Received: 8 September 2020 | Revised: 6 November 2020 | Accepted: 10 November 2020

DOI: 10.1002/hbm.25296

RESEARCH ARTICLE

WILEY

Testing the reinforcement learning hypothesis of social conformity

Marie Levorsen[1] | Ayahito Ito[1,2] | Shinsuke Suzuki[3] | Keise Izuma[1,2,4]

[1]Department of Psychology, University of Southampton, Southampton, UK

[2]Research Center for Future Design, Kochi University of Technology, Kochi, Japan

[3]Brain, Mind and Markets Laboratory, Department of Finance, Faculty of Business and Economics, the University of Melbourne, Parkville, Australia

[4]School of Economics and Management, Kochi University of Technology, Kochi, Japan

Correspondence
Keise Izuma, School of Economics and Management, Kochi University of Technology, 2-22 Eikokuji, Kochi City, Kochi 780-8515, Japan.
Email: izuma.keise@kochi-tech.ac.jp

Abstract

Our preferences are influenced by the opinions of others. The past human neuroimaging studies on social conformity have identified a network of brain regions related to social conformity that includes the posterior medial frontal cortex (pMFC), anterior insula, and striatum. Since these brain regions are also known to play important roles in reinforcement learning (i.e., processing prediction error), it was previously hypothesized that social conformity and reinforcement learning have a common neural mechanism. However, although this view is currently widely accepted, these two processes have never been directly compared; therefore, the extent to which they shared a common neural mechanism had remained unclear. This study aimed to formally test the hypothesis. The same group of participants (*n* = 25) performed social conformity and reinforcement learning tasks inside a functional magnetic resonance imaging (fMRI) scanner. Univariate fMRI data analyses revealed activation overlaps in the pMFC and bilateral insula between social conflict and unsigned prediction error and in the striatum between social conflict and signed prediction error. We further conducted multivoxel pattern analysis (MVPA) for more direct evidence of a shared neural mechanism. MVPA did not reveal any evidence to support the hypothesis in any of these regions but found that activation patterns between social conflict and prediction error in these regions were largely distinct. Taken together, the present study provides no clear evidence of a common neural mechanism between social conformity and reinforcement learning.

KEYWORDS

fMRI, MVPA, prediction error, reinforcement learning, social conformity

1 | INTRODUCTION

Humans are highly sensitive to social influence, and our everyday decisions are often guided by the opinions of others. One of the best known forms of social influence is social conformity, which refers to the act of changing one's judgments, attitudes, and preferences to align with the expectations of others (Cialdini & Goldstein, 2004). The neural mechanism underlying this important social phenomenon has been investigated over the past two decades using functional magnetic resonance imaging (fMRI).

A seminal study by Klucharev et al. (Klucharev, Hytonen, Rijpkema, Smidts, & Fernandez, 2009) found that the posterior medial frontal cortex (pMFC) and ventral striatum play important roles in social conformity. They asked participants to rate the attractiveness of female faces, and after rating each face, the participants were presented with the ratings of the same face by another group of people. They found that the larger the difference between an individual's rating and the group rating, the higher the activity in the pMFC, while the opposite pattern was found in the ventral striatum (i.e., the closer the individual and group ratings, the higher the activity in the ventral striatum). In a study on social conformity that used transcranial magnetic stimulation, it was shown that the pMFC plays a causal role in preference change (Klucharev, Munneke, Smidts, & Fernandez, 2011). Using the similar experimental paradigm, several studies have replicated the original fMRI findings (e.g., Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010; Izuma & Adolphs, 2013; Korn et al., 2014; Korn, Prehn, Park, Walter, & Heekeren, 2012; Wake, Aoki, Nakahara, & Izuma, 2019), and a recent meta-analysis revealed that the insula and pMFC are consistently positively related to social conflict (i.e., the difference between individual and group opinions) and that the ventral striatum is negatively related to social conflict (Wu, Luo, & Feng, 2016).

As these brain regions (especially the pMFC and striatum) are known to play pivotal roles in reinforcement learning, Klucharev et al. (2009) proposed an interesting hypothesis that a complex social phenomenon of social conformity may share common neural mechanisms as a simple nonsocial reward-based learning process (Izuma, 2013, 2017). It has been reported in several human neuroimaging studies and animal neurophysiology studies that ventral striatum activity tracks reward prediction error (i.e., the difference between expected and actual outcomes) and that the pMFC and insula are involved in processing unsigned prediction error (i.e., the absolute degree of deviation from expectations) (see Fouragnan, Retzler, & Philiastides, 2018 for a recent meta-analysis of human neuroimaging studies).

In addition to the commonly activated regions reported in human neuroimaging studies, social conformity and reinforcement learning are similar in at least the three ways presented below. First, there is a conceptual similarity between social conformity and reinforcement learning as they are processes that involve the adjustment of the behavior of an individual (e.g., rating or choice) based on received feedback (e.g., group opinion or reward) (Izuma, 2017). Second, the neurotransmitter dopamine is known to play a role in both processes. It is well established that dopamine neurons in the midbrain, which is heavily interconnected to the ventral striatum, signal reward prediction error (Schultz, 2015). Furthermore, a pharmacological study with human participants has shown that social conformity effect is modulated by methylphenidate, which indirectly increases extracellular dopamine levels in the brain (Campbell-Meiklejohn et al., 2012). Third, several electroencephalogram (EEG) studies have reported an EEG signal over the pMFC called feedback-related negativity (FRN), which is related to prediction error (Holroyd & Coles, 2002), and several EEG studies on social conformity have reported a similar signal over the pMFC that correlates with the difference between individual and group opinions (e.g., Chen, Wu, Tong, Guan, & Zhou, 2012; Kim, Liss, Rao, Singer, & Compton, 2012; Shestakova et al., 2012).

However, importantly, there are at least two reasons to believe that the hypothesis is too simplistic. First, in the context of the original social conformity task (Klucharev et al., 2009), the idea that social conflict is the same as prediction error implies that participants expect that the group's rating is always the same as their ratings, which seems highly unlikely. In fact, using a similar social conformity paradigm, we had previously asked participants to guess the group ratings, but their expectations of group ratings were not related to their own ratings (Izuma & Adolphs, 2013). In other words, the social conformity task was not perceived as a learning task (i.e., learning preferences of a group) by participants. If so, underlying computations are likely to be different between social conformity and reinforcement learning. Second, another important difference is that while striatal activity is positively related to signed prediction error during the reinforcement learning task (i.e., actual vs. expected rewards), it is negatively related to social conflict during the social conformity task (i.e., the absolute difference between one's vs. group's rating). This negative correlation with social conflict in the social conformity task corresponds to a negative correlation with unsigned prediction error in the reinforcement learning task (i.e., the absolute difference between actual vs. expected rewards). Thus, although the same striatal region is involved in both tasks, signals related to its activity are conceptually different between the two tasks.

The reinforcement learning hypothesis of social conformity suggests an interesting possibility of bridging two previously-unrelated literatures of neuroscience research on reinforcement learning and psychology research on social conformity and could significantly advance the understandings of the neural and psychological mechanisms of how we are influenced by others. However, while the reinforcement-learning hypothesis is currently most widely accepted as the neural mechanism of social conformity (Campbell-Meiklejohn et al., 2012; Chen et al., 2012; Izuma, 2013, 2017; Kim et al., 2012; Klucharev et al., 2009; Shestakova et al., 2012), to the best of our knowledge, these two processes have never been directly compared to each other. Thus, evidence in support of a common neural mechanism is still insufficient. Thus, the aim of this study was to rigorously test the reinforcement learning hypothesis of social conformity by asking the same sample of participants to perform social conformity and reinforcement learning tasks inside an fMRI scanner.

A question of whether social vs. nonsocial processes share the same neural mechanism remains an important topic in neuroscience (Lockwood, Apps, & Chang, 2020). Drawing on the idea of Marr's three levels (Marr, 1982), Lockwood et al. (2020) argued that social vs. nonsocial processes may be similar or distinct at each of three different levels: (a) computational level, (b) algorithmic level, and (c) implementational level. In the present study, we aimed to contribute to this important theoretical debate and test whether social conformity and reinforcement learning are similar at least at the implementational level (i.e., whether social conflict and reward prediction error are processed in the same brain region). Importantly, it is increasingly being recognized that even if there are activation overlaps in the same sample of participants, activation overlaps based on traditional univariate fMRI data analysis cannot be considered strong evidence of a common neural mechanism (e.g., Woo et al., 2014). Therefore, we used multi-voxel pattern analysis (MVPA) to obtain more compelling evidence to support or refute the hypothesis (Peelen & Downing, 2007).

