Social Cognitive and Affective Neuroscience Advance Access published July 4, 2012

doi:10.1093/scan/nss064 SCAN (2012) 1 of 8

Electrophysiological precursors of social conformity

Anna Shestakova,[1,2] Jörg Rieskamp,[3] Sergey Tugin,[1] Alexey Ossadtchi,[1,4] Janina Krutitskaya,[1] and Vasily Klucharev[3]

[1]Department of Higher Nervous Activity and Psychophysiology, Saint Petersburg State University, Saint Petersburg, Russia, [2]Centre for Magnetoencephalography, The Moscow State University of Psychology and Education, Moscow, Russia, [3]Department of Psychology, University of Basel, Basel, Switzerland, and [4]Institute for Problems of Mechanical Engineering, Russian Academy of Sciences, Saint Petersburg, Russia

Humans often change their beliefs or behavior due to the behavior or opinions of others. This study explored, with the use of human event-related potentials (ERPs), whether social conformity is based on a general performance-monitoring mechanism. We tested the hypothesis that conflicts with a normative group opinion evoke a feedback-related negativity (FRN) often associated with performance monitoring and subsequent adjustment of behavior. The experimental results show that individual judgments of facial attractiveness were adjusted in line with a normative group opinion. A mismatch between individual and group opinions triggered a frontocentral negative deflection with the maximum at 200 ms, similar to FRN. Overall, a conflict with a normative group opinion triggered a cascade of neuronal responses: from an earlier FRN response reflecting a conflict with the normative opinion to a later ERP component (peaking at 380 ms) reflecting a conforming behavioral adjustment. These results add to the growing literature on neuronal mechanisms of social influence by disentangling the conflict-monitoring signal in response to the perceived violation of social norms and the neural signal of a conforming behavioral adjustment.

Keywords: conformity; social influence; feedback-related negativity (FRN); medial frontal cortex; reinforcement learning

INTRODUCTION

People's decisions are often guided by social norms and the behavior of others (Ajzen and Fishbein, 1980; Cialdini and Goldstein, 2004). Recent neuroimaging studies have begun to uncover the neural mechanisms of various forms of social influence (Berns *et al.*, 2005; Behrens *et al.*, 2008; Klucharev *et al.*, 2008; Klucharev *et al.*, 2009; Berns *et al.*, 2010; Burke *et al.*, 2010a; Campbell-Meiklejohn *et al.*, 2010; Biele *et al.*, 2011; Klucharev *et al.*, 2011). In this study, we further explored the neuronal mechanisms of conformity, that is the act of changing one's behavior to match the behavior or opinions of other people (Cialdini and Goldstein, 2004).

Recent neuroimaging studies have suggested that conformity and other forms of social influence involve the activity of reward- and performance-monitoring neural circuitry (Klucharev *et al.*, 2009; Burke *et al.*, 2010b; Campbell-Meiklejohn *et al.*, 2010). Klucharev *et al.* (2009), for instance, demonstrated that conformity is associated with a neuronal response in the posterior medial frontal cortex and the ventral striatum areas known to be involved in reward monitoring, reinforcement learning and the evaluation of behavioral outcomes. Other functional magnetic resonance imaging (fMRI) studies showed that activity of the posterior medial frontal cortex reflects individuals' tendencies to change their opinion in the presence of others' opinions (Berns *et al.*, 2010; Campbell-Meiklejohn *et al.*, 2010) or others' advice (Behrens *et al.*, 2008). Interestingly, the posterior medial frontal cortex is also involved in cognitive dissonance—an important cognitive mechanism underlying social influence (van Veen *et al.*, 2009; Izuma *et al.*, 2010). Overall, there is a growing support for the hypothesis that the reward- and performance-monitoring neural circuitry (including the posterior medial frontal cortex) is involved in various forms of social influence.

Previous fMRI and event-related potential (ERP) studies suggested that the posterior medial frontal cortex has a specific role in performance monitoring. Activity of the posterior medial frontal cortex reflects a need for behavioral adjustments when the goal of an action was not achieved (Kerns *et al.*, 2004; Ridderinkhof *et al.*, 2004; Brown and Braver, 2005; Cohen and Ranganath, 2007; di Pellegrino *et al.*, 2007). Importantly, the magnitude of the activity of the posterior medial frontal cortex has also been shown to predict the strength of subsequent behavioral adjustments during simple choice decisions (O'Doherty *et al.*, 2003; Kerns *et al.*, 2004; Cohen and Ranganath, 2007). The reinforcement learning theory of performance monitoring suggests that medial frontal cortex activity indicates whether an action outcome is worse or better than expected (Holroyd and Coles, 2002). A 'prediction error' signal at the medial frontal cortex can be measured as a negative ERP on the scalp that has been called feedback-related negativity (FRN; see, e.g. Miltner *et al.*, 1997; Cohen and Ranganath, 2007). The FRN amplitude tends to correlate strongly with a negative prediction error and only marginally with a positive prediction error (Chase *et al.*, 2011). In general, FRN is a negative shift in the ERP occurring 200–400 ms after receiving negative performance feedback (Miltner *et al.*, 1997). FRN shares a functional and spatial relationship with ERN (error related negativity)—a negative ERP associated with error processing after the commission of an incorrect response in forced choice reaction time tasks (e.g. Gentsch *et al.*, 2009). Both source localization and fMRI studies have confirmed that FRN/ERN is generated in the posterior medial frontal cortex (rostral cingulate zone; Gehring and Willoughby, 2002; Holroyd *et al.*, 2004; van den Bos *et al.*, 2009; Roger *et al.*, 2010). Interestingly, the same area is also involved in conformity and general behavioral adjustments (e.g. Ridderinkhof *et al.*, 2004; Klucharev *et al.*, 2009).

Here, we studied how individual judgments of facial attractiveness are modulated by the group opinion. Past research on social influence has shown that people systematically change behavior and opinions in line with the normative opinion of a group to receive the group's approval and support (Cialdini and Goldstein, 2004). Thus, according to the *social influence hypothesis*, people should on average show a tendency to adjust subjective judgments of facial attractiveness when

Received 29 November 2011; Accepted 26 May 2012

This study was supported by the Switzerland–Russian Scientific & Technological Cooperation Program, the Russian Targeted Federal Program 'Scientific and scientific-pedagogical personnel of innovative Russia' (contracts 02.740.11.5233, 14.740.11.0232, and 02.740.11.5148), by a grant from the Russian Foundation for Basic Research (RFBR 11-06-00449-a) to Anna Shestakova and by a grant from the Swiss National Science Foundation (#100014_130352) to Vasily Klucharev and Jörg Rieskamp. We thank Daniel Kislyuk and Natalia Shemyakina for their recommendations and expertise in task programming, subject recruitment, and data collection. We also thank the referees for their valuable and instructive comments.

