meta-Analysis of Functional Imaging Studies of Social Rejection. *Scientific Reports*, 3, 2027. <https://doi.org/10.1038/srep02027>
- Caruana, N., Lissa, P. de, & McArthur, G. (2017a). Beliefs about human agency influence the neural processing of gaze during joint attention. *Social Neuroscience*, 12(2), 194–206. <https://doi.org/10.1080/17470919.2016.1160953>
- Caruana, N., McArthur, G., Woolgar, A., & Brock, J. (2017b). Simulating social interactions for the experimental investigation of joint attention. *Neuroscience & Biobehavioral Reviews*, 74, 115–125. <https://doi.org/10.1016/j.neubiorev.2016.12.022>
- Caruana, N., Spirou, D., & Brock, J. (2017c). Human agency beliefs influence behaviour during virtual social interactions. *PeerJ*, 5, e3819. <https://doi.org/10.7717/peerj.3819>
- Chaminade, T., Rosset, D., Fonseca, D. D., Nazarian, B., Lutter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, 6.
- Cremers, H. R., Demenescu, L. R., Aleman, A., Renken, R., van Tol, M.-J., van der Wee, N. J. A., ... Roelofs, K. (2010). Neuroticism modulates amygdala—prefrontal connectivity in response to negative

- emotional facial expressions. *NeuroImage*, 49(1), 963–970. <https://doi.org/10.1016/j.neuroimage.2009.08.023>
- Dennett, D. C. (1989). *The intentional stance*. MIT press.
- Dumas, G., Guzman, G. C. de, Tognoli, E., & Kelso, J. A. S. (2014). The human dynamic clamp as a paradigm for social interaction. *Proceedings of the National Academy of Sciences*, 111(35), E3726–E3734. <https://doi.org/10.1073/pnas.1407486111>
- Eisenberger, N. I., Inagaki, T. K., Muscatell, K. A., Haltom, K. E. B., & Leary, M. R. (2011). The neural sociometer: Brain mechanisms underlying state self-esteem. *Journal of Cognitive Neuroscience*, 23, 3448–3455.
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905. <https://doi.org/10.1073/pnas.1602413113>
- Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104. <https://doi.org/10.1145/2818717>
- Gazzola, V., Spezio, M. L., Etzel, J. A., Castelli, F., Adolphs, R., & Keysers, C. (2012). Primary somatosensory cortex discriminates affective significance in social touch. *Proceedings of the National Academy of Sciences*, 109(25), E1657–E1666.
- Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, 521(15), 3371–3388. <https://doi.org/10.1002/cne.23368>
- Harlé, K. M., Chang, L. J., van 't Wout, M., & Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *NeuroImage*, 61, 32–40.
- Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Phil. Trans. R. Soc. B*, 371(1693), 20150366. <https://doi.org/10.1098/rstb.2015.0366>
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303(5664), 1634–1640. <https://doi.org/10.1126/science.1089506>
- Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, 28(10), 2539–2550. <https://doi.org/10.1523/JNEUROSCI.5487-07.2008>
- Heleven, E., & Overwalle, F. V. (2018). Neural representations of others in the medial prefrontal cortex do not depend on our knowledge about them. *Social Neuroscience*, 0(0), 1–14. <https://doi.org/10.1080/17470919.2018.1472139>
- Hepper, E. G., Hart, C. M., Gregg, A. P., & Sedikides, C. (2011). Motivated expectations of positive feedback in social interactions. *The Journal of Social Psychology*, 151, 455–477.
- Herbert, C., Ethofer, T., Anders, S., Junghofer, M., Wildgruber, D., Grodd, W., & Kissler, J. (2009). Amygdala activation during reading of emotional adjectives—An advantage for pleasant content. *Social Cognitive and Affective Neuroscience*, 4, 35–49.
- Herbert, C., Herbert, B. M., & Pauli, P. (2011). Emotional self-reference: Brain structures involved in the processing of words describing one's own emotions. *Neuropsychologia*, 49, 2947–2956.
- Hughes, B. L., & Beer, J. S. (2012). Protecting the Self: The Effect of Social-evaluative Threat on Neural Representations of Self. *Journal of Cognitive Neuroscience*, 25(4), 613–622. https://doi.org/10.1162/jocn_a_00343
- Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, 58, 284–94. <https://doi.org/10.1016/j.neuron.2008.03.020>
- Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, 22, 621–631.