To test the hypothesis, we selected particular social conformity and reinforcement learning tasks, which are the most representative and appropriate to test the neural correlates of social conflict and prediction error signals. More specifically, for the social conformity task, we employed the same face attractiveness rating task as that in the original study by Klucharev et al. (2009) (see Figure 1a). The same (or conceptually similar) paradigm has been used in a number of previous studies (e.g., Campbell-Meiklejohn et al., 2012; Izuma & Adolphs, 2013; Klucharev et al., 2011; Shestakova et al., 2012; Zaki, Schirmer, & Mitchell, 2011). For the reinforcement learning task, we adopted a probabilistic reward learning task from Cooper, Dunne, Furey, and O (2012) in which participants were asked to pick one of two slot machines in each trial (Figure 1b). We selected this particular task because of its simplicity, and this task is conceptually similar to that used in other neuroimaging studies, in which the neural correlates of reward prediction error were demonstrated (e.g., Burke, Tobler, Baddeley, & Schultz, 2010; Fouragnan, Retzler, Mullinger, & Philiastides, 2015; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006).

2 | MATERIALS AND METHODS

Twenty-nine right-handed female students with no history of psychiatric disorders were recruited from the University of Southampton (mean age = 22.12 years). As in the original study (Klucharev et al., 2009), only female participants were recruited for the study as previous research suggests that there are gender differences in neural activity related to attractiveness rating (e.g., Cloutier, Heatherton, Whalen, & Kelley, 2008). Data from four participants were not included in the final analyses for the following reasons: excessive head movement (>3 mm) in one participant, strong doubts regarding the social conformity manipulation in two participants (see below for more details), and technical problems with the response box in one participant (this participant could not complete all the fMRI tasks). The final sample consisted of 25 participants (mean age = 22.1 years). Written consent was obtained from all the participants prior to the experiment, and the study was approved by the University of Southampton ethics committee.

2.1 | Stimuli

For the social conformity task, 100 digital color photographs of Caucasian women (aged 18–35) were used as stimuli. The images were taken from the set used in the study by Klucharev et al. (2009). All the women in the photographs were moderately attractive and had a moderate smile.

2.2 | Experimental procedure

The experiment consisted of two parts, namely an fMRI session and a behavioral session. Prior to the experiment, participants were given

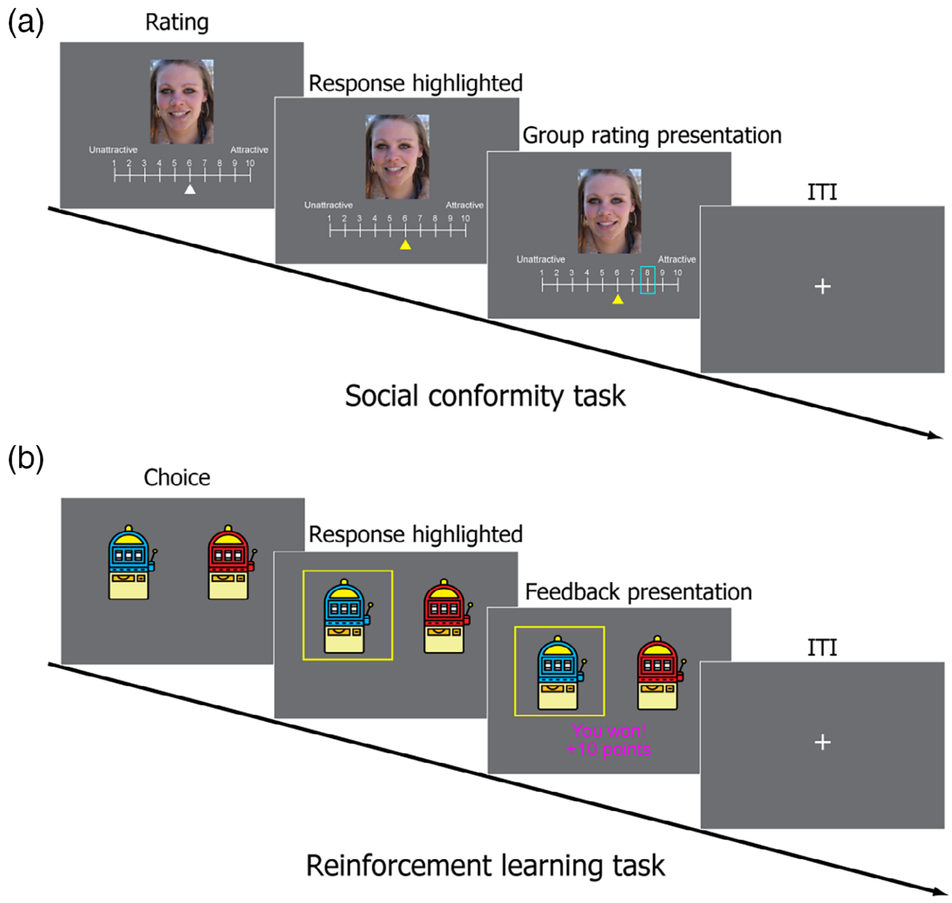

FIGURE 1 Experimental tasks. (a) Social conformity task. Participants were shown images of female faces and a 10-point scale and asked to rate the attractiveness of each face. After the participants submitted their ratings, they were shown the ratings of each face by other people (in blue frame) for 2 s. (b) Reinforcement learning task (probabilistic reward learning task). Participants were presented with two slot machines and asked to pick 1 of them. After the participants made their decisions, they were presented with a win or loss outcome

instructions and practice trials on the tasks. During the fMRI session, participants were asked to perform the following two tasks: (a) social conformity task and (b) reinforcement learning task. Both tasks were programmed using Psychtoolbox (http://psychtoolbox.org/) with Matlab software (version 2018b, http://www.mathworks.co.uk). Participants completed two runs of each task (a total of four fMRI runs) and each run consisted of 50 trials. The order of the tasks was counterbalanced across participants.

Social Conformity task We adopted the social conformity task used in the original study (Klucharev et al., 2009). For each trial, participants were presented with an image of a female face and a 10-point scale (Figure 1a). The participants were asked to rate the attractiveness of each face on a scale of 1 (least attractive) to 10 (most attractive). Each trial consisted of the following three phases: (a) rating phase (no time limit, but participants were encouraged to answer as quickly as they can), (b) highlight of response phase (1–5 s, mean = 2 s), (c) group rating presentation phase (2 s). There was an inter-trial interval (ITI) (1–7 s, mean = 2.5 s) between trials. The participants were asked to indicate their answers using response button handles (with one handle held in each hand). They used both of their index fingers to move a white cursor through the rating scale (e.g., a right index finger button press moved the cursor 1 point to the right). The white cursor remained invisible until the participant pressed a button, and it appeared below the scale when participants started scrolling. The initial position of the cursor in each trial was randomly determined. The participants were asked to use the cursor location to indicate their chosen rating and to press the right thumb button to select their chosen rating. In the highlight of response phase, the rating chosen by the participant was highlighted with a yellow cursor.

In the group rating presentation phase, the participants were presented with a rating in a blue frame that represented "the mean group rating of other students at the University of Southampton" (Figure 1a). During the instruction, the participants were informed about the meaning of the blue frame and were led to believe that the group rating was an actual group mean rating based on responses from other students at the University of Southampton. In reality, it was systematically manipulated such that the group rating matched the rating of the individual participant in 17.5–25% of the trials and such that the group rating was roughly equally less than or greater than the rating of the participant in 75–82.5% of the trials. The group rating did not deviate from the participant's rating by more than 3 points. The order of the images was randomized for each participant.