Correspondence should be addressed to Anna Shestakova, Department of Higher Nervous Activity and Psychophysiology, Saint Petersburg State University, 199034, University Emb.7-9, Saint Petersburg, Russia. E-mail: shestako@mappi.helsinki.fi

their judgments do not match the normative group opinion. Furthermore, as we have described above, we argue that social influence could be based on a general performance-monitoring mechanism. Accordingly, it can be hypothesized that when a person's behavior does not match others' behavior, this should be perceived as negative feedback with a similar neural response (i.e. FRN) to the response that can be observed for an individual learning problem. Thus, according to the *learning hypothesis of social influence*, observed conformity behavior should involve activity of the posterior medial frontal cortex that generates an FRN signal. Consequently, activity of the posterior medial frontal cortex should also predict the subsequent conforming adjustment of the behavior.

To test these hypotheses, we used a paradigm in which a person's initial judgments, that is attractiveness ratings of faces, were open to the social influence of the opinion of a group (Klucharev *et al.*, 2009; Zaki *et al.*, 2011). Female participants rated the attractiveness of female faces and after each rating they were informed about an 'average group rating' of the face. Actual group ratings were systematically manipulated during the experiment. We assumed that group opinion (group ratings) signaled descriptive group norms of facial attractiveness. With this procedure, we introduced a conflict between a person's own judgment and a group opinion. To detect subsequent conformity with the group, participants rated the same set of faces again but without the normative (group) ratings.

First, to identify the neural activity related to 'social (normative) conflict' we compared the evoked responses calculated over trials in which the group rating differed from the participant's rating (conflict trials) with all no-conflict trials. Second, to model subsequent conformity effects we separately averaged conflict trials followed by conformity (i.e. where perceived facial attractiveness subsequently changed in line with the group rating) and conflict trials not followed by conformity (where perceived facial attractiveness did not change). Overall, the excellent time resolution of the ERP method used allowed us to investigate for the first time the temporal overlap of ERPs to normative conflicts of opinion and ERPs indicating conforming adjustments of the judgment of facial attractiveness.

MATERIALS AND METHODS

The study was approved by the local ethics review committee. Prior to the start of the experiment, each participant gave her informed consent in writing.

Participants

Sixteen young Russian right-handed female students (aged 17–26 years, mean 19.9 years) were recruited for a small compensation (equivalent of 10 US dollars). They participated in two experimental sessions: an ERP session and a behavioral session separated by ~15 min. None of the subjects reported a history of drug abuse, head trauma or neurological or psychiatric illness. The data of one participant were discarded from the group analysis due to excessive electroencephalogram (EEG) artifacts.

Stimuli

A set of 222 digital photos of Caucasian females (aged 18–35 years, from free Internet sources) were used as stimuli. Color portraits of moderately attractive [mean = 4.2, standard deviation (SD) = 1.2 on an eight-point scale] females and moderate smiles were selected, all of a highly similar photographic style and appearance. We selected only female portraits to be presented to the female participants because cross-gender rating of attractiveness is related to mate selection, which has very specific neural mechanisms (Cloutier *et al.*, 2008). In contrast, within-gender ratings of attractiveness can be generalized to other types of conforming behavior.

Experimental procedure

Each experimental session started with the experimenter informing the participants about the experimental procedure. Participants were told that they were participating in a project entitled 'Seeing Beauty' to study human perception of attractiveness. During an EEG session (details described below), participants were exposed to a series of 222 photographs of female faces (stimuli duration = 2 s, inter-trial interval = 2.5–3.0 s, overall duration of the session = 38 min).

Participants were instructed to rate each face on an eight-point scale ranging from 1 (*very unattractive*) to 8 (*very attractive)*; for details see also Klucharev *et al.*, 2009. Participants indicated their rating by pressing the appropriate button. The participant's rating (initial rating, green rectangular frame) was visualized on the screen immediately after the face stimulus. At the end of each trial, the participant was informed (with a blue rectangular frame) about the average rating of the same face given by a large group of students from the same Russian university (group rating). The difference between the participant's and the group rating was additionally indicated by a score above the scale (0, ±2 or ±3 points). Importantly, the frame and the number indicating the deviance from the group opinion appeared for both 'conflict' and 'no-conflict' trials.

Actual group ratings were programmed using the following criteria: in 33% of the trials the group ratings agreed with the participant's ratings, whereas in 67% of the trials the group ratings were pseudo-randomly above or below the participant's ratings by ±2 or ±3 points. This was performed using an adaptive algorithm that kept the overall ratio of 'more negative' or 'more positive' group ratings approximately equal during the experiment for every participant. We informed participants that group ratings that matched their own rating within the range of ±1 point would produce a frame of the group rating that would visually overlap with the frame of the participant's own rating. Participants were not informed about the real purpose of the experiment or the manipulation of the group ratings. All participants were debriefed after the experiment. All photographs were randomized across participants and conditions. They were presented on a 14 inch computer monitor at a distance of ~60 cm from the participant's face. Fifteen minutes after the ERP session in an unannounced subsequent behavioral session, participants were instructed to again rate (self-paced) the attractiveness of the same faces presented in a new randomized order without the normative ratings (subsequent rating).

Our experimental design follows social psychological studies investigating persuasion, where participants are informed about a dominant behavior in a group (Cialdini, 2007). In this study, we investigated descriptive social norms sending the message 'If a lot of people are doing this, it's probably a wise thing to do'. Importantly, attractiveness is a socially important facial feature (Langlois *et al.*, 2000); judgments of facial attractiveness are fast, effortless and consistent across people (Willis and Todorov, 2006). Therefore, a mismatch between individual judgments of facial attractiveness and group opinion should create a strong normative conflict. Despite the formal structure, our task has a social nature, as demonstrated by previous studies (Klucharev *et al.*, 2009).

At the beginning of the experiment, the participants were asked to fill out the Edinburgh Handedness Inventory (Oldfield, 1971) and the Russian version of Spielberger's State–Trait Anxiety Inventory to assess handedness and the level of anxiety, respectively (Spielberger *et al.*, 1970). Previous studies demonstrated that FRN is modulated by individual level of anxiety (Hajcak *et al.*, 2003; Gu *et al.*, 2010a,

2010b). However, the multiple regression analysis of the anxiety score and the magnitude of ERPs obtained in our study revealed a non-linear relationship ($R^2 \leq 0.1$). Therefore, we did not use the state- and trait-anxiety scores obtained in the surveys as covariates in the statistical analysis of ERP data. Prior to the EEG session, we asked participants to sit comfortably in the experimental chair so as to limit their movements in order to reduce possible artifacts. They were also instructed to blink as little as possible.

