



‘The Last Shot’—the shared and distinct brain regions involved in processing unexpectedness of success and failure in the context of social cooperation

Peng Li, ^{1,*} Jing Wang,^{2,3} and Yi Liu ^{4,*}

¹Brain Function and Psychological Science Research Center, Shenzhen University, Shenzhen 518060, China

²School of Management, Shenzhen Polytechnic, Shenzhen 518055, China

³Department of Biomedical Sciences of Cells & Systems, University Medical Center Groningen, Groningen, AW 9713, the Netherlands

⁴School of Psychology, Northeast Normal University, No. 5268, Renmin Avenue, Changchun, Jilin 130024, China

*Correspondence should be addressed to Yi Liu, School of Psychology, Northeast Normal University, No. 5268, Renmin Avenue, Changchun, Jilin 130024, China.

E-mail: liuy930@nenu.edu.cn.

Abstract

Individual success and failure in social cooperation matter not only to oneself but also to teammates. However, the common and distinct neural activities underlying salient success and failure in social cooperation are unclear. In this functional magnetic resonance imaging (fMRI) study, participants in the social group (Experiment one) cooperated with two human beings during a dice-gambling task, whereas those in the nonsocial group (Experiment two) cooperated with two computers. The social group reported more pride in success and more guilt in failure. The fMRI results in Experiment one demonstrate that left temporoparietal junction (LTPJ) activation increased exclusively with linearly changing unexpected success, whereas increasing anterior cingulate cortex (ACC) activation was only coupled with increasing unexpectedness of failure. Moreover, the dorsal medial prefrontal cortex (dMPFC) and left anterior insula were recruited in both success and failure feedback conditions. Dynamic causality model analysis suggested that the dMPFC first received information from the LTPJ and ACC separately and then returned information to these regions. The between-experiment comparison showed more dMPFC activity in social vs nonsocial contexts irrespective of success and failure feedback. Our findings shed light on the common and distinct neural substrates involved in processing success and failure feedback in social cooperation.

Key words: social cooperation; feedback learning; success; failure; dorsal medial prefrontal cortex

Introduction

Michael Jordan, the famous basketball player, led his team to several last-minute victories in the National Basketball Association (NBA) courtside. Not as lucky as Michael, football players may sometimes miss a penalty kick in the last minute and cause their team to lose an important competition. Both examples illustrate different situations in which group outcome is impacted by individual success or failure when the social contexts predict reward in small or large probabilities, respectively. In such cases, it is truly important for individuals to evaluate the outcome of their actions with the aim of improving performance and contributing to the group when their performance affects the group's interests. Investigation of outcome evaluation in a social context not only improves our current understanding of the neural mechanism of domain-general reinforcement learning but also advances our knowledge of social consequences and emotional responses in interpersonal interaction (Koban and Pourtois, 2014).

In social contexts, when an individual's performance impacts the interests of others, neural responses to that individual's success or failure have been found to be more intense than in a nonsocial context (i.e. when individual performance matters only for self-interest) (Koban and Pourtois, 2014). In the reinforcement learning literature, the term ‘reward prediction error’ (RPE) (Schultz et al., 1997; Holroyd and Coles, 2002) describes the violation between expected and experienced outcomes. Studies on event-related potential (ERP) show that the feedback-related negativity (FRN) amplitude, which reflects RPE during outcome evaluation, is modulated by social contexts (Li et al., 2010; Koban et al., 2012; Beyer et al., 2017). For instance, Koban et al. (2012) found that feedback in social cooperation elicited a larger FRN amplitude than that in competition. Similarly, in our previous ERP study, we found that feedback elicited a larger FRN in a condition of concentrated responsibility than that in the diffusion of responsibility condition in a three-person cooperative task (Li et al., 2010).

Received: 6 September 2021; Revised: 6 August 2022; Accepted: 13 August 2022

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Enhanced FRNs were also observed when participants performed tasks in the presence of others vs those performed in private (Huang and Yu, 2018). Another line of evidence came from the functional magnetic resonance imaging (fMRI) literature. Radke et al. (2011) compared brain region responses between errors that mattered only to the participant and those that mattered to others. They found that both types of errors activated the posterior medial frontal cortex and the anterior insula (AI), whereas the latter error only recruited the MPFC (Radke et al., 2011). Another fMRI study demonstrated that when decisions to hurt others (thermal stimulation) were self-generated, the empathic neural responses of the left AI and dorsolateral PFC were stronger than those generated by others (Koban et al., 2013). Together, these studies suggest that more regions or stronger activation in related regions are involved in outcome evaluation or performance monitoring in social contexts.

Consequences leading to harm or benefit to others could cause dramatically different outcomes and social emotions; thus, it is highly important to distinguish the neural basis of success or failure in the social context when outcomes affect others' interests. Previous studies in social psychology have shown that feedback valence triggers different causal attribution patterns in achievement tasks (Weiner, 1985), particularly when the outcome is unexpected (Kanazawa, 1992). Accordingly, casual attribution of success or failure generally arouses different attribution-related emotions, such as pride vs guilt (Weiner, 1985; Tracy and Robins, 2004). People felt guilt or shame when their behavior led to pain for others, and several important brain regions—such as the anterior cingulate cortex (ACC) and AI—were associated with these attribution-related emotions (Yu et al., 2014; Zhu et al., 2019). Our previous study showed that the temporoparietal junction (TPJ) was associated with the processing of group success when participants made more contributions to the group (Li et al., 2013). Despite these advances in the literature, to the best of our knowledge, very few studies have been conducted to investigate the possible neural mechanisms of success and failure in social cooperation. In this study, we focus on both the distinct and common neural mechanisms of success and failure feedback modulated by varied reward expectations in social interaction.

We manipulated the expectancies of the two types of outcomes in a cooperative task to investigate the psychological and neural activities associated with success and failure in social cooperation. On the one hand, reward expectancy modulates reward experience, and a mismatch between expected and actual outcomes causes RPE signals. Human fMRI studies have found that the ventral striatum could represent linear variations of RPE (Bayer and Glimcher, 2005; Abler et al., 2006; Preusschoff et al., 2006); however, contradictory results have been observed in other studies (Tobler et al., 2008; Hsu et al., 2009). In contrast, the expectancy of events influences causal attribution (Hastie, 1984; Weiner, 1985). For example, a large RPE would be elicited when someone finishes a task with a lower chance of success, which in turn could yield stronger positive attribution-related emotions, such as pride (Weiner, 2010). In this fMRI study, we developed a novel three-person collaboration dice-gambling task, wherein the expectancies of success and failure outcomes were manipulated by linearly changing the theoretical probability of team success (Experiment one). Importantly, a nonsocial control study was also conducted to test whether the neural responses to outcomes in Experiment one were socially specific (Experiment two). With parametric analysis, the present paradigm also allowed us to test which brain regions serve to code the gradually varying negative and positive RPEs in a social context.

We predicted that success and failure feedback with varying expectancies would recruit some common and distinct brain regions. First, results of previous studies suggest that feedback with different valences would recruit different brain regions. Specifically, we predicted that the unexpectedness of failure feedback would be associated with the activation of ACC and AI—associated with errors or negative feedback in a social context (Behrens et al., 2008; Eisenberger et al., 2003; Cooper, Dunne, Furey, and O'Doherty, 2014; Yu, Hu, Hu, and Zhou, 2014)—and success feedback with more personal contribution would activate the TPJ, which is related to pride (Li et al., 2013). Second, given that one's performance matters to others' interests, we also predicted that both unexpected success and failure feedback would activate brain regions related to mentalizing and self-referential processing, such as the dorsal medial prefrontal cortex (dMPFC; Van Overwalle, 2009; Radke et al., 2011; Declerck et al., 2013; Koban et al., 2013).