Reinforcement Learning task This was a probabilistic reward learning task adapted from the study by Cooper et al. (2012). In each trial, participants were asked to pick one of two slot machines (Figure 1b). There was a certain probability of winning 10 points (equivalent to £1) on each of the two slot machines. There were independent probabilities of winning on each trial on the two slot machines, and the probabilities changed gradually over the 50 trials in each run to ensure that learning continued throughout the task so that the magnitude of prediction errors varied widely. Specifically, each slot machine's reward probability followed a sine curve set to drift between 0 and 100% probability. A starting point was randomly determined, and half-period was randomly set between 0.87 and 1.67 times the number of trials per run (i.e., 50 trials). The reward probabilities of the two slot machines were constrained to be correlated with each other at less than r = .02. Finally, a small amount of Gaussian noise was added to each trial (M = 0, *SD* = 6%) before scaling sines to have a range of 0–100% (Cooper et al., 2012). There were four different slot machines (red, blue, green, and purple), and two runs of the reinforcement learning task were performed with different combinations of two slot machines. The combinations of the slot machines and the location of the two slot machines were counterbalanced across participants.

Each reinforcement learning trial consisted of the following three phases (Figure 1b): (a) choice phase (participant's response, <2 s), (b) highlight of response phase (1–7 s, mean = 2.5 s), and (c) outcome phase (2 s). Trials were separated by an ITI (1–7 s, mean = 2.5 s). In the choice phase, the participants were asked to choose the slot machine they thought would result in a win outcome within 2 s. They were informed that the probability of winning associated with each slot machine might change gradually throughout the experiment. The participants were also told that two trials would be randomly selected (1 from each of the 2 runs) at the end of the experiment and that they would receive a cash bonus depending on the outcome of the two trials.

The participants were asked to indicate their answers by pressing the response handle buttons using the left or right index finger. If they did not respond within 2 s, an error message ("Too slow!!!") was shown, and the trial was repeated. In the highlight of response phase, the chosen slot machine image was highlighted by a yellow frame. In the outcome phase, 1 of the 2 following possible outcomes was shown: "You won! +10 points" or "You lost. 0 points" (Figure 1b). The outcome messages were written in magenta or cyan font color, and combinations of font colors (magenta or cyan) and outcomes (win or loss) were counterbalanced across participants.

2.3 | Behavioral session

After the scan, to measure how participants' face ratings were affected by group opinion (i.e., social conformity effect), they were unexpectedly (unannounced during the initial instruction) asked to rate each face again, this time, without the group rating. They rated the same 100 faces again in a new randomized order. Next, the participants were asked to complete a demographic questionnaire. On the questionnaire, participants were asked to indicate with a "yes" or a "no" if they had any doubts about the group ratings presented during the fMRI face-rating task. If the indication is a "yes," the participant is asked to explain the doubt in a follow-up interview. As earlier mentioned, two participants were excluded for their strong doubts about the group rating (both of them explicitly said that they did not believe the group ratings presented to them during the fMRI scanning). Lastly, the participants were paid and debriefed.

2.4 | fMRI data acquisition

All the images were obtained using a Siemens 3.0 Tesla Skyra scanner. For functional imaging in both task sessions, T2*-weighted gradient-echo echo-planar imaging (EPI) sequences were used with the following parameters: time repetition = 2,500 ms, echo time = 25 ms, flip angle = 90°, field of view = 220 mm, and voxel dimension = 3.0 × 3.0 × 3.0 mm. Forty-four contiguous slices with a thickness of 3 mm were acquired in an interleaved order. A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

2.5 | fMRI data preprocessing

Analysis of the fMRI data was performed using SPM12 (Welcome Department of Imaging Neuroscience) implemented in Matlab (Math Works). To allow for T1 equilibration, the first four volumes were discarded before preprocessing and data analysis. The SPM12 realignment program was used to correct for head motion. Following realignment, the volumes were normalized to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each individual participant to the EPI template using an affine transformation (resliced to a voxel size of 2.0 × 2.0 × 2.0 mm). The normalized data was spatially smoothed with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum). For MVPA, spatial smoothing was not applied so as to preserve fine-grained activation patterns.

2.6 | Behavioral analysis

Social conformity effect Multiple regression analysis was performed for each participant to investigate the effect of group rating on individual conformity (rating change). The following two predictor variables were included: (a) gap (group rating - participant's first rating) and (b) participant's first rating. The dependent variable was rating change (participant's second rating - first rating). The first rating was considered as one of the predictor variables to control for the regression-to-the-mean effect (Izuma & Adolphs, 2013; Wake et al., 2019). On the rare occasion (0.32% across all participants), seven participants pressed the decide button (right thump button) before the right or left key (i.e., they submitted their rating, most likely accidentally, when the cursor was still invisible). These missed trials were not included in the behavioral data analysis and were modeled as a regressor of no interest in the fMRI data analysis (see below).

Prediction error estimation To estimate prediction error signals in each trial of the reinforcement learning task, we fitted a standard *Q*-learning model (Sutton & Barto, 1998) to the participants' choice behaviors.

In the *Q*-learning model, in the choice phase of each trial, an agent chose an option (say *A*) over the other (say *B*) with the probability $q(\mathrm{A}) = 1/[1 + \exp(-\beta\ (Q(\mathrm{A}) - Q(\mathrm{B})))]$, where *Q* denotes the value of each option and β denotes the degree of stochasticity in the choices (called inverse temperature). In the outcome phase, the agent updated the value of the chosen option based on reward experience. Suppose that the option *A* is chosen, then the value is updated by the reward prediction error $\delta = R - Q(A)$, where *R* denotes the reward outcome (coded 1 for reward and 0 for no reward) as follows: $Q(\mathrm{A}) \leftarrow Q(\mathrm{A}) + \alpha\delta$. Here, the parameter α is the learning rate. In the first trial of the task, option values were set at 0.5 (as the agent seemed to have no prior belief in the reward probabilities).

We fitted this model to each participant's choice data. In the model fitting, to avoid any unreasonable individual fits (Niv, Edlund, Dayan, & O, 2012), we employed a maximum a posteriori (MAP) approach in which the learning rate was constrained to a range of 0 to 1 with a Beta (2,2) prior distribution and the inverse temperature was constrained to be positive with a Gamma (2,3) prior distribution.

2.7 | fMRI data analysis: Univariate analysis

Two general linear models (GLMs) were used to analyze the fMRI data. The first GLM was set up to assess brain activation correlated to the absolute gap (the difference between a participant's first rating and the group rating) in the social conformity task. The second GLM was set up to assess brain activation correlated to signed and unsigned prediction error values in the reinforcement learning task.

In the first GLM (social conformity task analysis), the absolute gap in the social conformity task was quantified by calculating the absolute difference between the individual and group ratings in each trial. A parametric modulation analysis was performed to assess the correlation between trial-by-trial absolute gap scores and brain activation. The model included the following three regressors: (a) trial regressor (onset = trial onset, duration = subject's response time), (b) feedback regressor (onset = feedback onset, duration = 2 s), and (c) feedback regressor modulated by absolute gap between individual and group ratings. As stated above, missed trials were modeled as an additional regressor of no interest for the seven participants.

In the second GLM (reinforcement learning task analysis), signed and unsigned prediction errors were quantified using the computational model described above. A parametric modulation analysis was performed to assess the correlation between trial-by-trial signed/unsigned prediction error and brain activation. The model included the following four regressors: (a) trial regressor (onset = trial onset, duration = subject's response time), (b) feedback regressor (onset = feedback onset, duration = 2 s), (c) feedback regressor modulated by signed prediction error values, and (d) feedback regressor modulated by unsigned prediction error values. If there were missed trials, they were separately modeled as a regressor of no interest.

In both GLMs, the regressors were calculated using a box-car function convolved with a hemodynamic-response function. Other regressors of no interest, such as six motion parameters, session effect, and high-pass filtering (128 s), were also included.

We aimed to follow up findings based on the univariate analyses (i.e., activation overlaps) with MVPA to further test the hypothesis. To avoid false negative results at the initial univariate analysis stage, we used a statistical threshold of $p < .005$ voxelwise (uncorrected for multiple comparisons) with a cluster size of 20 voxels within the three anatomical regions of interest (ROIs, see below) (Lieberman & Cunningham, 2009). Outside the ROIs, the statistical threshold was set at $p < .001$ voxelwise (uncorrected) and cluster $p < .05$ (FWE corrected for multiple comparisons). All reported p-values for both behavioral and fMRI data analyses were based on one-tailed tests.