Analysis of behavioral results

To detect conformal behavioral adjustments, we analyzed changes of ratings between the two sessions: the mean differences between the second and the first ratings were calculated separately for conflict and no-conflict trials. The effect of group opinion on conformal adjustments was analyzed using a one-way analysis of variance (ANOVA) with *changes in attractiveness ratings* as the dependent variable and the three-level within-subject factor *group rating* (more positive, more negative and consistent group rating). In addition, the probability of conformal changes in each condition was calculated. Both mean size and probability of conformal behavioral adjustments were submitted to two-way ANOVA controlling for the sign of the change of attractiveness rating with respect to the sign of the conflict: positive *vs* negative conflicts and small *vs* large conflicts. To study the effect of the stimulus ambiguity on social conformity, we selected faces with low and high variance of the initial ratings as unambiguous and ambiguous faces, respectively (see Results for details). To study the effect of ambiguity, we used a two-way ANOVA with *ambiguity* (ambiguous *vs* less ambiguous) and *group ratings* as two within-subject factors. The data were analyzed using the software STATISTICA (StatSoft, Inc.).

ERP recording and analysis

EEG data were recorded at 250 Hz from 19 Nikolet gold-cup scalp electrodes and two ocular electrodes (one in the corner of the eye and another above the right eye) using Mitsar Medical Diagnostic Equipment, EEG-201. EEG electrodes were on-line referenced to the average of all scalp electrodes and later off-line referenced to the average of the two mastoids. Scalp channels including Fp1, Fpz, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, P3, Pz, P4, O1, Oz and O2 were set according to the 10–20 system. Two referent electrodes were set over the mastoids. Data were recorded with a band-pass filter (0.1–70 Hz) and later refiltered with the filter at 0.5–20 Hz. Electrode resistance was kept below 10 kΩ.

Trials containing blinks or other artifacts or having voltage amplitudes greater than $\pm 100\,\mu$V were discarded from averaging (mean number of discarded trials = 55.4, SD = 5.3). Prior to averaging, the EEG data were spatially filtered in order to remove or minimize ocular artifacts (http://www.sourcesignal.com). The artifacts were manually separated (segmented) from the clean (artifact-free) data. Once artifacts were identified, the filter subtracted artifacts from all channels where it was detected (e.g. see Tremblay *et al.*, 2008; West *et al.*, 2011 for the same preprocessing routine). Overall, the approach is based on a spatial filter (including all EEG channels and optional electrooculogram (EOG) channels) that projects the data into the orthogonal complement of an identified artifact subspace after spatially whitening the data with respect to the covariance statistics of artifact-free EEG. This approach is known to minimally disturb clean EEG recordings. Correction rank did not exceed 2. EEG preprocessing, artifact removal and ERP analysis were performed with the EMSE Software suite by Source Signal Imaging, Inc. (San Diego, CA, USA).

Statistical analyses were performed by entering individually averaged ERPs from predefined latency windows as the dependent variable into two repeated-measures ANOVAs. The first ANOVA had the two main within-subject factors of *conflict* (conflict trials *vs* no-conflict trials) and *electrode* (19 electrode loci) in the 190 to 230 ms window. The second ANOVA had the two within-subject factors of *conformity* (conflict trials followed by conformity *vs* conflict trials not followed by conformity) and *electrode* (19 electrode loci) in the 300 to 380 ms window. The Greenhouse–Geisser (G–G) correction was applied to compensate for the lack of homogeneity in the repeated-measure variance.

The peaks were chosen from the Fz electrode, where the ERP responses indicating both social conflict and conformity effects were maximal. The frontocentral distribution of the components of interest can be seen on the topographical maps. ERPs were averaged across the 40 ms (in the case of a broader ERP for the effect of conflict) and 20 ms (narrower ERP response to the conformity effect) time windows because average amplitude measures are believed to be less sensitive to noise and therefore provide more reliable measures.

RESULTS

Behavioral results

Overall, participants rated faces as moderately attractive (first session: mean attractiveness = 4.5, SD = 1.9; second session: mean attractiveness = 4.4, SD = 1.7). In line with the social influence hypothesis, participants changed their ratings of attractiveness to align themselves with the group ratings (Figure 1). On average, participants decreased their attractiveness ratings when the group ratings were more negative than their own initial rating, whereas they increased their attractiveness ratings when the group ratings were more positive than their own initial rating (see Table 1 for details). A one-way ANOVA with changes in attractiveness ratings as the dependent variable and the three-level within-subject factor group rating revealed that the observed changes correspond to a significant main effect, $F(2,14) = 72.01$, $P < 0.0001$, $\eta^2 = 0.83$. Therefore, group opinion effectively modulated individuals' judgments of attractiveness. The conformity effect was moderately stronger for large conflicts with the group opinion (Figure 1). A two-way ANOVA (positive/negative conflicts and small/large conflicts) revealed a main effect of the factor conflict size, $F(1,14) = 9.66$, $P < 0.001$, $\eta^2 = 0.07$. The effect of the conflict direction (positive/negative) was not significant: $F(1,14) = 0.03$, $P = 0.85$. In sum, our study revealed a strong conformity effect according to which the

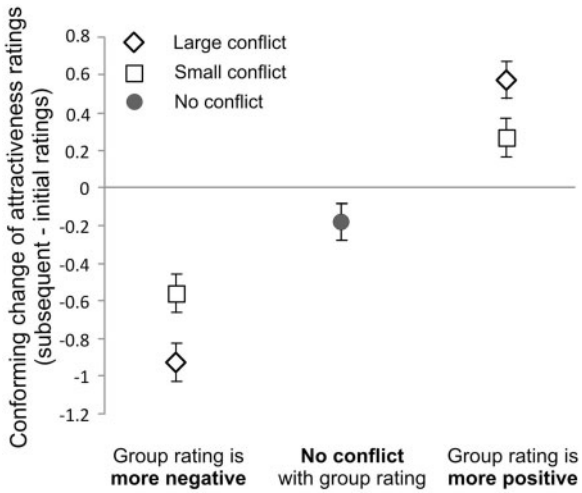

Fig. 1 Mean behavioral conformity effect after large and small conflicts with the group opinion. The graph illustrates the change in the faces' attractiveness measured during the behavioral session when compared with the initial ratings during the ERP session. Error bars indicate 1 standard error of the mean.

attractiveness ratings for faces were substantially changed due to the social influence of a group rating.