Materials and methods

Experiment one

Participants

Twenty-four participants [11 females, mean age = 21.5, standard deviation (s.d.) = 1.9 years] were recruited as paid volunteers. No participants reported medical or psychiatric disorders or medication/drug/alcohol abuse. All were right-handed and had normal or corrected-to-normal vision. Informed consent was obtained prior to participation in the study, and the study was approved by the local ethics committee. Participants received CNY 60 (equal to US\$8.5) as a basic reward for participating, and an equal share of the group reward was divided between partners after the gambling task as a bonus (CNY 20). One participant's reaction time (RT) data were missing, so his/her RT data were not included in the final analysis.

Stimuli and procedure

Participants were told to conduct a novel cooperative gambling task in a local area network with two partners (disguised study assistants, one female). The whole group would win the game when their score was higher than 10 (success from 11 to 18) and lose when it was lower than 11 (failure from 3 to 10). The two partners assigned to each participant were strangers to the real participant. At the beginning of this experiment, the real participant met the two partners, and all three drew lots to decide their dice-throwing order. By prearrangement, the real participant was always determined to be the third person to conduct the task in the fMRI scanner. Before the formal experiment, the three players practiced this game for 10 trials outside the scanner room. The two partners were then assigned to their personal computers in a room next to the scanner room, and this process was intentionally arranged to be observed by the real participant. Thereafter, the real participant went into the scanner to start the formal experiment.

At the beginning of the formal experiment in each trial, a black figure appeared against a gray background and lasted 1 s (Figure 1). Three boxes appeared with the agents' labels on them; the labels 'A', 'B' and 'You' represented the first and second players and the real participant, respectively. The phrase 'waiting for A' was displayed on the top of the boxes to indicate that it was A's turn. After a short delay of 1–3 s (randomly selected), player A's dice number was presented in the corresponding box. The same procedure was repeated for the second player. Finally, the phrase 'it's your turn' was displayed to alert the real participant

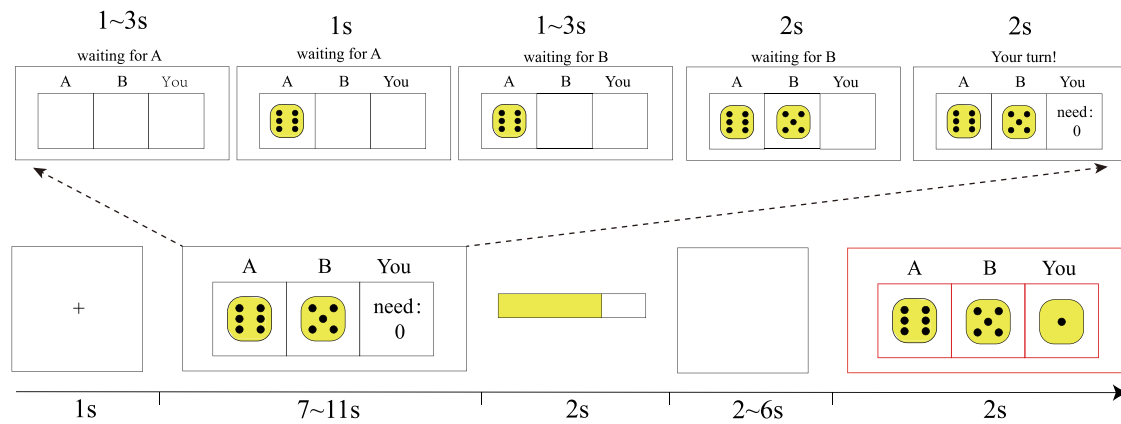


Fig. 1. An illustration of the experimental procedure in one trial.

Table 1. The experimental conditions and hypotheses based on the sum of the partners' scores

SUM _{partners}	Positive feedback		Negative feedback	
	Condition	Unexpectedness	Condition	Unexpectedness
>9	Certain Win	0	N/A	N/A
9	Need 2 to win	1	Need 2 to win	5
8	Need 3 to win	2	Need 3 to win	4
7	Need 4 to win	3	Need 4 to win	3
6	Need 5 to win	4	Need 5 to win	2
5	Need 6 to win	5	Need 6 to win	1
<5	N/A	N/A	Certain Loss	0

that it was their turn to throw the dice. To avoid the mathematical calculation of the sum of the other teammates' dice scores, the program automatically displayed how many points the participant needed to help the group win the game. Subsequently, a black box containing a rapidly increasing yellow bar was displayed on the screen. A cover story was set here to make participants believe that they had some control over this gambling task. They were asked to estimate the length of the dynamically increasing yellow bar and press a button as soon as they believed the yellow bar was close to 3.5 cm. They were also told that the more accurate their estimate, the higher the dice score they would obtain. Moreover, they had to respond within 2 s; otherwise, a negative group feedback was provided by the program. Notably, the difference between 2 s and their RT was automatically filled in with the aim of keeping each event aligned with the repetition time of fMRI scanning. After the response, a blank screen was displayed for 2, 4 or 6 s (randomly selected), followed by the final feedback. The feedback, presented for 2 s, displayed the three players' dice numbers with a colored box. A green box indicated that the group had won the current trial, whereas a red box indicated loss. The meanings of the colors were counterbalanced among the participants.

The full experiment consisted of four runs with 45 trials in each run. Based on the sum of the two partners' numbers (SUM_{partners}), 12 conditions for positive or negative feedback were determined (Table 1). It should be noted that the 'unexpectedness' in Table 1 indicates the level of unexpected success or failure feedback we manipulated in this study. Equal numbers of success and failure feedback were provided. To balance the ecological validity (objective probability of events in reality) and lasting duration in each run, the total numbers of trials for the 12 conditions were

preset as follows: 31 trials for 'Certain Win', 30 trials for 'Certain Loss', 8 trials for 'Need six to win' (and won), 10 trials for 'Need six to win' (but lost), 12 trials for 'Need five to win' (and won), 10 trials for 'Need five to win' (but lost), 19 trials for 'Need four to win' (and won), 21 trials for 'Need four to win' (but lost), 10 trials for 'Need three to win' (and won), 12 trials for 'Need three to win' (but lost), 9 trials for 'Need two to win' (and won) and 8 trials for 'Need two to win' (but lost).

The participants were informed that the group could win or lose CNY 15 (CNY 5 for each player) depending on success or failure in the task. They were also informed that the program would randomly select the outcome of four trials as the final bonus. The feedback during the formal experiment was pseudo-randomly set to a 1:1 ratio of win (success) vs loss (failure). However, unbeknown to participants, only four win trials were finally selected to ensure that each participant obtained a certain amount of bonus (CNY 20 in total).

After scanning, participants were asked to complete a 7-point Likert scale to rate their subjective sense of pride and guilt when receiving success and failure feedback separately. They were also instructed to rate the extent to which their dice points would influence the final results when they observed their teammates' performance. Specifically, they were instructed to choose '1' for least pride and '7' for extreme pride on a 7-point Likert scale for pride rating, choose '1' for least guilt and '7' for extreme guilt on a 7-point Likert scale for guilt rating, and rate the degree of influence on the same scale. For simplification, we only required them to rate their emotions (pride and guilt) and influence scores in 'Certain Win', 'Certain Loss', 'Need two to win', 'Need four to win' and 'Need six to win' conditions. If they did not feel any emotion under each condition, they were to select '0'.

fMRI data acquisition

Brain images were acquired using a 3.0T Siemens Magnetom Trio scanner (Siemens, Munich, Germany) with a standard head coil. Functional images were acquired using T2-weighted, gradient-echo, echo-planar imaging sequences sensitive to blood oxygenation level dependent (BOLD) contrast ($64 \times 64 \times 32$ matrix with $3.4 \times 3.4 \times 3$ mm³ spatial resolution, repetition time = 2000 ms, echo time = 30 ms, flip angle = 90° and field of view = 22×22 cm). A high-resolution T1-weighted structural image ($256 \times 256 \times 180$ matrix with a spatial resolution of $0.47 \times 0.47 \times 1.0$ mm³, repetition time = 8.204 ms, echo time = 3.22 ms and flip angle = 9°) was acquired before the functional scans. Data from the first five functional volumes were used to ensure steady state and were excluded from further analysis.

fMRI data preprocessing and GLM analysis

Functional images were preprocessed using SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK). Head movements were corrected within each run, and six movement parameters (translation: x, y and z and rotation: pitch, roll and yaw) were extracted for further analysis in the statistical model. The anatomical image was coregistered with the mean realigned functional image and normalized to the standard Montreal Neurological Institute (MNI) template. The functional images were resampled to $3 \times 3 \times 3$ mm³ voxels, normalized to the MNI space using the parameters of anatomical normalization and then spatially smoothed using an isotropic 8 mm full-width at half-maximum Gaussian kernel.