2.8 | fMRI data analysis: MVPA

Correlation-based MVPA When the univariate analyses revealed activation overlaps, we further investigated if activation patterns in each of the overlapped regions were similar between social conflict and signed/unsigned prediction errors (Figure 2a). In this correlation-based MVPA, the following four contrast images were used: (a) those representing positive sensitivity to absolute gaps between individual and group ratings, (b) those representing negative sensitivity to absolute gaps between individual and group ratings, (c) those representing positive sensitivity to unsigned prediction error values, and (d) those representing positive sensitivity to signed prediction error values. We calculated voxel-by-voxel correlations between contrast images 1 and 3 (representing positive social conflict and unsigned prediction error, respectively) for each of the overlapped regions within the pMFC and insula. Similarly, we calculated voxel-by-voxel correlations between contrast images 2 and 4 (representing negative social conflict and signed prediction error, respectively) for each of the overlapped regions within the striatum. These within-subject correlation values were Fisher-z-transformed and submitted to a one-sample t test to test for significantly positive correlation. A positive correlation indicates that the pattern of voxelwise sensitivity to social conflict is similar to the pattern of voxelwise sensitivity to signed and unsigned prediction errors, thus providing a support for the hypothesis.

Classifier-based MVPA While the correlation-based MVPA described above assessed the similarities in activation (or sensitivity) between social conflict and signed/unsigned prediction errors, the aim of classifier-based MVPA was to determine if the two patterns were significantly distinct. We used a linear support vector machine, which was performed using Matlab in combination with LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) (Wake & Izuma, 2017), with a cost parameter of c = 1 (default). For this analysis, separate contrast images were created for each of the two fMRI runs, and classification performances were evaluated using a leave-one-run-out cross-validation procedure. Thus, using the contrast images from the first run of each task, we trained a classifier that discriminates activation patterns between social conflict and signed/unsigned prediction error. Then, using the contrast images from the second run of each task, we tested if the classifier could discriminate between social conflict and signed/unsigned prediction error (Figure 2b). The procedure was repeated using data from the second run as training data and data from the first run as test data. Two classification accuracy values were averaged for each participant and the average classification accuracy values were submitted to a Wilcoxon signed-rank test to determine if the classification accuracy was significantly higher than the theoretical chance level (i.e., 50%; note that we also conducted permutation tests [1,000 times] to estimate the empirical chance level in each ROI, but the results were virtually the same). Significantly high classification accuracy indicates that the pattern of voxelwise sensitivity to social conflict is distinct from that to signed/unsigned prediction error.

Searchlight analysis Further, we performed searchlight analysis (Kriegeskorte, Goebel, & Bandettini, 2006) to more thoroughly depict the activation profiles of each local region within the pMFC, insula, and striatum using the correlation-based and classifier-based MVPA procedures described above. We used a radius of three voxels so that each searchlight included a maximum of 123 voxels (and less voxels at the boundaries of each ROI). In each searchlight, a correlation between social conflict and unsigned/signed prediction error was computed for the correlation-based MVPA, and classification accuracy was computed for the classifer-based MVPA. The correlation maps and classification accuracy maps were entered into a second-level permutation-based analysis

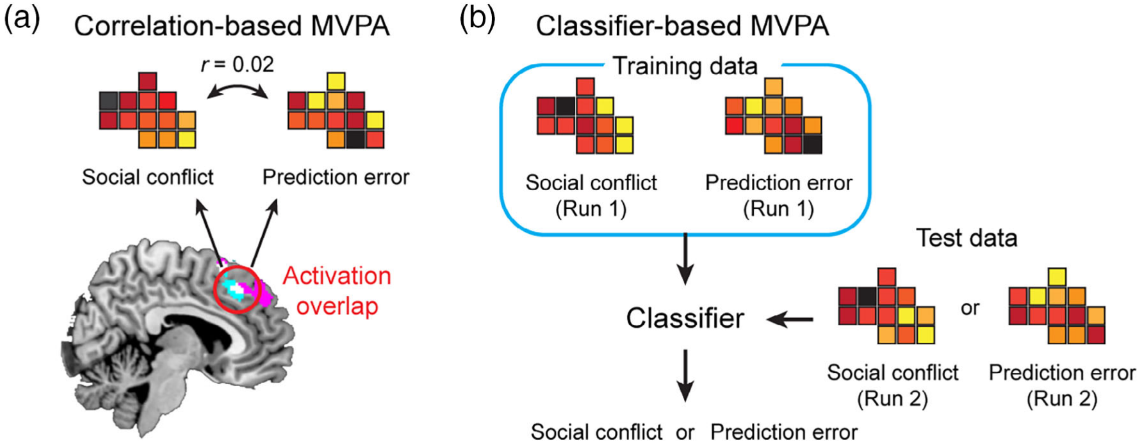

FIGURE 2 Schematic illustrations of two types of MVPA. (a) Correlation-based MVPA tests if the two patterns are significantly similar. (b) Classifier-based MVPA tests if the two patterns are significantly distinct

with 5,000 permutations using the Statistical NonParametric Mapping toolbox for SPM (Nichols & Holmes, 2002). A statistical threshold (i.e., voxel level) was set at $p < .005$ and a cluster-level threshold was set at $p < .0125$ (FWE corrected; separate second-level analyses were conducted for each of the four ROIs so that the cluster-level threshold was further corrected for four comparisons).

2.9 | Regions of interest

Anatomical ROIs To test the hypothesis, we focused on the following three anatomical ROIs: (a) pMFC, (b) insula, and (c) striatum. In a recent meta-analysis, it was reported that these ROIs were consistently positively associated with social conflict (pMFC and insula) and consistently negatively associated with social conflict (striatum) (Wu et al., 2016). These ROIs were also consistently associated with signed/unsigned prediction errors (Fouragnan et al., 2018). These anatomical ROIs were defined using a WFU pickatlas toolbox for SPM (dilation factor = 2) (Maldjian, Laurienti, Kraft, & Burdette, 2003). The pMFC included the superior frontal gyrus (*Frontal_Sup_Medial*), anterior cingulate cortex (ACC), middle cingulate cortex, and supplementary motor area (SMA). The striatum ROI included the caudate nucleus, putamen, and globus pallidus. We tested if the same areas within the pMFC and insula ROIs were activated by social conflict (positive correlation) and unsigned prediction error. Similarly, we tested if the same areas within the striatum ROI were activated by social conflict (negative correlation) and signed prediction error.

Functional ROIs For the subsequent MVPAs, we defined functional ROIs as overlapped clusters between social conflict and unsigned prediction error in the pMFC and insula anatomical ROIs and overlapped clusters between social conflict and signed prediction error in the striatum anatomical ROIs with a threshold of $p < .005$ with more than 20 voxels. For the searchlight MVPA, we used the same definition of functional ROIs above but with a more lenient threshold of $p < .05$ (uncorrected).

3 | RESULTS

3.1 | Behavioral results

On average, participants took 5.14 s (*SD* = 1.65) to rate a face in the social conformity task. During the reinforcement learning task, participants selected a slot machine within 2 s for most trials (average number of missed trials = 0.2) and the average reaction time was 0.67 s (*SD* = 0.18).

Consistent with the previous works, we found a significant conformity effect; the second ratings of the participants were significantly influenced by the group rating even after the regression-to-the-mean effect was controlled ($t[24] = 2.18$, $p = .02$, $d = 0.43$). We also found highly significant regression-to-the-mean effect ($t[24] = -15.81$, $p < .001$, $d = 3.16$), which is consistent with our previous studies (Izuma & Adolphs, 2013; Wake et al., 2019).

During the reinforcement learning task, participants selected options with a higher reward probability 59.9% of the trials, which is significantly higher than the chance (50%; $t[24] = 6.22$, $p < .001$, $d = 1.24$), indicating that the participants were generally able to accurately keep track of the fluctuating reward probabilities based on reward outcomes.

Further, our data showed that the reinforcement learning model explains participant behavior better than a model that assumes that an individual selects the right option with a fixed probability (p). To compare the goodness-of-fit of the models, we computed the Laplace approximated log model evidence (MacKay, 2003) of the two models. The values were compared using the Bayesian Model Selection (BMS) method from the study by Stephan, Penny, Daunizeau, Moran, and Friston (2009), which treats model identity as a random effect. Exceedance probabilities from this analysis indicated that the reinforcement learning model has a 100% chance of being the more common of the two models in the population.