On average, conformity changes occurred in 49.8% of the conflict trials. Figure 2A shows that the proportion of trials followed by conforming changes was significantly higher when a large conflict occurred as opposed to when a small conflict occurred. A two-way ANOVA (positive/negative conflicts, small/large conflicts) led to a main effect of the factor conflict size, $F(1,15) = 20$, $P < 0.001$, $\eta^2 = 0.18$. The proportion of trials followed by conforming changes was slightly higher when the group ratings were more negative than participants' own ratings as opposed to when group ratings were more positive: we found a main effect of the factor positive/negative group ratings, $F(1,15) = 5.44$, $P = 0.034$, $\eta^2 = 0.11$.

Previous studies have robustly demonstrated that social influence is most effective in ambiguous situations (Cialdini and Goldstein, 2004). Therefore, conformity effects should be particularly strong for highly ambiguous faces, that is for faces whose initial ratings vary greatly across participants. To determine the ambiguity level of each face stimulus, we analyzed the SD of the initial ratings in the first session for each face across all participants. The SD varied between 0.6 and 6.3. Faces with low variance (SD $\leq$ 2.78, $n = 86$, up to the 40th percentile) and high variance (SD $\geq$ 3.36, $n = 89$, from the 60th percentile; a slight asymmetry is caused by a rounding of values) were selected for further analysis as ambiguous and unambiguous faces, respectively. The size of the conformity effect (the absolute change of attractiveness ratings due to a conflict with the group rating) should be higher for ambiguous faces than for unambiguous faces. In line with this hypothesis, conforming changes were larger in the case of ambiguous when compared with less ambiguous faces (Figure 2B). A two-way ANOVA with ambiguity (ambiguous *vs* less ambiguous) and group ratings (more positive, more negative and consistent group rating) as two within-subject factors revealed a significant interaction effect, $F(2,14) = 8.33$, $P = 0.011$, $\eta^2 = 0.03$. In summary, the behavioral results show that social normative influence induced significant conforming adjustments of the judgment of facial attractiveness.

Table 1 Conformity effects and SDs for different levels of conflict

Mean group ratings (SD)				
More negative		Equal	More positive	
−3	−2	0	+2	+3
−0.92 (0.44)	−0.56 (0.40)	−0.18 (0.27)	0.27 (0.33)	0.58 (0.30)

ERP results

Figure 3 shows ERPs for conflict trials in which the group ratings were in conflict with the participants' own ratings and ERPs for no-conflict trials in which the group ratings were not in conflict with the participants' own ratings, as well as the difference curve. We found a significant difference between the brain responses in conflict and no-conflict trials at a latency of 200 ms. A two-way ANOVA (conflict/no conflict, electrode) led to a main effect of the factor conflict, $F(1,14) = 6.24$, G–G adjusted $P = 0.026$, $\eta^2 = 0.64$. The ERPs in the 'conflict' trials were significantly more negative than the ERPs in the no-conflict trials. A least-significance difference *post hoc* test revealed a significant effect only at the Fpz ($P = 0.025$), Fp1 ($P = 0.009$), F7 ($P = 0.002$), F3 ($P = 0.002$), Fz ($P < 0.001$), F4 ($P = 0.032$), T3 ($P = 0.020$) and C3 ($P = 0.006$) locations and thus supports the hypothesis of a frontal (dorsal cingulate) origin of the observed conflict-related effect. We also compared ERPs with the large (± 3 points) and small (± 2 points) conflicts with the group ratings. We found a trend of significant difference between the large and small conflicts at a latency around 250 ms. A two-way ANOVA revealed an interaction of conflict (large/small) $\times$ electrode, $F(18,252) = 2.18$, G–G adjusted $P = 0.108$. In sum, the results support the learning hypothesis of social influence and show that conflicts with the group opinion triggered a neural response in the frontocentral areas which appears similar to FRN, which is often associated with a performance-monitoring and reinforcement-learning error signal.

Next, we examined whether ERP components exist that are predictive of conforming changes in participants' ratings of facial attractiveness (conformity effect). We compared ERPs with the conflicting group ratings that were followed by changes in perceived attractiveness

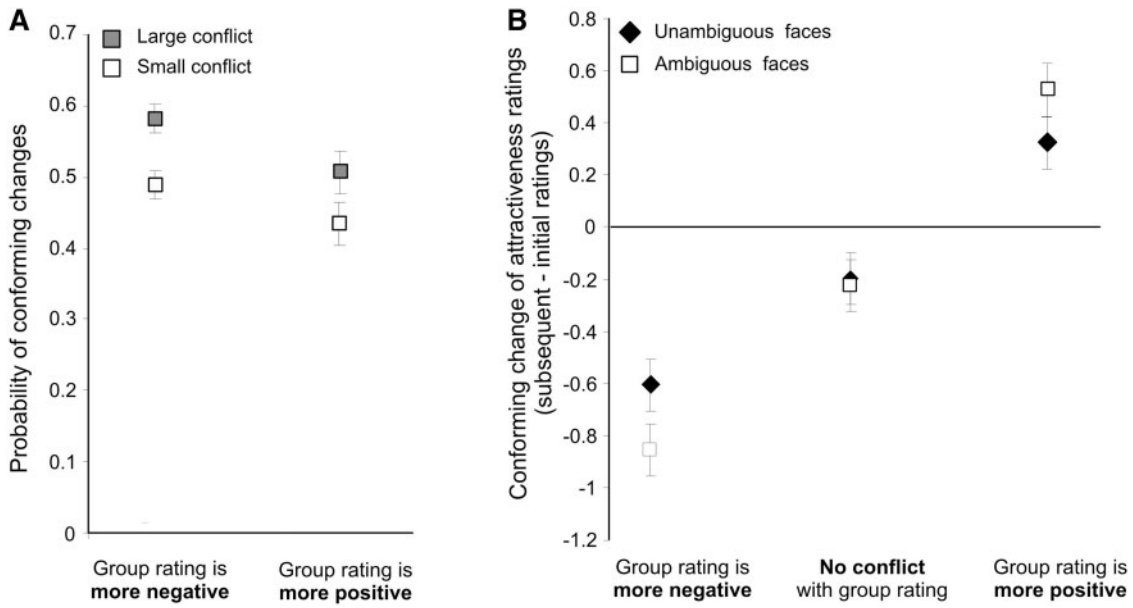

Fig. 2 (**A**) Large conflict with normative opinion led to a higher proportion of trials in which conforming adjustments were made. (**B**) Conformity was stronger for ambiguous than for unambiguous faces. Error bars indicate 1 standard error of the mean.