- Kim, H., Somerville, L. H., Johnstone, T., Polis, S., Alexander, A., Shin, L. M., & Whalen, P. J. (2004). Contextual modulation of amygdala responsivity to surprised faces. *Journal of Cognitive Neuroscience*, 16(10), 1730–1745.
- Kircher, T., Blumel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., ... Krach, S. (2009). Online mentalising investigated with functional MRI. *Neuroscience Letters*, 454, 176–81. <https://doi.org/10.1016/j.neulet.2009.03.026>
- Klucken, T., Kruse, O., Wehrum-Osinsky, S., Hennig, J., Schweckendiek, J., & Stark, R. (2015). Impact of COMT Val158Met-polymorphism on appetitive conditioning and amygdala/prefrontal effective connectivity. *Human Brain Mapping*, 36(3), 1093–1101. <https://doi.org/10.1002/hbm.22688>
- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience*, 32, 16832–16844.
- Kross, E., Berman, M. G., Mischel, W., Smith, E. E., & Wager, T. D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences*, 108(15), 6270–6275.
- Kruse, O., Tapia León, I., Stark, R., & Klucken, T. (2017). Neural correlates of appetitive extinction in humans. *Social Cognitive and Affective Neuroscience*, 12(1), 106–115. <https://doi.org/10.1093/scan/nsw157>
- Kuhlen, A. K., Bogler, C., Brennan, S. E., & Haynes, J.-D. (2017). Brains in dialogue: decoding neural preparation of speaking to a conversational partner. *Social Cognitive and Affective Neuroscience*, 12(6), 871–880. <https://doi.org/10.1093/scan/nsw018>
- Laeger, I., Dobel, C., Dannlowski, U., Kugel, H., Grotegerd, D., Kissler, J., ... Zwaninger, P. (2012). Amygdala responsiveness to emotional words is modulated by subclinical anxiety and depression. *Behavioural Brain Research*, 233(2), 508–516.
- Lasswell, H. D. (1948). The structure and function of communication in society. *The Communication of Ideas*, 37, 215–228.
- Lee, P. (2016). Learning from Tay's introduction. Retrieved July 15, 2016, from <http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>
- Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, 58, 259–89.
- Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2014). Anterior Insula Activity Reflects the Effects of Intentionality on the Anticipation of Aversive Stimulation. *Journal of Neuroscience*, 34(34), 11339–11348. <https://doi.org/10.1523/JNEUROSCI.1126-14.2014>
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100.
- Masten, C. L., Eisenberger, N. I., Borofsky, L. A., Pfeifer, J. H., McNealy, K., Mazziotta, J. C., & Dapretto, M. (2009). Neural correlates of social exclusion during adolescence: Understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience*, 4, 143–157.
- Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for “social pain” and subsequent prosocial behavior. *NeuroImage*, 55, 381–388.
- Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, 214(5–6), 655–667. <https://doi.org/10.1007/s00429-010-0262-0>
- Michael, K. (2016). Science Fiction Is Full of Bots That Hurt People: ...But these bots are here now. *IEEE Consumer Electronics Magazine*, 5(4), 112–117. <https://doi.org/10.1109/MCE.2016.2590218>
- Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 65, 276–291. <https://doi.org/10.1016/j.neubiorev.2016.03.020>

- Niedenthal, P. M. (2007). Embodying emotion. *Science*, 316(5827), 1002–1005. <https://doi.org/10.1126/science.1136930>
- Nieuwenhuys, R. (2012). The insular cortex: a review. In *Progress in brain research* (Vol. 195, pp. 123–163). Elsevier.
- Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences*, 8, 102–7.
- 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(1), 440–457. <https://doi.org/10.1016/j.neuroimage.2005.12.002>
- Peterburs, J., Sandrock, C., Miltner, W. H. R., & Straube, T. (2016). Look who's judging—Feedback source modulates brain activation to performance feedback in social anxiety. *NeuroImage*. <https://doi.org/10.1016/j.neuroimage.2016.03.036>
- Pfeiffer, U. J., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A. L., Bente, G., & Vogeley, K. (2014). Why we interact: On the functional role of the striatum in the subjective experience of social interaction. *NeuroImage*, 101, 124–137. <https://doi.org/10.1016/j.neuroimage.2014.06.061>
- Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences*, 107, 13099–13104.