Finally, we computed the across-subject correlation between the social conformity effects (beta values from the multiple regression analyses) and learning rate parameters (α estimated using the reinforcement learning model), but they did not correlate with each other ($r[23] = -.09$, $p = .66$). This indicates that individual differences in the susceptibility to social influence during the social conformity task are unrelated to individual differences in the sensitivity to reward outcome during the reinforcement learning task.

3.2 | Univariate results

We first successfully replicated the findings of previous studies on social conformity. The pMFC (dmPFC [dorsomedial prefrontal cortex] and pre-SMA [presupplementary motor area]) and bilateral anterior insula activities were positively correlated with social conflict (i.e., absolute difference between participant and group ratings) (Figure 3a), whereas the striatum activities were negatively correlated with social conflict (Figure 3b and Table 1). All activated areas outside the ROIs are listed in Table 2.

We also successfully replicated the findings of previous studies on reinforcement learning. During the reinforcement learning task, the pMFC and bilateral anterior insula activities were positively correlated with unsigned prediction error (Figure 3c), whereas the striatum activities were positively correlated with signed prediction error (Figure 3d and Table 3). All areas outside the ROIs that were significantly positively related to unsigned or signed prediction error are listed in Table 4 (note that no area was significantly negatively related to signed or unsigned prediction error).

Consistent with the hypothesis, in each of the three anatomical ROIs (i.e., the pMFC, insula, and striatum), there were a total of seven activation overlaps (18–158 voxels; Table 5). We found two overlapped clusters in the pMFC; one in the dmPFC and the other in the pre-SMA (Figure 3e and Table 5). These overlapped areas in the pMFC and bilateral insula were sensitive to social conflict (positively related) and unsigned prediction error. Similarly, we found three

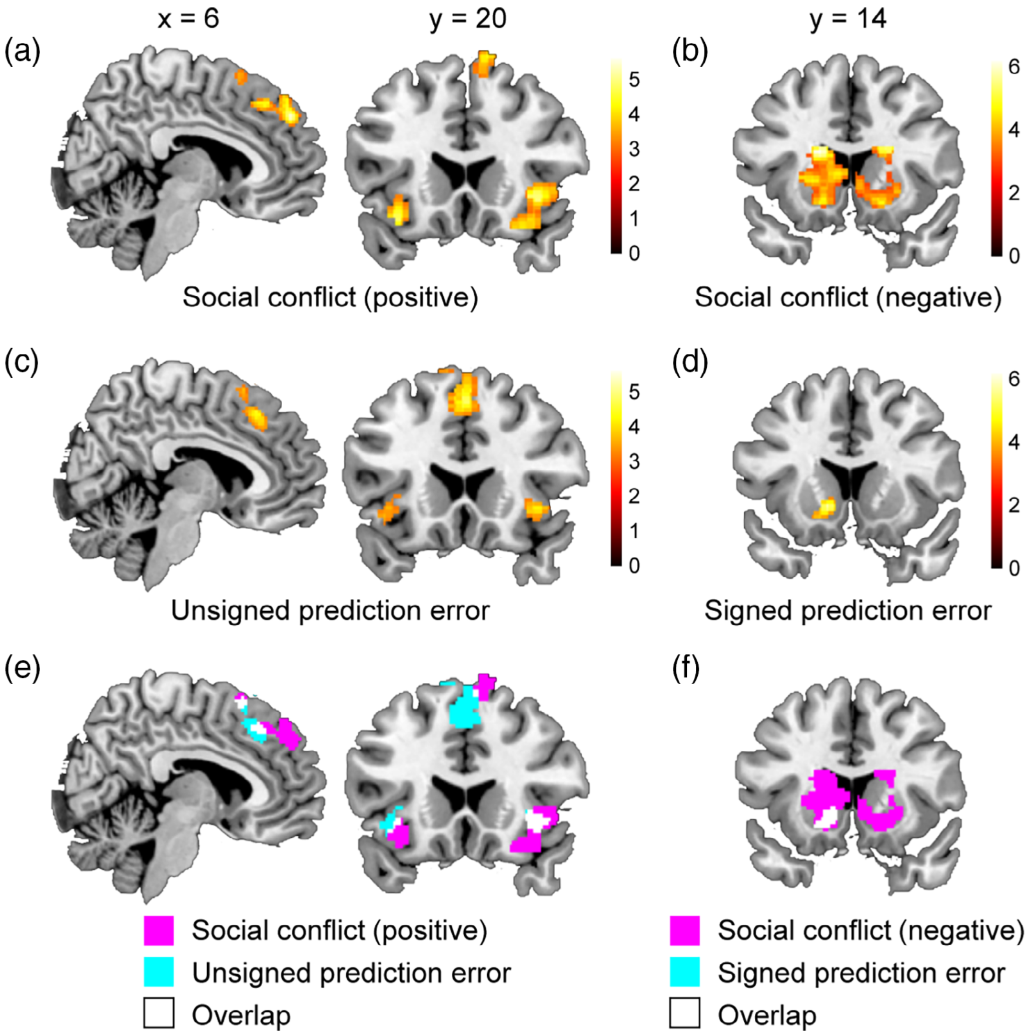
x = 6
y = 20
y = 14
(a)
(b)
5 4 3 2 1 0
6 4 2 0
Social conflict (positive)
Social conflict (negative)
(c)
(d)
5 4 3 2 1 0
6 4 2 0
Unsigned prediction error
Signed prediction error
(e)
(f)
🟪 Social conflict (positive)
🟦 Unsigned prediction error
□ Overlap
🟪 Social conflict (negative)
🟦 Signed prediction error
□ Overlap

FIGURE 3 fMRI results from univariate analyses. (a) pMFC and insula regions positively related to social conflict (i.e., absolute difference between participant and group ratings). (b) Striatum regions negatively related to social conflict. (c) pMFC and insula regions positively related to unsigned prediction error. (d) Striatum regions positively related to signed prediction error. (e) Activation overlaps between social conflict (panel a) vs. unsigned prediction error (panel c) related regions. (f) Activation overlaps between social conflict (panel b) vs. unsigned prediction error (panel d) related regions

separate overlapped clusters within the striatum (Figure 3f and Table 5), and these areas were sensitive to social conflict (negatively related) and signed prediction error.

3.3 | MVPA results

3.4 | ROI-based analyses

Correlation-based MVPA To obtain more compelling evidence of a common neural mechanism between social conformity and reinforcement learning, we conducted correlation-based MVPA which investigates whether social conflict and signed/unsigned prediction error evoked similar activation patterns in each of the seven overlapped areas (Table 5). A high correlation indicates that voxels sensitive to social conflict are also sensitive to reward prediction error and therefore supports the hypothesis. However, there were no significant correlations in any of the seven overlapped areas. The average correlations in the four overlapped areas in the pMFC and insula were not significantly positive even at p < .05 uncorrected level (all ps > .60; Table 6). Similarly, all three overlapped regions in the striatum showed nonsignificant correlation (all ps > .41; Table 6). These results suggest that, although the same brain regions are involved in social conformity and reinforcement learning, the underlying neural populations may be distinct.

Classifier-based MVPA We further attempted to find more direct evidence that refutes the hypothesis and conducted classifier-based MVPA to determine whether activation patterns are distinct between social conflict and signed/unsigned prediction error (i.e., whether a classifier is able to distinguish patterns associated with social conflict and reward prediction error). The dmPFC cluster showed significant classification accuracy, which indicates that the patterns evoked in the dmPFC by social conflict and unsigned prediction error are distinct. Classification

TABLE 1 ROI activation during the social conformity task

	MNI coordinate				
Location	**x**	**y**	**z**	**Z**	**Cluster size**
Areas in the pMFC and insula positively related to social conflict					
dmPFC	8	48	38	4.26	337
Right anterior insula	38	22	-4	4.07	388
Left anterior insula	–34	20	–16	3.76	128
Pre-SMA	12	20	68	3.48	191
Areas in the striatum negatively related to social conflict					
Left striatum (caudate body)	–18	–8	20	5.13	1,714
Left caudate tail	–14	–20	22	4.55	
Left putamen	–32	–16	6	4.22	
Left caudate head	–18	22	14	4.18	
Left NAcc	–18	10	–8	4.61	
Right striatum (caudate head)	14	28	0	4.85	1,682
Right caudate body	20	16	20	4.53	
Right putamen	36	–8	–4	4.3	
Right caudate tail	22	–24	22	4.27	
Right NAcc	16	10	–12	3.99	

Abbreviations: dmPFC, dorsomedial prefrontal cortex; NAcc, nucleus accumbens; pMFC, posterior medial frontal cortex; pre-SMA, presupplementary motor area.