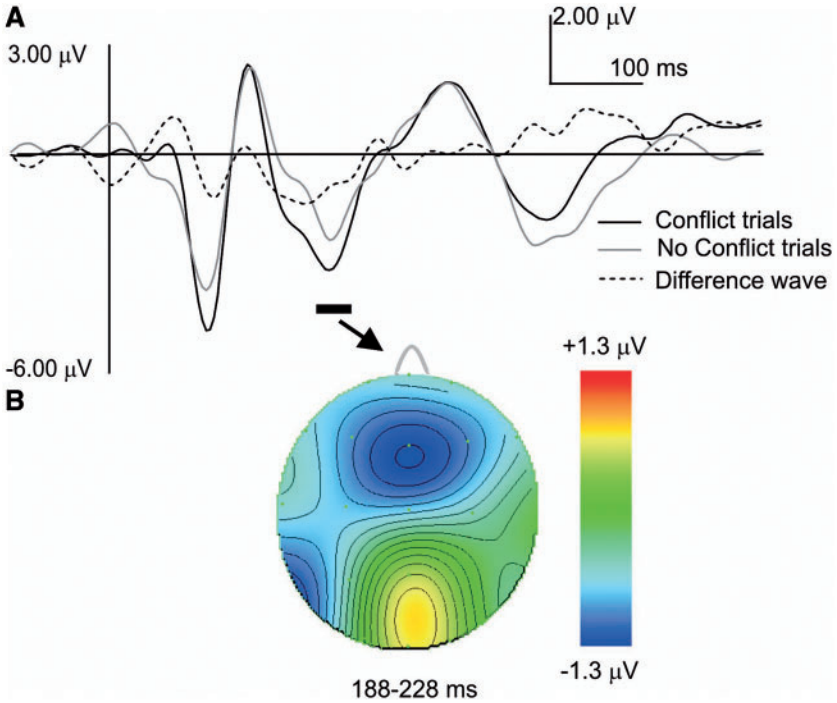

Fig. 3 Social conflict effect. (**A**) Grand-averaged ERPs (top) are presented according to whether participants' ratings of attractiveness of the faces agreed with the group opinion (gray line) or disagreed with the group opinion (black line). The dotted line (subtracted curve) indicates the difference between the agreement and disagreement processes. (**B**) Topographical map of a voltage distribution of the subtracted curve: blue indicating negative, red indicating positive voltages.

of faces in line with group ratings (conformity trials) with ERPs to the conflicting group ratings that were not followed by changes in perceived attractiveness (non-conformity trials). As illustrated in Figure 4, an ERP deflection of interest that reflected a conformity effect consisted of two components: amplitudes of early P310 and late P380 were larger for conformity trials than for non-conformity trials. In order to examine both components, we divided the interval into two even windows featuring both peaks. As indicated by the scalp topographies (Figure 4), both ERP components had a frontocentral maximum. The conformity effect was significant for the early component [two-way ANOVA, conformity × electrode, $F(18,252) = 5.38$, $P = 0.00001$, G–G adjusted $P = 0.002$, $\eta^2 = 0.035$] and for the late component [conformity × electrode, $F(18,252) = 2.63$, $P = 0.00045$, G–G adjusted $P = 0.05$, $\eta^2 = 0.045$]. We also examined whether a conformity effect exists at the latency where the effect of conflicts with the group ratings was initially found (Figure 4). The analysis of amplitudes revealed neither a significant main effect of conformity nor its interaction with the electrode location ($P > 0.1$).

To further examine and confirm the ERP signatures of conforming behavioral changes, we compared ERPs with the conflicting group ratings that were followed by changes in perceived attractiveness *in line* with group ratings (conformity trials, i.e. conforming behavioral changes) with ERPs to the conflicting group ratings that were followed by changes in the perceived attractiveness of faces in the *opposite* direction to group ratings ('opposite' behavioral changes). Additional analysis showed that the early conformity-related ERP component peaking at 310 ms was non-specific; that is it did not differ between conforming and 'opposite' behavioral changes. The later one peaking at 380 ms is a specific precursor of behavioral adjustments in line with the group opinion that is supported by a significant interaction 'direction of behavioral changes' × electrode, $F(18,252) = 3.83$, $P = 0.000001$, G–G adjusted $P = 0.0038$, $\eta^2 = 0.045$. Overall, our results indicate that conforming behavioral adjustments are hallmarked by a late frontocentral cortical activity peaking around 380 ms.

GENERAL DISCUSSION

Starting with the seminal work of Solomon Asch (1951), past research on social influence has demonstrated that people often change their behavior in light of other people's behavior or opinions. In general, people are motivated to win approval and avoid rejection by conforming to others' expectations (Chaiken *et al.*, 1996). Furthermore, others' opinions can often also provide useful information to improve one's own judgments (e.g. Festinger, 1954). Recently, researchers have progressed in examining the neurobiological underpinnings of social influence. Neuroimaging results suggest that conformity and other forms of social influence modulate neural activity in reward- and performance-monitoring neural circuitry (Behrens *et al.*, 2008; Klucharev *et al.*, 2009; Burke *et al.*, 2010a; Campbell-Meiklejohn *et al.*, 2010; Biele *et al.*, 2011; Zaki *et al.*, 2011). Nevertheless, the timing of neuronal activity underlying social conformity has been unknown. Our results for the first time show that a conflict with a normative group opinion triggers a sequence of neuronal responses (peaking around 200–380 ms) reflecting a conflict with normative opinion and a conforming behavioral adjustment.

In this study, we influenced individual opinion by introducing a descriptive norm of facial attractiveness that could be either consistent or inconsistent with a person's own opinion. The behavioral data in our experiment clearly illustrated how the group opinion systematically changed people's judgments. In line with previous research, the conforming behavioral adjustments were especially strong when greater conflicts with the group opinion occurred or when the stimuli were rather ambiguous (Cialdini and Goldstein, 2004).