Fixed effect analyses were conducted by applying a general linear model (GLM) to event-related fMRI data. Each trial comprised the following phases: Partners' results (the presenting phase of partner B's result); Anticipation (anticipation for self-performance, i.e. the "Your turn!" phase in Figure 1) and Feedback (feedback of group outcome). Our analyses focused on the Feedback phase of trials with responses that required the participants' input to decide the group outcome. Therefore, trials in which the other two partners' cumulative scores ($SUM_{partners}$) were below 5 or above 9 were not included in the analyses. The Feedback phase was modeled for the presentation period (2 s) of the group outcomes. Thus, two conditions (success and failure) were included in the GLM model. To explore which brain regions showed increased activation with increasing unexpectedness (manipulated using the parametrically varied points that the participants needed), each condition also had an associated parametric modulator (regressor) coding for the linear unexpectedness (experiencing success or loss when 2, 3, 4, 5 or 6 points were needed). For the success and failure feedback conditions, unexpectedness and the number of points needed were positively and negatively correlated, respectively. The design matrix also included the Partners' results and Anticipation phases for each trial. In addition, we included the Anticipation and Feedback events of the 'no response' trials as two regressors, as well as the Partners' results and Anticipation and Feedback events of the trials in which the participants' contributions were not required as three regressors of no interest in the model, as well as the realignment parameters. A delta function was used to convolve the canonical hemodynamic response in each condition. We defined regions encoding the unexpectedness of success and failure in the Feedback phase using the corresponding parametric maps. Whole-brain random effect analyses were conducted using one-sample t-tests with beta images for the two parametric modulators to define the brain regions encoding the unexpectedness

of success and failure conditions. Moreover, paired-sample t-tests were conducted on the beta images of success and failure conditions to examine the possible distinct neural correlates of success and failure (using the parametric map of success or failure as an inclusive mask). Conjunction analysis was also conducted to identify the shared brain regions. Brain activations in the whole-brain analyses were defined using a threshold of $P < 0.001$ uncorrected at the single voxel level and $P < 0.05$ with false discovery rate (FDR) correction at the cluster level. The beta values of the parametric regressors coding for the unexpectedness of success and failure in these regions were extracted using MarsBaR (<http://marsbar.sourceforge.net>) to conduct correlation analyses with subjective ratings of pride and guilt. For visualization purposes, beta values for each level of unexpectedness were extracted using another GLM in which trials with different points served as different regressors.

Functional connectivity analysis

According to the results of the whole-brain analyses, the distinct [left TPJ (LTPJ) and ACC] and shared (dMPFC and left anterior insula (LAI)) brain regions coding for unexpectedness of success and failure were used as the regions of interest (ROIs) in the following connectivity analyses. The generalized psychophysiological interaction (gPPI) analysis (McClaren et al., 2012) was performed to test whether the functional connectivity between the distinct and shared brain regions also increased parametrically with the unexpectedness of success and failure, respectively. Because we only focused on the connectivity of LTPJ with dMPFC and LAI for success and ACC with dMPFC and LAI for failure, we used LTPJ and ACC as seed regions and dMPFC and LAI as target ROIs. The coordinates of the peak voxels in the LTPJ and ACC regions across all participants served as seed voxels, defined as a sphere with a 5 mm radius centered at the peak voxels. The time series of each seed region was then extracted. gPPI produces a connectivity map with seed regions separately for each condition (regressor), including parametric regressors. The beta values of the PPI regressor (the interaction of the parametric regressor of unexpectedness and time series of the seed region) were extracted and subjected to one-sample t-tests to test whether the functional connectivity between the shared (dMPFC and LAI) and the distinct (LTPJ for success and ACC for failure) brain regions varied with the unexpectedness of the outcome.

DCM analysis

A dynamic causal model (DCM) analysis was performed to investigate information flow between the shared (dMPFC) and distinct (LTPJ or ACC) brain regions encoding the unexpectedness of success and failure. Since success and failure occurred in different trials and induced different feelings (pride or guilt), and the models needed to be simplified, the connections between the LTPJ and ACC were not modeled. The parametric regressors (unexpectedness) in the Feedback phase were used as the driving inputs and moderators for the connections between brain regions. We systematically varied the inputs on the shared (dMPFC) or distinct regions (LTPJ for the success and ACC for the failure condition), directions of connections between these regions (directions of dMPFC-LTPJ connections for the success condition and dMPFC-ACC connections for the failure condition), and moderation of unexpectedness of success or failure on these connections, resulting in seven model families of 64 models in total for each participant. To test information flow between the shared and distinct

brain regions encoding the unexpectedness of success and failure, the seven model families were defined according to the input and direction of information flow. Specifically, in families 1 and 2, information flow was defined from the shared (dMPFC) to the distinct (LTPJ or ACC) regions, while the inputs of the two families were different. In families 3 and 4, information flow was defined from the distinct (LTPJ or ACC) to the shared (dMPFC) regions, while the inputs of the two families were different. In families 5, 6 and 7, information flow was bilateral between the shared (dMPFC) and distinct (LTPJ or ACC) regions, while the inputs of the three families were different (Figure 4A). A random effect analysis was used in Bayesian model selection to estimate and compare these models.

Two experiments for comparison analysis

Finally, two further analyses were conducted in addition to the parametric analysis. First, we compared neural responses to success and failure feedback in two separate experiments to replicate the classical finding that success feedback elicits stronger activation in reward-related regions compared with failure feedback. To do so, another GLM was constructed with success and failure feedback events as separate regressors regardless of the points needed to win. Second, in order to explore the evidence for distinct neural patterns between social and nonsocial contexts within the regions derived from the above-mentioned parametric analysis, we carried out multi-voxel pattern analysis (MVPA) to provide more information on activation patterns at the voxel level. Pattern analyses were performed using PRoNT to v2.0.1 (Schrouff et al., 2013). The detailed method and results are reported in the Supplementary Materials.

Results

Behavioral results

As shown in Figure 2A, a 2 (feelings: pride in success vs guilt in failure) \times 3 (points needed: 2 vs 4 vs 6) repeated-measures analysis of variance (ANOVA) on the behavioral ratings of feelings of pride and guilt showed a significant interaction ($F_{2,46} = 14.365$,

$P < 0.001$, $\eta^2 = 0.384$), while no significant main effect was found (P -values > 0.61). The simple effect revealed that the guilt feeling in the 'points needed 6' condition was significantly weaker than that in the 'points needed 4' and 'points needed 2' conditions, whereas the latter two were not significantly different from each other ($P = 0.15$). Meanwhile, the feeling of success in the 'points needed 6' condition was significantly stronger than that in the 'points needed 4' and 'points needed 2' conditions, whereas the latter two were not significantly different from each other ($P = 0.13$). This suggested a general pattern wherein in situations of group success, more points were needed to predict stronger feelings of pride, whereas in situations of group loss, more points were needed to predict weaker feelings of guilt. These results confirmed that we successfully induced feelings of pride and guilt using the number of points the participants needed to get when they experienced group success/failure.