- Ponz, A., Montant, M., Liegeois-Chauvel, C., Silva, C., Braun, M., Jacobs, A. M., & Ziegler, J. C. (2013). Emotion processing in words: a test of the neural re-use hypothesis using surface and intracranial EEG. *Social Cognitive and Affective Neuroscience*, nst034.
- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews. Neuroscience*, 6(7), 576–582. <https://doi.org/10.1038/nrn1706>
- Richter, M., Eck, J., Straube, T., Miltner, W. H., & Weiss, T. (2010). Do words hurt? Brain activation during the processing of pain-related words. *Pain*, 148(2), 198–205.
- Rotge, J.-Y., Lemogne, C., Hinfrey, S., Huguet, P., Grynszpan, O., Tartour, E., ... Fossati, P. (2014). A meta-analysis of the anterior cingulate contribution to social pain. *Social Cognitive and Affective Neuroscience*, nsu110.
- Sabatinelli, D., Bradley, M. M., Lang, P. J., Costa, V. D., & Versace, F. (2007). Pleasure Rather Than Salience Activates Human Nucleus Accumbens and Medial Prefrontal Cortex. *Journal of Neurophysiology*, 98(3), 1374–1379. <https://doi.org/10.1152/jn.00230.2007>
- Saito, D. N., Tanabe, H. C., Izuma, K., Hayashi, M. J., Morito, Y., Komeda, H., ... Sadato, N. (2010). “Stay Tuned”: Inter-individual neural synchronization during mutual gaze and joint attention. *Frontiers in Integrative Neuroscience*, 4. <https://doi.org/10.3389/fnint.2010.00127>
- Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4), 317–352. <https://doi.org/10.1016/j.neunet.2005.03.001>
- van Schie, C. C., Chiu, C.-D., Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2018). When compliments do not hit but critiques do: an fMRI study into self-esteem and self-knowledge in processing social feedback. *Social Cognitive and Affective Neuroscience*, 13(4), 404–417. <https://doi.org/10.1093/scan/nsy014>
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36(04), 393–414.
- Schindler, S., & Kissler, J. (2016). People matter: Perceived sender identity modulates cerebral processing of socio-emotional language feedback. *NeuroImage*, 134, 160–169. <https://doi.org/10.1016/j.neuroimage.2016.03.052>
- Schindler, S., & Kissler, J. (2018). Language-based social feedback processing with randomized “senders”: An ERP study. *Social Neuroscience*, 13(2), 202–213. <https://doi.org/10.1080/17470919.2017.1285249>
- Schindler, S., Wegrzyn, M., Steppacher, I., & Kissler, J. (2015). Perceived Communicative Context and Emotional Content Amplify Visual Word Processing in the Fusiform Gyrus. *The Journal of Neuroscience*, 35(15), 6010–6019. <https://doi.org/10.1523/JNEUROSCI.3346-14.2015>
- Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal Processes Involving Impression Regulation and Management. *Annual Review of Psychology*, 43(1), 133–168. <https://doi.org/10.1146/annurev.ps.43.020192.001025>
- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42, 9–34. <https://doi.org/10.1016/j.neubiorev.2014.01.009>
- Straube, T., & Schenkenberg, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin*, 86(3), 549.
- Singer, T., Kiehl, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron*, 41, 653–662.
- Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex*, 20, 3005–3013.
- Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4), 153–157.
- Veer, I. M., Oei, N. Y. L., Spinhoven, P., van Buchem, M. A., Elzinga, B. M., & Rombouts, S. A. R. B. (2011). Beyond acute social stress: Increased functional connectivity between amygdala and cortical midline structures. *NeuroImage*, 57(4), 1534–1541. <https://doi.org/10.1016/j.neuroimage.2011.05.074>
- Watson, L. A., Dritschel, B., Obonsawin, M. C., & Jentsch, I. (2007). Seeing yourself in a positive light: brain correlates of the self-positivity bias. *Brain Research*, 1152, 106–10.
- Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01663>
- Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the Minds of Others Influence How We Process Sensory Information. *PLOS ONE*, 9(4), e94339. <https://doi.org/10.1371/journal.pone.0094339>
- Wyvell, C. L., & Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward “wanting” without enhanced “liking” or response reinforcement. *Journal of Neuroscience*, 20(21), 8122–8130.