Electrophysiological correlates of conformity

Our ERP data suggest that conflicts with a normative group opinion trigger FRN—a frontocentral negative deflection with the maximum at 200 ms that had often been implicated in performance monitoring and signaling of negative reward prediction error

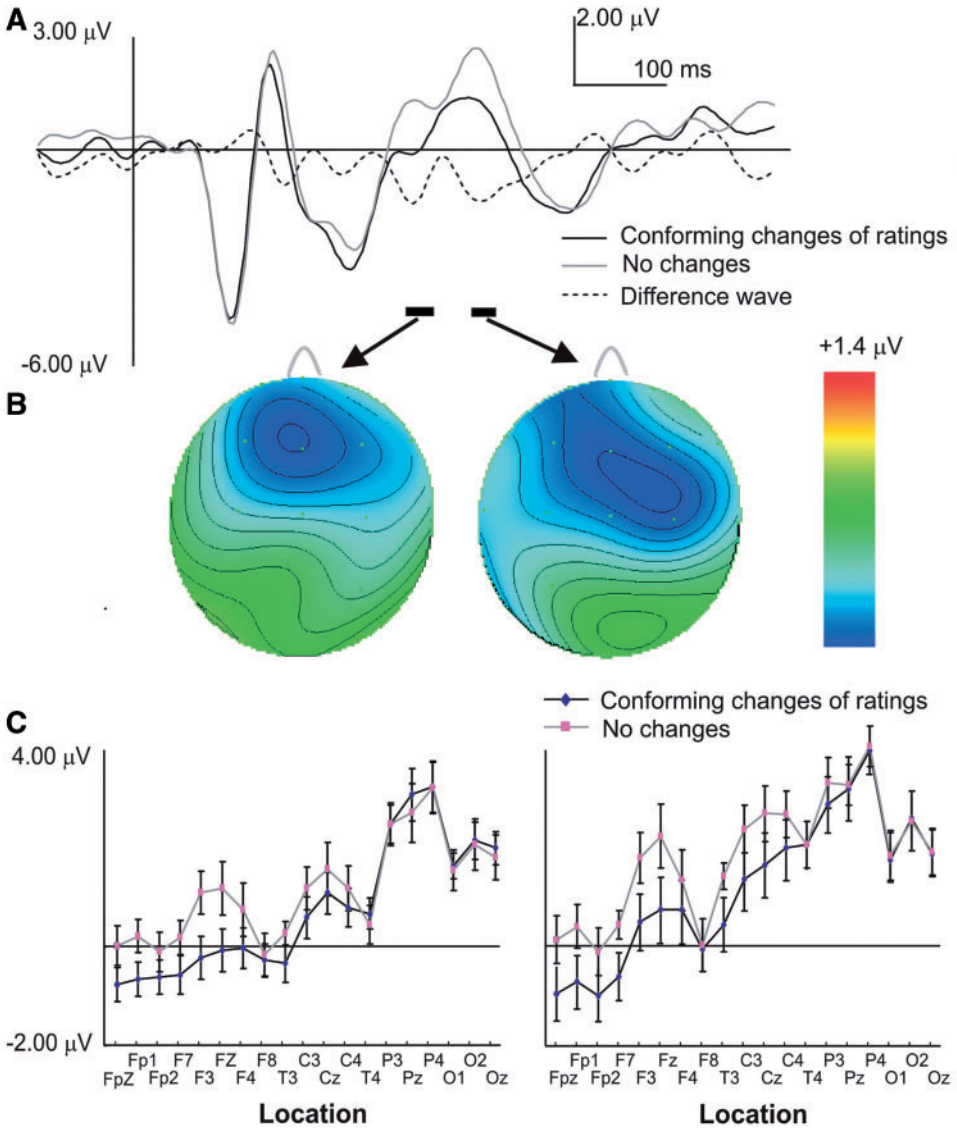

Fig. 4 Social conformity effects. (**A**) Grand-averaged ERPs are presented according to whether participants changed their opinion in line with the group opinion (black line) or did not change it at all (gray line). The dotted line (subtracted curve) indicates the difference between the ERPs followed by changes in line with the group opinion and no changes. (**B**) Topographical map of a voltage distribution of the subtracted curve: blue indicating negative, red indicating positive voltages. (**C**) Bar plots (means with standard errors) for the early (left) and late (right) conformity effects illustrating the interaction between *conformity* and *electrode*.

(Miltner *et al.*, 1997; Gehring and Willoughby, 2002; Holroyd *et al.*, 2002; Nieuwenhuis *et al.*, 2004; Nieuwenhuis *et al.*, 2007). Thus, the social influence of group norms could be based on a general performance-monitoring mechanism. Accordingly, deviations from descriptive norms are perceived as negative behavioral outcomes. FRN-like signals have also been previously recorded during the observation of others' errors in a modified Eriksen flanker task (van Schie *et al.*, 2004) or when observing the consequences of others' actions in a gambling task (Yu and Zhou, 2006). Our results suggest that people not only automatically monitor their own and others' performances, as previously demonstrated, but also continuously compare their own behavior with the 'normative' one. The ERP results show that FRN is triggered by the individual behavioral outcomes calculated relative to the group normative behavior.

We also demonstrated that conforming adjustments were preceded by a frontocentral waveform peaking at 380 ms. Unfortunately, the relatively limited spatial resolution of ERPs does not allow testing the hypothesis that both early and late components are generated in exactly the same brain area or by the same neural populations. However, similar frontocentral voltage distribution and previous fMRI studies (e.g. Klucharev *et al.*, 2009) have pointed to the possible involvement of similar posterior medial frontal areas in the early and the late response. Overall, our results suggest that a conflict with a group opinion triggers a sequence of neuronal responses in the posterior medial frontal cortex: from initial generation of FRN detecting a violation of descriptive norms at 200 ms to later neural activity, peaking at 380 ms after the conflict and relating to behavioral adjustments underlying conformity.

In previous studies, FRN was often but not necessarily always followed by a positive waveform (P3/Pe complex or error positivity, e.g. see Nieuwenhuis *et al.*, 2001), which is also associated with outcome evaluation, decision making and high-order behavioral adjustments (Nieuwenhuis *et al.*, 2001; Yeung and Sanfey, 2004; Hajcak *et al.*, 2005; Hajcak *et al.*, 2007). Two error-related components could represent different aspects of error processing with the later positive component probably reflecting deliberate processing of the error event

(Falkenstein *et al.*, 2000) or adjustment of behavior on the basis of explicit rules (Chase *et al.*, 2011). Importantly, the conformity effect peaking at 380 ms after the conflict reported in our study has a frontocentral maximum in contrast to the parietal maximum of the classical error positivity (Falkenstein *et al.*, 2000; Chase *et al.*, 2011). Previous studies suggested that the amplitude of the error positivity reflects adjustment of the response strategy and the subjective significance of the errors (Falkenstein *et al.*, 2000). In contrast, in our study the most effective trials followed by conforming adjustments evoked the smaller ERP than the trails followed by no adjustments: as indicated by the negative differential wave of the conformity effect. Thus, the conformity effect in our study is rather different from the classical error positivity that is likely to be due to the difference in spatial origin of the measured EEG signals. Neural activity peaking at 380 ms could represent an extended FRN overlapping with the later positive component. However, we cannot exclude that deliberate processing of conflicts with the group opinion (often associated with the classical error positivity) could contribute to conforming adjustments in our study.