RTs were subjected to ANOVA with points needed (0, 2, 3, 4, 5, 6 and >6) as a within-participant variable, showing a significant difference between different points ($F_{6,132} = 20.754$, $P < 0.001$, $\eta^2 = 0.485$), with faster responses for 0 and >6 points (certain outcomes) than for 2, 3, 4, 5 and 6 points (uncertain outcomes) (all $P < 0.001$, Figure 2B). No significant difference between the 'points needed 0' and ' >6 ' conditions ($P = 0.28$) as well as between any two uncertain outcome conditions were noted (all $P > 0.43$). The RT results suggest that participants provided faster responses in certain conditions than in uncertain conditions when the team required their contribution.

fMRI results

Whole-brain analysis of functional fMRI data allows us to determine which regions show a linear increase in activity as a function of the unexpectedness of success or failure in the Feedback phase. Group win resulted in feelings of pride, and brain activity in the TPJ, dMPFC, bilateral insular/inferior frontal gyrus, middle temporal gyrus and middle frontal gyrus increased linearly with the unexpectedness of success, whereas in the group failure

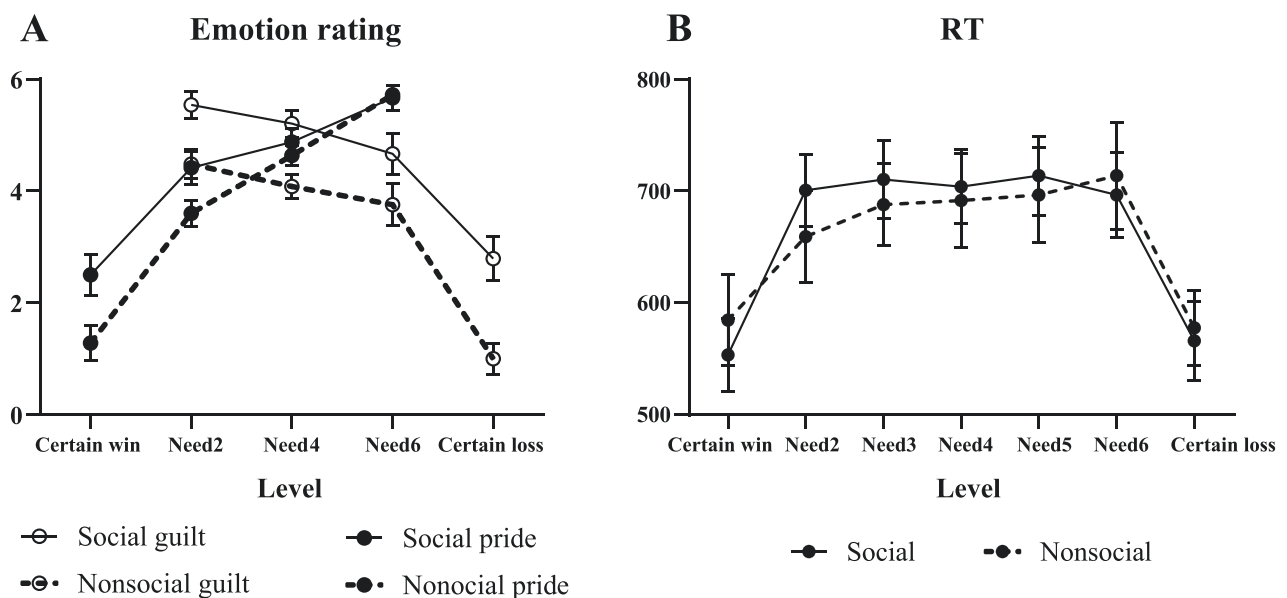


Fig. 2. Behavioral results. (A) Participants' post-experiment rating of guilt in failure conditions (Need2, Need4, Need6 and Certain Loss) and pride in success conditions (Need2, Need4, Need6 and Certain Win); (B) RT in different conditions for the two groups based on the sum of their partners' dice scores.

Table 2. Activations of distinct and shared brain regions sensitive to increased unexpectedness of success and/or failure

Region	MNI Coordinates			Cluster size	Peak Z
	x	y	z		
Increased with unexpectedness of success					
Left temporal-parietal junction	-57	-58	31	675	5.87
Left insular/inferior frontal gyrus	-48	26	-5	481	5.51
Middle temporal gyrus	60	-37	-2	391	4.33
dMPFC	-3	47	19	984	4.33
Right insular/inferior frontal gyrus	48	41	-5	378	4.59
Middle frontal gyrus	-39	17	46	84	3.80
Increased with unexpectedness of failure					
ACC/dMPFC	0	23	25	1237	4.85
Right insular/inferior frontal gyrus	45	20	7	113	4.49
Left insular/inferior frontal gyrus	-33	20	-11	238	4.29
Success > Failure					
Left temporal-parietal junction	-54	-61	37	219	4.32
Failure > Success					
ACC	9	26	34	185	4.21
Conjunction of success and failure					
dMPFC	-3	47	28	318	4.71
Left anterior insula	-36	23	-11	115	4.76

condition, the brain activity of the ACC and bilateral insular/inferior frontal gyrus increased linearly with the unexpectedness of failure (Table 2). To access the distinct neural correlates of success and failure feedback, we conducted paired-sample *t*-tests to find the brain regions specifically coding for success/failure but not for the other (using a parametric map of success/failure as a mask). The results showed that LTPJ ($x/y/z = -54/-61/37$, $z = 4.32$, $k = 219$) activity increased with success unexpectedness (Table 2 and Figure 3A) but not with failure unexpectedness, and ACC ($x/y/z = 9/26/34$, $z = 4.21$, $k = 185$) activity exclusively varied with the unexpectedness of failure but not of success (Table 2 and Figure 3B). Conjunction analysis showed that the activation of the dMPFC ($x/y/z = -3/47/28$, $z = 4.23$, $k = 318$) and LAI ($x/y/z = -36/23/-11$, $z = 4.76$, $k = 115$) increased with the unexpectedness of both success and failure (Table 2 and Figure 3C). In addition, we found that the average subjective susceptibility of pride and guilt was marginally correlated with the average dMPFC activation in the success and failure conditions ($r_{24} = 0.393$, $P = 0.058$); however, this was not the case for the LAI ($r_{24} = -0.106$, $P = 0.623$).

Once the brain regions distinctively coding for the unexpectedness of success (LTPJ) and failure (ACC) and the shared regions (dMPFC and LAI) for both success and failure feedback were identified, we examined how these four brain regions were functionally connected to each other for success (LTPJ with dMPFC and LAI) and failure (ACC with dMPFC and LAI), respectively. Thus, gPPI analyses were conducted with the LTPJ and ACC as seed regions and the dMPFC and LAI as target ROIs to investigate whether the shared regions for success and failure (dMPFC and LAI) were functionally connected to the regions exclusively coding for success (LTPJ) or failure (ACC) in a parametric manner that

was sensitive to changes in unexpectedness. The results showed that the functional connectivity between dMPFC and LTPJ in the success condition as well as between dMPFC and ACC in the failure condition increased linearly with unexpectedness (dMPFC–LTPJ connectivity: $t_{23} = 2.170$, $P = 0.041$, $d = 0.443$; dMPFC–ACC connectivity: $t_{23} = 2.405$, $P = 0.025$, $d = 0.491$), whereas that between the LAI and LTPJ ($t_{23} = 1.956$ and $P = 0.063$) as well as between the LAI and ACC ($t_{23} = 1.697$, $P = 0.103$) was not modulated by unexpectedness.

After confirming the functional connectivity between the LTPJ/ACC and dMPFC, we further investigated how the LTPJ and ACC were effectively connected with the dMPFC. Specifically, did the dMPFC serve as a manager and send the unexpectedness information of success or failure to the LTPJ or ACC, respectively, or did it integrate the unexpectedness information from the LTPJ and ACC? A DCM analysis was conducted to answer this question. With the unexpectedness information of success and failure as inputs and moderators, we constructed 64 models (seven model families) differing in input (to dMPFC or to LTPJ and ACC), modulatory effect and connectivity between the dMPFC and LTPJ/ACC (Figure 4A). As shown in Figure 4B and C, the winning model family added the unexpectedness of success and failure inputs to the LTPJ and ACC, respectively, with bilateral connections between the LTPJ/ACC and dMPFC (exceedance probability = 0.817). In the winning model, the unexpectedness of success linearly modulated information flow from the LTPJ to the dMPFC and feedback from the dMPFC to the LTPJ, whereas the unexpectedness of failure linearly modulated information flow from the ACC to the dMPFC and feedback from the dMPFC to the ACC (exceedance probability = 0.823). This result indicated that the unexpectedness information was entered into the LTPJ for success and into the ACC for failure, and the dMPFC subsequently received the unexpectedness information from these regions (LTPJ for success and ACC for failure) and also sent the information back to these regions. This whole process may be the neural mechanism of feedback processing in a social context.