TABLE 2 Activations outside the ROIs during the social conformity task

	MNI coordinate				
Location	**x**	**y**	**z**	**Z**	**Cluster size**
Areas positively related to social conflict					
No significant region					
Areas negatively related to social conflict					
Left IPL	–58	–30	38	5.39	6,834
Left paracentral lobule	–12	–20	60	4.83	
Posterior cingulate cortex	0	–34	42	4.66	
Left lateral prefrontal cortex	–42	50	12	4.88	499
Left inferior temporal gyrus	–54	–58	–8	4.5	751
Right cerebellum	46	–70	–44	4.2	240
Right posterior insula	38	–8	–4	4.13	362
Right STG	62	–18	–2	4.13	
Left DLPFC	–36	34	–32	3.92	206
Left MFG	–22	–24	52	3.86	262
Left STG	–62	–28	4	3.51	204

Abbreviations: DLPFC, dorsolateral prefrontal cortex; IPL, inferior parietal lobule; MFG, middle frontal gyrus; STG, superior temporal gyrus.

accuracies in the right insula, left insula, and pre-SMA were not significant (although classification accuracy in the left insula was significant at p < .05 uncorrected level; Table 6). Similarly, the right putamen cluster showed significant classification accuracy, which indicates that the patterns evoked in this area by social conflict and signed prediction error were distinct. Classification accuracies in the left putamen and left nucleus accumbens (NAcc) were not significant (Table 6).

Overall, our ROI based MVPA analysis did not find any support for the hypothses. On the contrary, we found evidence refuting the hypothesis, especially in the dmPFC and the right putamen, whereas

TABLE 3 ROI activation during the reinforcement learning task

	MNI coordinate				
Location	x	y	z	Z	Cluster size
Areas in the pMFC and insula positively related to unsigned prediction error					
mPFC	16	64	2	4.17	55
Pre-SMA/dmPFC	−2	16	54	3.99	632
Left anterior insula	−44	16	−10	3.55	89
Right anterior insula	38	18	−4	3.30	96
Areas in the striatum positively related to signed prediction error					
Left putamen	−34	−4	2	4.10	183
Right putamen	32	−6	14	3.89	146
Left NAcc	−12	14	−4	3.85	224

Abbreviations: dmPFC, dorsomedial prefrontal cortex; mPFC, medial prefrontal cortex; NAcc, nucleus accumbens; pMFC, posterior medial frontal cortex; pre-SMA, presupplementary motor area.

TABLE 4 Activations outside the ROIs during the reinforcement learning task

	MNI coordinate				
Location	x	y	z	Z	Cluster size
Areas positively related to unsigned prediction error					
Right IPL	50	−40	48	5.28	2,706
Right inferior temporal gyrus	58	−30	−22	5.03	491
Left lateral prefrontal cortex	−28	46	10	4.95	175
Left IPL	−38	−46	42	4.45	653
Right DLPFC	46	32	42	4.18	505
Right VLPFC	40	58	−10	3.74	
Areas positively related to signed prediction error					
mPFC	6	62	8	4.55	2,420
dmPFC	−14	42	54	4.39	
ACC	−2	44	2	4.24	
vmPFC	−6	48	−14	3.41	
Left lingual gyrus	−20	−90	6	4.52	424
Right lingual gyrus	28	−86	10	4.39	891

Note: The mPFC cluster positively related to signed prediction error did not overlap with the dmPFC region related to social conflict.
Abbreviations: ACC, anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; IPL, inferior parietal lobule; mPFC, medial prefrontal cortex; VLPFC, ventrolateral prefrontal cortex; vmPFC, ventromedial prefrontal cortex.

there was no clear evidence for or against the hypothesis in the pre-SMA, left and right insula, left putamen, and left NAcc.

3.5 | Searchlight analysis

Although we conducted MVPA analyses on each of the seven overlapped clusters (Table 6), the size of an overlapped cluster depends on the thresholds and smoothing kernels (e.g., see Deen, Koldewyn, Kanwisher, & Saxe, 2015), and this makes our choice of functional ROIs somewhat arbitrary. Therefore, we conducted searchlight analysis to more thoroughly depict the activation profiles of each local area within the pMFC, insula, and striatum. We defined the functional ROIs (i.e., univariate activation overlaps) more broadly by using a $p < .05$ (uncorrected) threshold in each of the three anatomical ROIs. This procedure revealed 917 overlapped voxels in the pMFC, 369 in the right insula, 199 in the left insula, and 3,054 across three separate clusters in the striatum (Figure 4a).

We performed searchlight MVPA of each of the overlapped clusters to determine if any area showed similar activation patterns between social conflict and prediction error (i.e., correlation-based MVPA). However, we found no such areas in any of the ROIs. We also

TABLE 5 Overlapped activations between social conflict and signed/unsigned prediction error

		Peak MNI coordinate		
	Size of overlap (voxel)	x	y	Z
Overlap between areas positively related to social conflict and areas positively related to unsigned prediction error				
dmPFC	27	8	28	44
Pre-SMA	26	6	16	62
Right anterior insula	65	38	22	–4
Left anterior insula	18	–36	20	–8
Overlap between areas negatively related to social conflict and areas positively related to signed prediction error				
Right posterior putamen	66	32	–12	2
Left posterior putamen	74	–30	–12	6
Left NAcc	158	–18	10	–8

Note: The peak MNI coordinates reported here are based on the social conflict contrasts. There were two more overlapped clusters (both in the left putamen), but these clusters consist of less than three voxels and were not investigated further.
Abbreviations: dmPFC, dorsomedial prefrontal cortex; pre-SMA, presupplementary motor area; NAcc, nucleus accumbens.

TABLE 6 MVPA results

		Correlation-based MVPA		Classifier-based MVPA	
Regions	ROI size (voxel)	Average correlation	*p*-value (uncorrected)	Average classification accuracy (%)	*p*-value (uncorrected)
Overlap between areas positively related to social conflict and areas positively related to unsigned prediction error					
dmPFC	27	–0.02	0.62	59	.004*
Pre-SMA	26	–0.02	0.61	56	.078
Right insula	65	–0.03	0.70	52	.250
Left insula	18	–0.01	0.60	56	.031
Overlap between areas negatively related to social conflict and areas positively related to signed prediction error					
Right putamen	66	–0.02	0.72	56	.016*
Left putamen	74	0.01	0.41	51	.500
Left NAcc	157	–0.02	0.72	52	.375

*p < .05 (Bonferroni correction). Significant results of correlation-based MVPA mean that activation patterns are similar between social conflict and prediction error (i.e., evidence supporting the hypothesis), while those of classifier-based MVPA mean that they are distinct (i.e., evidence against the hypothesis).

performed classifier-based MVPA using the same searchlight procedure and found a total of nine significant clusters across the four ROIs (Figure 4b and Table 7). Thus, no evidence of a shared neural mechanism between social conformity and reinforcement learning was found in the searchlight analysis. On the contrary, searchlight analysis found that the patterns between social conformity and reinforcement learning in several local areas within each ROI were significantly distinct.