In contrast to our results, those of previous studies showed that relatively early activity of the posterior medial frontal cortex underlies the ability to adjust decision-making behavior. For example, Cohen and Ranganath (2007) examined behavior in the 'matching pennies' game (i.e. a coordination game) and found that the magnitude of FRN after losing to a computer opponent predicted whether people would change their decisions on the subsequent trial. However, unlike the fast transient changes of decisions in the game situation, we investigated longer lasting conforming adjustments measured some 15–45 min after the normative conflict. Therefore, it appears plausible that immediate behavioral adjustments could be reflected by an earlier neural response in the cingulate cortex, whereas a longer lasting effect might need some form of plastic change underlined by much later neural responses. Interestingly, correlates of long-lasting adjustments (e.g. subsequent memory effects) are often reported in the interval between 400 and 1100 ms (for a review see Friedman and Johnson, 2000). However, the neural circuitry of this process remains to be studied in detail.

A general mechanism of social influence

Montague and Lohrenz (2007) suggested that conformity with social norms requires an 'error' signal indicating deviations from norms. Perhaps such an 'error' signal shares the same neural mechanism as the standard 'reward prediction error' underlying reinforcement learning. A single exposure to a social influence in our study makes it virtually impossible to apply conventional reinforcement learning models to describe conforming behavior. Nevertheless, one can speculate that social influence could work on a similar mechanism; that is a conflict with a group opinion might generate a 'social' reward prediction error signal. More precisely, a difference between a person's attractiveness rating and the group's opinion could be perceived as an error. In many real-life situations, our opinions are affected by a single exposure to social feedback: for example, a reviewer's opinion or a medical doctor's recommendation. In these cases, people might compare their own opinion or expectation with the social feedback, and this difference could be reflected as a prediction error. This difference could then be used to adjust one's own belief, depending on how much weight it is given.

Interestingly, studies on the spatial overlap of brain regions involved in social influence and reinforcement learning provide additional arguments for a similarity of the underlying mechanism (Behrens *et al.*, 2008; Klucharev *et al.*, 2009; Berns *et al.*, 2010; Campbell-Meiklejohn *et al.*, 2010; Falk *et al.*, 2010; Klucharev *et al.*, 2011). Importantly, experiments that were specifically designed to model reward prediction error (Behrens *et al.*, 2008; Burke *et al.*, 2010a; Biele *et al.*, 2011) demonstrated a prediction-error-like signal generated by some forms of social influence. It is interesting to note that classical psychological studies explain conformity by the rewarding value of social approval or affiliation with others (Cialdini and Goldstein, 2004); behavioral economists also highlight the effects of social punishment for violations of the group norm (Fehr and Fischbacher, 2004a,b). In fact, both explanations of conforming behavior are consistent with a general reinforcement learning mechanism; that is compliance with social norms and conforming behavioral adjustments to others are reinforced.

One possible alternative explanation of our results is that normative group pressure triggers anxiety or emotional/cognitive dissonance (van Veen *et al.*, 2009; Berns *et al.*, 2010). Accordingly, people adjust their opinion to reduce negative emotional states. However, the FRN observed in our study indicates a general performance-monitoring mechanism of behavioral adjustment (Holroyd and Coles, 2002; Ridderinkhof *et al.*, 2004; Matsumoto *et al.*, 2007). Further studies are needed to clarify the exact role of the posterior medial frontal cortex in social influence. ERP studies of the time-estimation task suggested that the rostral anterior cingulate cortex (rACC) could be involved in the FRN generation (Nieuwenhuis *et al.*, 2005; Mies *et al.*, 2011). According to this view, the posterior medial cortex is primarily involved in the processing of feedback validity, whereas the rACC is primarily involved in the processing of feedback valence (Mies *et al.*, 2011). A high-density EEG study could improve the localization of the observed electrophysiological precursors of social conformity. In addition, different mechanisms can underlie conformity (Cialdini and Goldstein, 2004). For example, informational conformity (as contrasted with normative conformity) serves an informational function helping to be accurate and can be underlined by an attention-related neural mechanism (e.g. study by Berns *et al.*, 2005). More studies are clearly needed to determine all mechanisms of conformity. Current results should be interpreted with caution because we investigated a female population only. A high-density EEG study could further improve localization of the observed electrophysiological precursors of social conformity. Further studies will help to generalize the observed mechanisms to the male population and other social situations leading to conformity.

Taken together, our behavioral results clearly show that people continuously change their opinion in light of a different normative opinion of the group. A mismatch between individual and group opinions triggered a frontocentral negative deflection similar to FRN, implicated in individual learning. Furthermore, the FRN was followed by brain activity underlying conforming behavioral adjustment and peaking around 380 ms. This work complements earlier high spatial resolution fMRI studies with the complex temporal structure of the neural underpinnings of conforming behavioral adjustments. In general, our results support the hypothesis that forms of social influence are mediated by activity of the posterior medial frontal cortex as a part of the general performance-monitoring circuitry.

REFERENCES

Ajzen, I., Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow, H., editor. *Groups, Leadership and Men Research in Human Relations*. Pittsburgh, PA: Carnegie Press, pp. 177–90.

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

Berns, G.S., Capra, C.M., Moore, S., Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage*, *49*, 2687–96.

Berns, G.S., Chappelow, J., Zink, C.F., Pagnoni, G., Martin-Skurski, M.E., Richards, J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. *Biological Psychiatry*, *58*, 245–53.

Biele, G., Rieskamp, J., Krugel, L.K., Heekeren, H. (2011). The neural basis of following advice. *PLoS Biology*, *9*, e1001089.

Brown, J.W., Braver, T.S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*, 1118–21.

Burke, C.J., Tobler, P.N., Baddeley, M., Schultz, W. (2010a). Neural mechanisms of observational learning. *Proceedings of the National Acadamy of Sciences of the United States of America*, *107*, 14431–6.

Burke, C.J., Tobler, P.N., Schultz, W., Baddeley, M. (2010b). Striatal BOLD response reflects the impact of herd information on financial decisions. *Frontiers in Human Neuroscience*, *4*. Article 48.

Campbell-Meiklejohn, D., Bach, D., Roepstorff, A., Dolan, R., Frith, C. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, *20*, 1165–70.

Chaiken, S., Wood, W., Eagly, A.H. (1996). Principles of persuasion. In: Higgins, E.T., Kruglansk, I.W., editors. *Social Psychology: Handbook of Basic Principles*. New York, NY: Guilford Press, pp. 702–42.