In addition, the results showed that success feedback activated more reward regions than failure feedback, including the ventral striatum and medial prefrontal cortex (Figure 5, see Supplementary Material for more details).

Experiment two

Participants

Twenty-five participants (12 females, mean age = 22.4, s.d. = 1.8 years) were recruited as paid volunteers. The inclusion and exclusion criteria for participation and reward rules were the same as those in Experiment one. Informed consent was obtained prior to participation in the study. The local ethics committee approved the study.

Stimuli and procedure

Experiment two adopted the same task as Experiment one, with the exception that participants played the game with two computers instead of two human partners.

fMRI data acquisition and analysis

Although Experiment two was conducted 7 years later, the same scanner, scanning sequence and GLM model were used in Experiment two to acquire and measure the BOLD signals as in Experiment one. Given that the parametric analysis results did not pass the multiple comparison thresholds, no further

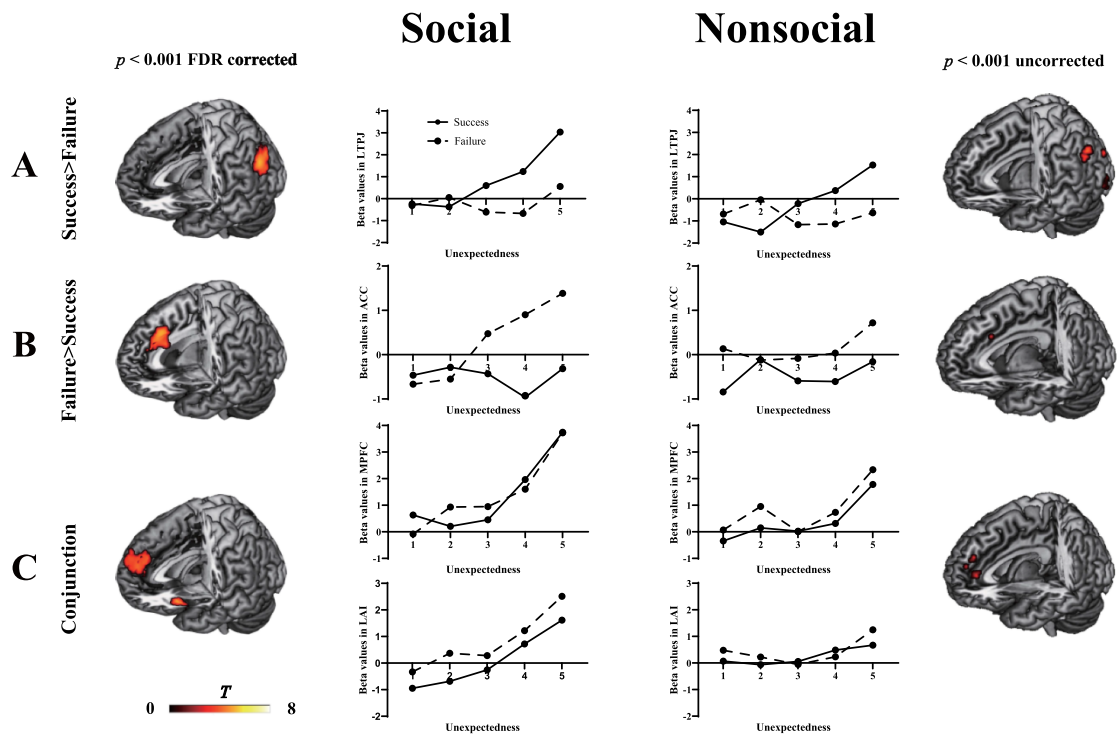


Fig. 3. Distinct and shared brain regions coding the unexpectedness of success and failure (A), the unexpectedness of failure > success conditions (B) and the conjunction between the unexpectedness of success and failure (C). The left panels show brain activation results in the social group, whereas the right panels show the results in the nonsocial group. The line charts illustrate beta values as a function of the unexpectedness of success (solid line) and failure (dashed line).

analyses—such as gPPI and DCM analyses—were conducted in Experiment two.

Results

Behavioral results

As in Experiment one, a 2 (feelings: pride in success vs guilt in failure) \times 3 (points needed: 2 vs 4 vs 6) repeated-measures ANOVA was performed on the behavioral ratings of feelings of pride/guilt. The main effects of both feelings ($F_{1,24} = 8.6$, $P < 0.01$, $\eta^2 = 0.26$) and points needed ($F_{2,48} = 7.36$, $P < 0.005$, $\eta^2 = 0.24$) reached significance, whereas a significant interaction between feelings and points needed was noted ($F_{2,48} = 21.38$, $P < 0.001$, $\eta^2 = 0.47$). The simple effect analysis revealed no significant difference between any two ‘points needed’ conditions when the group lost the game (all $P > 0.09$); however, participants felt prouder in the ‘points needed 6’ condition than in the ‘points needed 4’ condition ($P < 0.001$) as well as in the ‘points needed 4’ condition than in the ‘points needed 2’ condition ($P < 0.001$) when the group won the game. These results demonstrated that in situations of a group win, more points were needed to predict stronger feelings of pride. However, no significant changes in feelings of guilt were reported, regardless of the points needed (Figure 2A).

A one-way, repeated-measures ANOVA was conducted on the RT with the seven conditions (points needed: ‘0’, ‘2’, ‘3’, ‘4’, ‘5’, ‘6’ and ‘>6’) as independent variables. The results revealed that the main effect of points needed was significant ($F_{2,72,65.33} = 10.97$, $P < 0.001$, $\eta^2 = 0.31$). Pairwise comparisons revealed faster responses for 0 and >6 points (certain outcomes) than for 2, 3, 4, 5 and 6 points (uncertain outcomes) (all $P < 0.03$; Figure 2B). No significant difference between the ‘points needed

0’ and ‘>6’ conditions ($P = 0.71$) as well as between any two uncertain outcome conditions were noted (all $P > 0.06$), except that participants provided faster responses in the ‘need 2’ condition than in the ‘need 4’, ‘need 5’ and ‘need 6’ conditions (all $P < 0.02$). These data suggest that participants responded faster when group results were determined by partners compared with when the participant made a contribution (Figure 2B).

fMRI results

With the same threshold ($P < 0.05$ with FDR, $P < 0.001$ uncorrected at the single voxel level) for correcting multiple comparisons as in Experiment one, parametric analysis did not yield any significant regions in either the success or failure conditions in Experiment two. However, the data showed a similar pattern when the results were inspected by liberally uncorrected P -values. Specifically, the TPJ was activated in the success feedback condition ($x/y/z = 57/-61/43$, uncorrected $P < 0.001$ at voxel level), and the ACC was also activated in the success feedback condition ($x/y/z = 3/47/13$, uncorrected $P < 0.005$ at voxel level), as shown in the right panel of Figure 3.

As predicted, a direct T-contrast between success and failure regardless of points needed showed reward-related regions, including the ventral striatum and ventral medial prefrontal cortex (Figure 5, see Supplementary Materials for more details).