4 | DISCUSSION

The present study investigated whether social conformity and reinforcement learning have a common neural mechanism. The behavioral results showed a robust social conformity effect during the social conformity task and that the reinforcement learning model can adequately explain the behavior of participants during the reinforcement learning task. Furthermore, univariate fMRI analyses successfully replicated the findings of earlier studies that reported the involvement of the pMFC, insula, and striatum in the processing of prediction error signals and social conflict signals (i.e., the difference between one's and group ratings) during the reinforcement learning task and social conformity task, respectively. Further, we found that, in the pMFC and anterior insula, areas positively related to social conflict (i.e., absolute difference between participant and group ratings) overlapped with areas sensitive to unsigned prediction error. We also found that areas of the striatum negatively related to social conflict overlapped with areas of the striatum sensitive to signed prediction

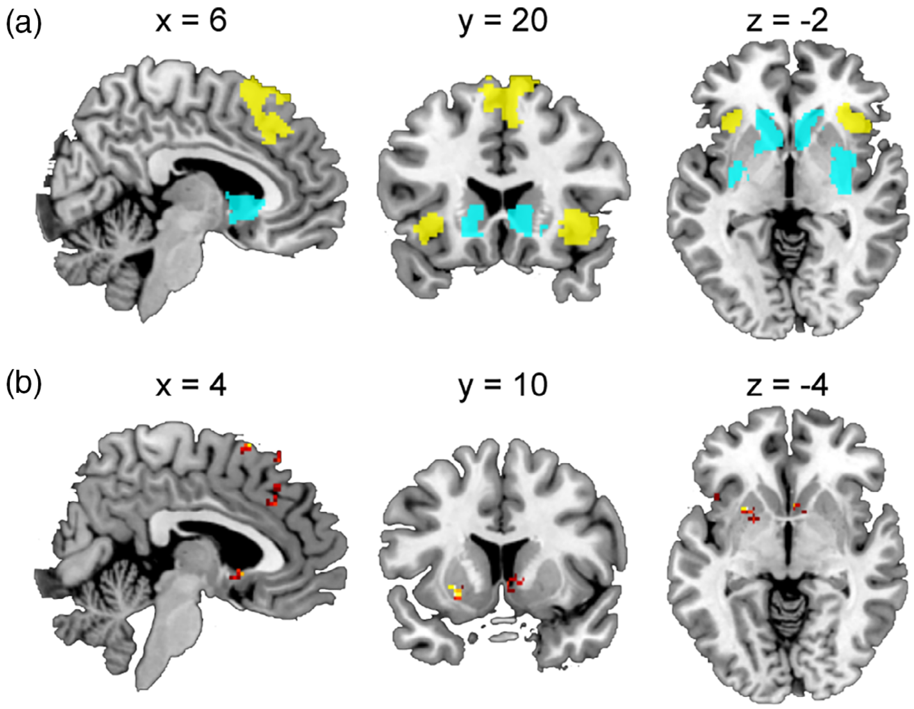

FIGURE 4 Searchlight MVPA. (a) Functional ROIs used in the searchlight MVPAs. Yellow color denotes regions sensitive to social conflict (positively related) and unsigned prediction error. Cyan color denotes regions sensitive to social conflict (negatively related) and signed prediction error. (b) Searchlight MVPA results (classifier-based MVPA). Each panel depicts areas that showed significantly distinct activation patterns between social conflict and signed/unsigned prediction error

TABLE 7 Searchlight MVPA results

	MNI coordinate				
Location	x	y	z	Cluster size	Cluster *p*-value
Correlation-based MVPA					
No significant region					
Classifier-based MVPA					
pMFC ROI					
dmPFC	16	24	60	54	<.001
dACC	–2	30	38	42	<.001
Pre-SMA	4	16	66	12	.009
Right insula ROI					
Right anterior insula	44	20	–12	5	.001
Left insula ROI					
Left anterior insula 1	–28	24	0	9	.001
Left anterior insula 2	–42	16	–4	2	.012
Striatum ROI					
Left posterior putamen	–24	10	–4	29	<.001
Right NAcc	4	12	–4	24	.005
Left anterior putamen	–32	2	4	30	.003

error. This is the first unequivocal evidence that social conflict and signed/unsigned prediction error activate the same areas in the pMFC, insula, and striatum.

However, follow-up MVPA did not provide evidence of a common neural mechanism. It revealed that patterns of sensitivity to social conflict were not similar to patterns of sensitivity to unsigned/signed prediction error in any of the overlapped areas. Although this negative result could be explained by the high noise of the data, classifier-based MVPA could successfully distinguish between social conflict vs. unsigned prediction error related activation patterns in the pMFC, and this is evidence of distinct (nonsimilar) neural mechanisms in social conformity and reinforcement learning at least within the pMFC. Searchlight analyses further confirmed these results, and there was no evidence of a common neural mechanism in any of the ROIs. On the contrary, overall pictures of the searchlight results show largely distinct, rather than common, activation patterns in social conflict and signed/unsigned prediction error. Thus, despite our efforts to minimize the possibility of false negatives, we found no evidence to support the hypothesis that social conformity and reinforcement learning have a common neural mechanism. These results suggest that the reinforcement learning hypothesis may be too simplistic to explain the neural mechanism of social conformity.

Based on the theoretical framework provided by Lockwood et al. (2020), our results showed that social conformity and reinforcement learning are different at least at the implementational level. Furthermore, we speculate that they may be different even at the algorithmic and/or computational levels. For example, while activities in the pMFC and insula may reflect the degree of surprise (or the absolute deviation from expectations) during the reinforcement learning task, they may express a negative feeling (e.g., unpleasantness or anxiety due to the recognition that one is different from others) during the social conformity task. This in turn motivates the individual to reduce the negative feeling by conforming to group opinion. It is unlikely that individuals think their rating will always be the same as the group rating (see Izuma & Adolphs, 2013). In other words, the degree of social conflict is not related to the degree of surprise (deviation from expectations). This seems to be a critical difference between social conflict in the social conformity paradigm and reward prediction error (including various forms of social prediction error (Suzuki & O'Doherty, 2020). Furthermore, the notion that the activities in the pMFC and insula during the social conformity task reflect negative emotion is consistent with reports of previous studies stating that these regions also play a role in cognitive dissonance (Izuma et al., 2010; van Veen, Krug, Schooler, & Carter, 2009) (for reviews, see Izuma, 2013; Izuma & Murayama, 2019), which is considered to be a negative feeling caused by inconsistency between attitude and behavior (Festinger, 1957). This notion is also supported by our previous finding that agreement, rather than disagreement, with a disliked group activated the pMFC and insula (and disagreement with a disliked group activated the striatum) (Izuma & Adolphs, 2013).

Similarly, activity in the striatum, which is negatively related to social conflict, may reflect a positive subjective feeling (Campbell-Meiklejohn et al., 2010; Nook & Zaki, 2015), which results from the realization that the group has the same opinion as the participant. Further, activity related to signed prediction error reflects a learning signal but not a positive subjective feeling. It is well known that striatum activity tracks subjective pleasantness induced by various stimuli such as faces and foods (e.g., Izuma et al., 2010; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009). However, it should also be noted that the ventromedial prefrontal cortex (vmPFC) is also known to be robustly related to subjective pleasantness (e.g., Ito et al., 2015; Lebreton et al., 2009; Suzuki, Adachi, Dunne, Bossaerts, & O, 2015; Suzuki, Cross, & O, 2017), but we did not find any activation negatively related to social conflict in the vmPFC.

Although the present study used an experimental paradigm similar to that of the social conformity study that originally proposed the reinforcement learning hypothesis of social conformity (Klucharev et al., 2009), it should be noted that social conformity is not a unitary phenomenon and that some forms of social conformity may be more similar to reinforcement learning. At least three types of motivation for attitude change based on social influence have been identified in psychological studies (Cialdini & Goldstein, 2004; Petty & Cacioppo, 1981), and they include the following: (a) motivation to be accurate, (b) motivation to obtain social approval from others, and (c) motivation to maintain a positive self-concept (which includes attitude change following cognitive dissonance). The majority of previous neuroimaging studies on social conformity used the face rating task (or a similar task which involves subjective ratings of stimuli) (Izuma, 2013). The conformity effect reported in these studies can be explained by the motivation to obtain social approval and/or the motivation to maintain a positive self-concept, but it cannot be explained by the motivation to be accurate as there is no right or wrong answer in facial attractiveness rating. In a situation where individuals strongly believe that group opinion is more accurate than their own opinion (e.g., group opinion ≈ correct performance feedback), social conflict can serve as a teaching signal just like prediction error in the reinforcement learning task. In fact, the pMFC, insula, and striatum are related to prediction error in a semantic learning paradigm where no reward is involved (i.e., acquiring new semantic knowledge based on performance feedback [correct or incorrect]) (Pine, Sadeh, Ben-Yakov, Dudai, & Mendelsohn, 2018). In this study, prediction error in each trial was calculated based on a subjective rating of confidence and on feedback received by participants (e.g., a positive prediction error signal is generated when individuals were not confident about their answer but received a correct feedback). Further, activities in the pMFC and insula were found to be positively related to unsigned prediction error, whereas striatum activities were found to be positively related to signed prediction error (Pine et al., 2018). Thus, how these brain regions process social conflict might be more similar to reinforcement learning in a different social conformity paradigm where a right answer can be objectively defined (i.e., where social conflict can serve as a strong teaching signal), and this is an important avenue for future research.