Chase, H.W., Swainson, R., Durham, L., Benham, L., Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*, 936–46.

Cialdini, R.B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, *72*, 263–8.

Cialdini, R.B., Goldstein, N.J. (2004). Social influence: compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Cloutier, J., Heatherton, T.F., Whalen, P.J., Kelley, W.M. (2008). Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of Cognitive Neuroscience*, *20*, 941–51.

Cohen, M.X., Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, *27*, 371–8.

di Pellegrino, G., Ciaramelli, E., Ladavas, E. (2007). The regulation of cognitive control following rostral anterior cingulate cortex lesion in humans. *Journal of Cognitive Neuroscience*, *19*, 275–86.

Falk, E.B., Berkman, E.T., Mann, T., Harrison, B., Lieberman, M.D. (2010). Predicting persuasion-induced behavior change from the brain. *Journal of Neuroscience*, *30*, 8421–4.

Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, *51*, 87–107.

Fehr, E., Fischbacher, U. (2004a). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Fehr, E., Fischbacher, U. (2004b). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*, 185–90.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–40.

Friedman, D., Johnson, R.Jr (2000). Event-related potential (ERP) studies of memory encoding and retrieval: a selective review. *Microscopy Research and Technique*, *51*, 6–28.

Gehring, W.J., Willoughby, A.R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*, 2279–82.

Gentsch, A., Ullsperger, P., Ullsperger, M. (2009). Dissociable medial frontal negativities from a common monitoring system for self- and externally caused failure of goal achievement. *Neuroimage*, *47*, 2023–30.

Gu, R.L., Ge, Y., Jiang, Y., Luo, Y.J. (2010a). Anxiety and outcome evaluation: the good, the bad and the ambiguous. *Biological Psychology*, *85*, 200–6.

Gu, R.L., Huang, Y.X., Luo, Y.J. (2010b). Anxiety and feedback negativity. *Psychophysiology*, *47*, 961–7.

Hajcak, G., Holroyd, C.B., Moser, J.S., Simons, R.F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*, 161–70.

Hajcak, G., McDonald, N., Simons, R.F. (2003). Anxiety and error-related brain activity. *Biological Psychology*, *64*, 77–90.

Hajcak, G., Moser, J.S., Holroyd, C.B., Simons, R.F. (2007). It's worse than you thought: the feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*, 905–12.

Holroyd, C.B., Coles, M.G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

Holroyd, C.B., Coles, M.G., Nieuwenhuis, S. (2002). Medial prefrontal cortex and error potentials. *Science*, *296*, 1610–1; author reply 1610–1.

Holroyd, C.B., Nieuwenhuis, S., Yeung, N., et al. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*, 497–8.

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 22014–9.

Kerns, J.G., Cohen, J.D., MacDonald, A.W.3rd, Cho, R.Y., Stenger, V.A., Carter, C.S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*, 1023–6.

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*, 140–51.

Klucharev, V., Munneke, M.A., Smidts, A., Fernandez, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience*, *31*, 11934–40.

Klucharev, V., Smidts, A., Fernandez, G. (2008). Brain mechanisms of persuasion: how 'expert power' modulates memory and attitudes. *Social Cognitive and Affective Neuroscience*, *3*, 353–66.

Langlois, J.H., Kalakanis, L., Rubenstein, A.J., Larson, A., Hallam, M., Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*, 390–423.

Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, *10*, 647–56.

Mies, G.W., van der Molen, M.W., Smits, M., Hengeveld, M.W., van der Veen, F.M. (2011). The anterior cingulate cortex responds differently to the validity and valence of feedback in a time-estimation task. *Neuroimage*, *56*, 2321–8.

Miltner, W.H.R., Braun, C.H., Coles, M.G.H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–98.

Montague, P.R., Lohrenz, T. (2007). To detect and correct:norm violations and their enforcement. *Neuron*, *56*, 14–18.

Nieuwenhuis, S., Holroyd, C.B., Mol, N., Coles, M.G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*, 441–8.

Nieuwenhuis, S., Ridderinkhof, K.R., Blom, J., Band, G.P., Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, *38*, 752–60.

Nieuwenhuis, S., Schweizer, T.S., Mars, R.B., Botvinick, M.M., Hajcak, G. (2007). Error-likelihood prediction in the medial frontal cortex: a critical evaluation. *Cerebral Cortex*, *17*, 1570–81.

Nieuwenhuis, S., Slagter, H.A., von Geusau, N.J., Heslenfeld, D.J., Holroyd, C.B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *The European Journal of Neuroscience*, *21*, 3161–8.

O'Doherty, J., Critchley, H., Deichmann, R., Dolan, R.J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *Journal of Neuroscience*, *23*, 7931–9.

Oldfield, R.C. (1971). Assessment and analysis of handedness–Edinburgh inventory. *Neuropsychologia*, *9*, 97–113.

Ridderinkhof, K.R., Ullsperger, M., Crone, E.A., Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*, 443–7.

Roger, C., Benar, C.G., Vidal, F., Hasbroucq, T., Burle, B. (2010). Rostral Cingulate Zone and correct response monitoring: ICA and source localization evidences for the unicity of correct- and error-negativities. *Neuroimage*, *51*, 391–403.

Spielberger, C.D., Gorsuch, R.L., Lushene, R.E. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Tremblay, P., Shiller, D.M., Gracco, V.L. (2008). On the time-course and frequency selectivity of the EEG for different modes of response selection: evidence from speech production and keyboard pressing. *Clinical Neurophysiology*, *119*, 88–99.

van den Bos, W., Guroglu, B., van den Bulk, B.G., Rombouts, S.A.R.B., Crone, E.A. (2009). Better than expected or as bad as you thought? The neurocognitive development of probabilistic feedback processing. *Frontiers in Human Neuroscience*, *3*. Article 52.

van Schie, H.T., Mars, R.B., Coles, M.G.H., Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*, 549–54.

van Veen, V., Krug, M.K., Schooler, J.W., Carter, C.S. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, *12*, 1469–74.

West, R., Langley, M.M., Bailey, K. (2011). Signaling a switch: neural correlates of task switching guided by task cues and transition cues. *Psychophysiology*, *48*, 612–23.

Willis, J., Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–8.

Yeung, N., Sanfey, A.G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*, 6258–64.

Yu, R., Zhou, X. (2006). Brain responses to outcomes of one's own and other's performance in a gambling task. *Neuroreport*, *17*, 1747–1751.

Zaki, J., Schirmer, J., Mitchell, J.P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, *22*, 894–900.