Between-experiments comparison

To investigate whether psychological and neural activities associated with success and failure at different levels of expectancy are socially specific, we ran a mixed-variables ANOVA with group (social vs nonsocial) as a between-participant variable on both behavioral data and BOLD signals. Since the main results of each

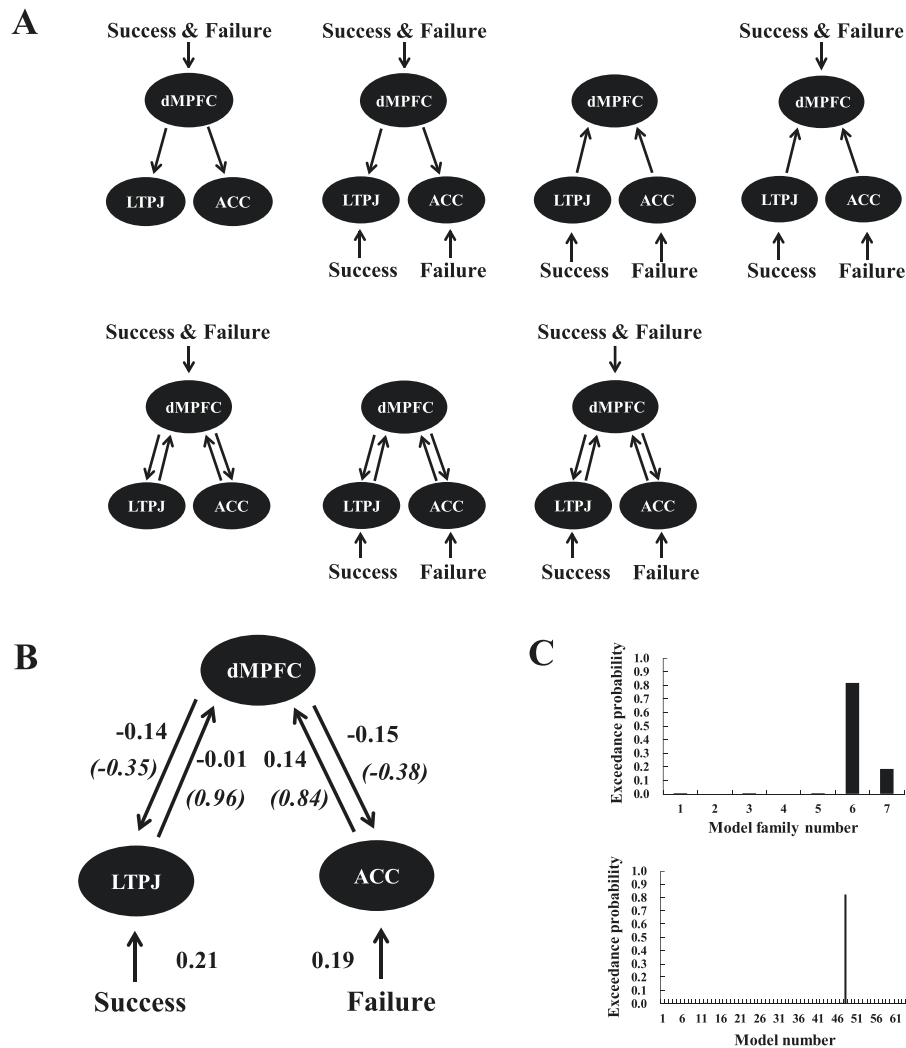


Fig. 4. DCM analysis of the network consisting of the LTPJ, ACC and dMPFC. (A) Sixty-four models grouped into seven model families differing in input and connectivity between the dMPFC and LTPJ/ACC. (B) Parameters (connectivity strength) of the winning model. The numbers without brackets indicate the strength of intrinsic connectivity, and the numbers with brackets indicate the strength of the modulatory effects of success (between the dMPFC and LTPJ) and failure (between the dMPFC and ACC). (C) The exceedance probabilities of model families (upper panel) and individual models (lower panel). Note that success/failure in the figure implies the unexpectedness of success/failure.

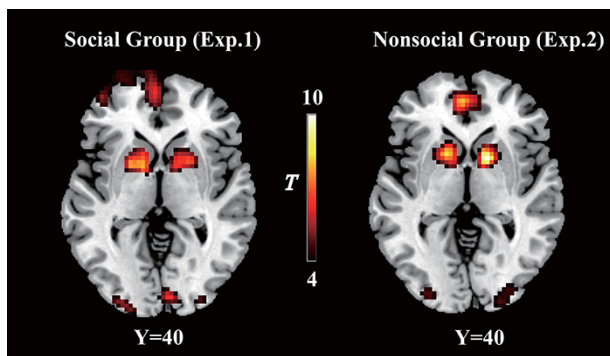


Fig. 5. Left and right panels illustrate the success > failure contrast activations in the social and nonsocial groups, respectively.

experiment have been reported, here we mainly report the group and interaction effects between the groups and other variables, if any.

Behavioral data

A two-way ANOVA on RT showed that the main effect of points needed was significant ($F_{3,3,151.73} = 29.53$, $P < 0.001$, $\eta^2 = 0.39$), suggesting that participants spent less time in certain conditions (points needed = 0 and >6) than in uncertain conditions (points needed = 2, 3, 4, 5 or 6). The pairwise comparisons revealed faster responses for 0 and >6 points (certain outcomes) than for 2, 3, 4, 5 and 6 points (uncertain outcomes) (all $P < 0.001$, Figure 2B). No significant difference between the 'points needed 0' and '>6' conditions ($P = 0.79$) and between any two uncertain outcome conditions was noted (all $P > 0.05$). However, neither the group effect nor the interaction between the groups and points reached significance (all $P > 0.05$). The RT data suggested that both groups spent comparable time on the task, which could rule out any movement-related activity in the difference in brain patterns between groups.

A three-way ANOVA on emotion rating showed that the main group effect was significant ($F_{1,47} = 18.04$, $P < 0.001$, $\eta^2 = 0.28$), indicating that the social group reported a stronger emotion

(4.46 ± 0.15) than the nonsocial group (3.57 ± 0.15). Importantly, the interaction effect between groups and valence was significant ($F_{1,47} = 5.33$, $P < 0.03$, $\eta^2 = 0.10$). Post hoc analyses showed that the social group reported more pride (4.37 ± 0.16) than the nonsocial group (3.81 ± 0.16 , $P < 0.02$) when they won the game and more guilt (4.55 ± 0.2) than the nonsocial group (3.32 ± 0.2 , $P < 0.001$) when they lost. In addition, the main effect of level ($F_{1,99,93.44} = 107.33$, $P < 0.001$, $\eta^2 = 0.70$) and interaction between level and valence ($F_{2,46,115.79} = 3.35$, $P < 0.03$, $\eta^2 = 0.07$) was significant, showing the same pattern as in each experiment. No other significant differences were observed.

We conducted a 2 (social vs nonsocial group) \times 5 (points needed: '0', '2', '4', '6' and '>6') mixed ANOVA to test the influence of the scores. The results revealed that the main effect of the group was significant ($F_{1,47} = 13.93$, $P = 0.001$, $\eta^2 = 0.23$); participants in the social group reported higher influence scores than those in the nonsocial group, possibly suggesting that self-involvement in the social context was higher than that in the nonsocial context. Importantly, the main effect of the 'points needed' condition was also significant ($F_{2,9,138.4} = 114.84$, $P < 0.001$, $\eta^2 = 0.71$). Post hoc analyses demonstrated that participants reported lower influence scores for 0 and >6 points (certain outcomes) than for 2, 4 and 6 points (uncertain outcomes) (all $P < 0.001$). No significant difference between the 'points needed 0' and '>6' conditions ($P = 1.0$), as well as between the '2' and '4' points needed conditions ($P = 0.29$) was noted. However, participants felt that their results would influence the group results more in the '6' than in the '2' and '4' points needed conditions (both $P < 0.03$). These results suggest that participants believed that their contributions would have more influence on the final results when the group needed more points to win the game.

fMRI data—mixed ANOVA

We conducted a two-way mixed ANOVA on the beta images of feedback parametrically modulated by feedback expectancy with group (social vs nonsocial) and valence (positive and negative) as between- and within-participant variables. No significant interaction effect was found between group and valence. The main effects of valence-activated regions, including the ventral striatum, ACC and dMPFC, are shown in Table 3. Importantly, the main group effects showed that activation of the LAI and dMPFC was stronger in the social group than in the nonsocial group ($P < 0.05$, FDR corrected at cluster level; Table 3 and Figure 6). As shown in Figure 5, these regions overlapped with the conjunction regions between success and failure in Experiment one. These results demonstrated that both dMPFC and LAI were involved in integrating feedback information of success and failure as well as distinguishing social and nonsocial contexts.