Finally, this study highlights the importance of directly comparing two tasks (cognitive processes) with the same sample of participants and the utility of the multivariate approach in interpreting univariate activation overlaps (Peelen & Downing, 2007). Although previous EEG studies (Chen et al., 2012; Kim et al., 2012; Shestakova et al., 2012) report the finding of a signal over the pMFC during the social conformity task that resembles the FRN signal found in the reinforcement learning task, it is important to compare these signals with the same participants to clearly determine if they are similar in terms of spatial location and timing.

In conclusion, this study investigated the reinforcement learning hypothesis of social conformity, which states that social conformity and reinforcement learning have a common neural mechanism. Using the representative tasks of social conformity and reinforcement learning, we found that the pMFC, bilateral anterior insula, and striatum were involved in processing reward prediction error and social conflict, which is consistent with the reports of previous studies. However, MVPA failed to find any clear evidence of a shared neural mechanism. Thus, our results suggest that the reinforcement learning hypothesis is likely to be too simplistic and caution against proposing a common neural mechanism based on univariate activation overlaps.

ACKNOWLEDGMENTS

We thank Christopher Watson, Chris Everitt, Angela Darekar, and other members of the MRI unit at Southampton University Hospital for their help with fMRI data collection. We also thank Dr. Vasily Klucharev for providing us with the face stimulus set and Dr. Ryuta Aoki for helpful comments on the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICAL STATEMENT

The study was approved by the University of Southampton ethics committee

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

ORCID

Ayahito Ito https://orcid.org/0000-0001-5217-7340
Keise Izuma https://orcid.org/0000-0003-0256-3571

REFERENCES

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, *20*(13), 1165–1170. https://doi.org/10.1016/j.cub.2010.04.055

Campbell-Meiklejohn, D. K., Simonsen, A., Jensen, M., Wohlert, V., Gjerloff, T., Scheel-Kruger, J., … Roepstorff, A. (2012). Modulation of social influence by methylphenidate. *Neuropsychopharmacology*, *37*, 1517–1525. https://doi.org/10.1038/npp.2011.337

Chen, J., Wu, Y., Tong, G. Y., Guan, X. M., & Zhou, X. L. (2012). ERP correlates of social conformity in a line judgment task. *BMC Neuroscience*, *13*, 43. https://doi.org/10.1186/1471-2202-13-43

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. https://doi.org/10.1146/Annurev.Psych.55.090902.142015

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*(6), 941–951. https://doi.org/10.1162/jocn.2008.20062

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience*, *24*(1), 106–118. https://doi.org/10.1162/jocn_a_00114

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Fouragnan, E., Retzler, C., Mullinger, K., & Philiastides, M. G. (2015). Two spatiotemporally distinct value systems shape reward-based learning in the human brain. *Nature Communications*, *6*, 8107. https://doi.org/10.1038/ncomms9107

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. https://doi.org/10.1002/hbm.24047

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis. Of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295x.109.4.679

Ito, A., Abe, N., Kawachi, Y., Kawasaki, I., Ueno, A., Yoshida, K., … Fujii, T. (2015). Distinct neural correlates of the preference-related valuation of supraliminally and subliminally presented faces. *Human Brain Mapping*, *36*(8), 2865–2877. https://doi.org/10.1002/hbm.22813

Izuma, K. (2013). The neural basis of social influence and attitude change. *Current Opinion in Neurobiology*, *23*(3), 456–462. https://doi.org/10.1016/j.conb.2013.03.009

Izuma, K. (2017). The neural bases of social influence on valuation and behavior. In J.-C. Dreher & L. Tremblay (Eds.), *Decision neuroscience: An integrative approach* (pp. 199–209). Cambridge, MA: Academic Press.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573. https://doi.org/10.1016/j.neuron.2013.03.023

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*(51), 22014–22019. https://doi.org/10.1073/Pnas.1011879108

Izuma, K., & Murayama, K. (2019). The neural basis of cognitive dissonance. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Progress on apivotal theory in social psychology* (2nd ed.). Washington, DC: American Psychological Association.

Kim, B. R., Liss, A., Rao, M., Singer, Z., & Compton, R. J. (2012). Social deviance activates the brain's error-monitoring system. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 65–73. https://doi.org/10.3758/S13415-011-0067-5

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Klucharev, V., Munneke, M. A. M., Smidts, A., & Fernandez, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *The Journal of Neuroscience*, *31*(33), 11934–11940. https://doi.org/10.1523/Jneurosci.1869-11.2011

Korn, C. W., Fan, Y., Zhang, K., Wang, C., Han, S., & Heekeren, H. R. (2014). Cultural influences on social feedback processing of character traits. *Frontiers in Human Neuroscience*, *8*, 192. https://doi.org/10.3389/fnhum.2014.00192

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*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. https://doi.org/10.1016/j.neuron.2009.09.040

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and*

Affective Neuroscience, *4*(4), 423–428. https://doi.org/10.1093/scan/nsp052

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is there a 'Social' brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239.

Marr, D. (1982). *Vision*, Cambridge, MA: MIT Press.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Nook, E. C., & Zaki, J. (2015). Social norms shift behavioral and neural responses to foods. *Journal of Cognitive Neuroscience*, *27*(7), 1412–1426. https://doi.org/10.1162/jocn_a_00795

Peelen, M. V., & Downing, P. E. (2007). Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends in Cognitive Sciences*, *11*(1), 4–5. https://doi.org/10.1016/j.tics.2006.10.009

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuation: Classic and contemporary approaches*. Dubuque, IA: Wm. C. Brown.

Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y., & Mendelsohn, A. (2018). Knowledge acquisition is governed by striatal prediction errors. *Nature Communications*, *9*, 1673. https://doi.org/10.1038/s41467-018-03992-5

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, *95*(3), 853–951. https://doi.org/10.1152/physrev.00023.2014

Shestakova, A., Rieskamp, J., Tugin, S., Ossadtchi, A., Krutitskaya, J., & Klucharev, V. (2012). Electrophysiological precursors of social conformity. *Social Cognitive and Affective Neuroscience*, *8*, 756–763. https://doi.org/10.1093/scan/nss064

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction. Adaptive computation and machine learning*. Cambridge, Massachusetts: MIT Press.

Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2015). Neural mechanisms underlying human consensus decision-making. *Neuron*, *86*(2), 591–602. https://doi.org/10.1016/j.neuron.2015.03.019

Suzuki, S., Cross, L., & O'Doherty, J. P. (2017). Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nature Neuroscience*, *20*(12), 1780–1786. https://doi.org/10.1038/s41593-017-0008-x

Suzuki, S., & O'Doherty, J. P. (2020). Breaking human social decision making into multiple components and then putting them together again. *Cortex*, *127*, 221–230. https://doi.org/10.1016/j.cortex.2020.02.014

van Veen, V., Krug, M. K., Schooler, J. W., & Carter, C. S. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, *12*(11), 1469–1474. https://doi.org/10.1038/nn.2413

Wake, S. J., Aoki, R., Nakahara, K., & Izuma, K. (2019). Elucidating the role of the posterior medial frontal cortex in social conflict processing. *Neuropsychologia*, *132*, 107124. https://doi.org/10.1016/j.neuropsychologia.2019.107124

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10), 1558–1564.

Woo, C. W., Koban, L., Kross, E., Lindquist, M. A., Banich, M. T., Ruzic, L., … Wager, T. D. (2014). Separate neural representations for physical pain and social rejection. *Nature Communications*, *5*, 5380. https://doi.org/10.1038/ncomms6380

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *71*, 101–111. https://doi.org/10.1016/j.neubiorev.2016.08.038

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, *22*(7), 894–900. https://doi.org/10.1177/0956797611411057

How to cite this article: Levorsen M, Ito A, Suzuki S, Izuma K. Testing the reinforcement learning hypothesis of social conformity. *Hum Brain Mapp*. 2021;42:1328–1342. https://doi.org/10.1002/hbm.25296