Table 3. Brain regions activated by the main effects of group and valence

Region	MNI Coordinates			Cluster size	Peak Z
	x	y	z		
Main effect of group					
Left anterior insula	-33	20	-11	89	4.68
dMPFC	-9	62	37	73	4.42
Occipital cortex	-9	-97	-2	72	3.78
Main effect of valence					
Ventral striatum (R)	24	2	-8	245	6.32
Ventral striatum (L)	-24	11	-2	189	5.56

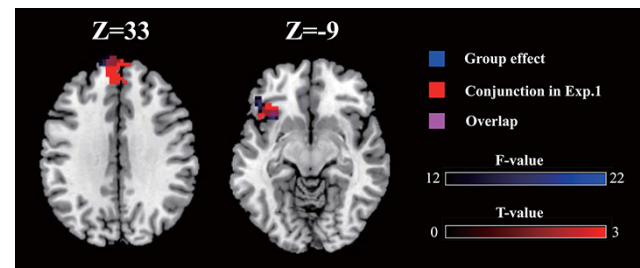


Fig. 6. Blue and red represent the regions activated by group effects (social > nonsocial) and regions from conjunction analysis between success and failure conditions in Experiment one. Violet regions indicate overlaps between the regions from these two analyses.

As we have reported the T-contrast between success and failure feedback for each of the aforementioned experiments, we performed another mixed ANOVA with group and valence as independent variables and mainly focused on the group and interaction effects. The main effects of group recruited regions included the occipital cortex, superior parietal lobule and lateral prefrontal cortex. Importantly, the region was activated by the interaction between the group and valence. Detailed results are reported in the [Supplementary Materials](#).

Discussion

This study used a novel paradigm to explore the neural basis of feedback processing with linearly increased reward expectations in a social cooperation context. In both experiments, the RT showed that participants took longer to make decisions in conditions of uncertain outcome compared with those in conditions when group outcomes were predefined by their teammates or computers. Moreover, participants in the social group reported an increased pattern of pride and guilt intensity as a function of group outcome and teammates' scores in uncertain conditions (needing 2, 4 or 6 to win). In addition, the social group reported stronger attribution-related emotions and more influence on the final outcome than the nonsocial group. Therefore, these behavioral results suggest that the current design successfully manipulates linearly increasing reward expectations during cooperation. More importantly, the fMRI results revealed that the LTPJ and ACC coded the success and failure outcomes with linearly increasing prediction errors, respectively, and that the dMPFC and LAI coded unexpected success and failure outcomes collectively in the social context.

The fMRI results in Experiment one showed that success and failure outcomes in social cooperation activated two different brain regions. Specifically, the ACC was selectively activated by the level of negative prediction error caused by the participants' performance. These results are consistent with previous findings wherein the ACC processed internal and external error signals and negative RPE (Holroyd et al., 2004; Matsumoto et al., 2007). The ACC has also been linked with social prediction errors (Behrens et al., 2008) and negative feelings of social exclusion and rejection (Eisenberger et al., 2003; Cooper, Dunne, Furey, and O'Doherty, 2014). The ACC was also found to be involved when participants made error responses with great responsibility and felt guilt in an interpersonal context (Yu, Hu, Hu, and Zhou, 2014). Although the between-experiment comparisons in this study did not suggest that ACC activation was stronger in the social group than in the nonsocial group, MVPA analysis (Supplementary Figure S1) showed that the activation pattern in this region was different,

which may serve to code for a stronger guilt feeling in the social vs nonsocial groups. The unique neural representation of guilt in the ACC was in line with previous studies that focused on interpersonal guilt (Li et al., 2020; Yu et al., 2020).

The LTPJ was found to be particularly sensitive to unexpected success feedback. Note that although only LTPJ activation was reported here, right TPJ activation was also observed in our data with a liberal threshold (uncorrected $P < 0.005$ at single voxel level). The function of the TPJ was suggested to be either social domain-specific—such as predicting others' mental states or future actions in social tasks (Saxe and Kanwisher, 2003; Carter et al., 2012) and social prediction error (Behrens et al., 2008)—or domain-general, that is, attention to unexpected stimuli (Krall et al., 2015). Neither univariate nor multivariate analyses found differences in LTPJ activation between the social and nonsocial groups, perhaps supporting the latter interpretation (that the LTPJ was related to unexpected outcomes when the group won the trial). Another possibility is that participants in the nonsocial group also felt a certain level of pride, even when they played with computers. Previous researchers have proposed that internal attributions for success tend to produce pride in such achievement-related tasks (Weiner, 1985). The function of pride may promote social status in achievement situations (Tracy and Robins, 2004), increasing the individual's visibility not only to partners in a social task but also to others, such as experimenters in a nonsocial task. However, the interpretation of the null effect of LTPJ activation between social and nonsocial contexts needs to be interpreted cautiously. Nevertheless, our study provided evidence that the TPJ was associated with successful outcomes, which is consistent with our previous study (Li et al., 2013).

Importantly, the unexpectedness of both success and failure feedback recruited the dMPFC, a region that has been widely found to represent mentalizing and theory of mind (Amodio and Frith, 2006; Frith and Frith, 2006; Blakemore, 2008; Van Overwalle, 2009). Our results also showed that the LAI was activated by both unexpected success and failure; however, only the dMPFC was functionally connected with both the LTPJ and ACC. Correspondingly, only the dMPFC was marginally correlated with post-experiment ratings of attribution-related emotions; further activation of the dMPFC was observed when participants felt stronger emotions (pride or guilt) regardless of their valence. Moreover, DCM analysis revealed that the dMPFC works to integrate information from the two separated pathways: one from the ACC, activation of which was driven by manipulation of unexpected failure, and another from the LTPJ, in which activation was associated with unexpected success. To our knowledge, this is the first study to reveal the hierarchical information processing of RPE signals in social cooperation, that is, as a key node of the social brain network, the dMPFC not only integrates information received from the LTPJ and ACC but also sends information back to these regions. Furthermore, compared with the nonsocial context, feedback in the social context induced stronger dMPFC activation. Thus, we demonstrated that the dMPFC serves a central role in dealing with salience prediction error signals and is associated with success and failure outcomes in a social context when one's performance affects others' interests.

Paralleling the activation pattern of the dMPFC, the LAI was also sensitive to the increasing unexpectedness of both success and failure feedback. Moreover, the LAI was also recruited to differentiate social and nonsocial contexts. This region has been commonly associated with empathy, pain or interpersonal guilt

(Singer et al., 2004; Koban et al., 2013; Li et al., 2020; Yu et al., 2020). Notably, an electric shock was introduced as an enforcer in these studies, thus potentially amplifying negative outcomes against positive outcomes. Our study used monetary feedback and further showed that the AI might serve not only for unexpected negative events but also for positive events in the present social interaction. In other words, AI seems to work for the detection of salient events in the paradigm we adopted here (see also Seeley et al., 2007; Sridharan et al., 2008). Koban and Pourtois (2014) proposed an integrative framework to account for overlaps in the activation of the dMPFC and LAI for coding social context and action monitoring. In line with this statement, our study also showed the important role of the dMPFC and LAI in processing the social cooperation context and unexpected outcomes. However, Koban and Pourtois (2014) mainly focused on the error response monitoring aspect. We argue that an updated framework is needed to consider positive outcomes as well.

Surprisingly, the ventral striatum was not recruited by the linearly varying RPE, and this region only served to process feedback valence in this study. The robust activation of the ventral striatum in response to feedback valence was in line with previous findings (O'doherty et al., 2004; Pagnoni et al., 2002; McClure et al., 2003; for a review, see Garrison et al., 2013). However, this region was not sensitive to the varying RPE magnitude, contrary to the findings from previous monkey studies (Montague et al., 1996; Schultz et al., 1997) and human fMRI studies (D'ardenne, McClure, Nystrom, and Cohen, 2008; McClure et al., 2003; Garrison et al., 2013, but see Hsu et al., 2009; Tobler et al., 2008). Note that all these studies only focused on reward processing in single-agent conditions. On the contrary, the feedback stimuli in the current paradigm conveyed performance information from the participants themselves as well as from their group partners. It is possible that the ventral striatum may not be flexible enough to represent varying RPE signals in a relatively complicated task that requires coordination with other agents in particular contexts. A control experiment with only individual economic tasks is needed to confirm our argument in future studies.

Our study also provides implications for understanding the common neural currency hypothesis. According to this hypothesis, a shared neural underpinning codes reward valuation and anticipation in both social and nonsocial decision-making (Ruff and Fehr, 2014; Gu et al., 2019). In this framework, the specific activation related to the social aspect of the environment was interpreted as an input for the shared reward neural circle (Ruff and Fehr, 2014). However, the DCM model and between-experiment comparison in our study seem to support the possibility that the dMPFC serves as a top-down control system to monitor social context information and motivate individuals' feedback learning in a social context. Further studies are necessary to discuss the interaction between common and distinct regions in social and nonsocial reward processing.

This study has several limitations. First, we asked participants to report their own ratings of specific emotions (guilt or pride) in each condition after scanning. This potentially led to a demand effect. Although we observed a marginally significant correlation between emotion intensity and dMPFC activation, listing different types of emotions and asking participants to select their salient emotion in each condition to rule out the demand effect might have been preferable. Second, Experiment two was conducted 7 years after Experiment one, thus potentially causing contamination of the between-experiment difference in this study. However, these two experiments were conducted using the same scanner, and participants were recruited from

the same university. We believe participants' behaviors and brain responses were less likely to be altered because RPE-based social emotions are fundamental mental processes in social interaction.

Conclusion

The present paradigm identified two common regions—the dMPFC and LAI, rather than the ventral striatum—that coded the linear prediction error elicited by both success and failure during interpersonal cooperation. Moreover, both these regions also coded the social context; however, only the dMPFC played an important role in integrating learning information from different feedbacks. Furthermore, we found that the ACC and LTPJ were recruited to process negative and positive prediction errors, respectively, suggesting that these regions worked on the first level of learning information from feedback. This study demonstrate that a hierarchical neural network serves to process salient success and failure in the context of social cooperation.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC 31671158) and the Guangdong Key Project in 'Development of new tools for diagnosis and treatment of Autism' (2018B030335001).

Conflict of interest

The authors declared that they had no conflict of interest with respect to their authorship or the publication of this article.

Supplementary data

Supplementary data are available at SCAN online.

Data availability

Our data were analyzed using open-source tools. Most of the fMRI data analyses, including the preprocessing, first level general linear model, second level group analyses and dynamic causality model analysis, were conducted using statistical parametric mapping (SPM12, <https://www.fil.ion.ucl.ac.uk/spm/>), and the multivariate pattern analysis was run using PRoNTTo (Schrouff et al., 2013, *Neuroinformatics*, 11(3), 319–337). In addition, the functional connectivity analysis was conducted using the gPPI toolbox (McClaren et al., 2012, *NeuroImage*, 61(4), 1277–1286). Codes are available online at https://github.com/PengSZU-Lab/social_FMRI_experiment, and MNI-registered functional image data are accessible at <https://osf.io/nfpwx/>.

References

- Abler, B., Walter, H., Erk, S., Kammerer, H., Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage*, **31**, 790–5. [10.1016/j.neuroimage.2006.01.001](https://doi.org/10.1016/j.neuroimage.2006.01.001).
- Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**, 268–77. [10.1038/nrn1884](https://doi.org/10.1038/nrn1884).
- Bayer, H.M., Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, **47**(1), 129–41. [10.1016/j.neuron.2005.05.020](https://doi.org/10.1016/j.neuron.2005.05.020).
- Behrens, T.E., Hunt, L.T., Woolrich, M.W., Rushworth, M.F. (2008). Associative learning of social value. *Nature*, **456**(7219), 245–9. [10.1038/nature07538](https://doi.org/10.1038/nature07538).
- Beyer, F., Sidarus, N., Bonicalzi, S., Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive Affective Neuroscience*, **12**(1), 138–45. [10.1093/scan/nsw160](https://doi.org/10.1093/scan/nsw160).
- Blakemore, S.J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, **9**(4), 267–77. [10.1038/nrn2353](https://doi.org/10.1038/nrn2353).
- Carter, M.C., Bowling, D.L., Reeck, C., Huettel, S.A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, **337**, 109–11. [10.1126/science.1219681](https://doi.org/10.1126/science.1219681).
- Cooper, J.C., Dunne, S., Furey, T., O'Doherty, J.P. (2014). The role of the posterior temporal and medial prefrontal cortices in mediating learning from romantic interest and rejection. *Cerebral Cortex*, **24**(9), 2502–11. [10.1093/cercor/bht102](https://doi.org/10.1093/cercor/bht102).
- D'Ardenne, K., McClure, S.M., Nystrom, L.E., Cohen, J.D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, **319**(5867), 1264–7.
- Declerck, C.H., Boone, C., Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition*, **81**(1), 95–117. [10.1016/j.bandc.2012.09.009](https://doi.org/10.1016/j.bandc.2012.09.009).
- Eisenberger, N.I., Lieberman, M.D., Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, **302**(5643), 290–2. [10.1126/science.1089134](https://doi.org/10.1126/science.1089134).
- Frith, C.D., Frith, U. (2006). The neural basis of mentalizing. *Neuron*, **50**(4), 531–4. [10.1016/j.neuron.2006.05.001](https://doi.org/10.1016/j.neuron.2006.05.001).
- Garrison, J., Erdeniz, B., Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, **37**(7), 1297–310. [10.1016/j.neubiorev.2013.03.023](https://doi.org/10.1016/j.neubiorev.2013.03.023).
- Gu, R., Huang, W., Camilleri, J., et al. (2019). Love is analogous to money in human brain: coordinate-based and functional connectivity meta-analyses of social and monetary reward anticipation. *Neuroscience and Biobehavioral Reviews*, **100**, 108–28. [10.1016/j.neubiorev.2019.02.017](https://doi.org/10.1016/j.neubiorev.2019.02.017).
- Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology*, **46**(1), 44–56. [10.1037/0022-3514.46.1.44](https://doi.org/10.1037/0022-3514.46.1.44).
- 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**(4), 679.
- 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**(5), 497–8. [10.1038/nn1238](https://doi.org/10.1038/nn1238).
- Hsu, M., Krajbich, I., Zhao, C., Camerer, C.F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience*, **29**(7), 2231–7. [10.1523/JNEUROSCI.5296-08.2009](https://doi.org/10.1523/JNEUROSCI.5296-08.2009).
- Huang, C., Yu, R. (2018). Making mistakes in public: being observed magnifies physiological responses to errors. *Neuropsychologia*, **119**, 214–22. [10.1016/j.neuropsychologia.2018.08.015](https://doi.org/10.1016/j.neuropsychologia.2018.08.015).
- Kanazawa, S. (1992). Outcome or expectancy? Antecedent of spontaneous causal attribution. *Personality & Social Psychology Bulletin*, **18**(6), 659–68. [10.1177/0146167292186001](https://doi.org/10.1177/0146167292186001).
- Koban, L., Pourtois, G., Bediou, B., Vuilleumier, P. (2012). Effects of social context and predictive relevance on action outcome monitoring. *Cognitive Affective Behavioral Neuroscience*, **12**, 460–78. [10.3758/s13415-012-0091-0](https://doi.org/10.3758/s13415-012-0091-0).
- Koban, L., Corradi-Dell'Acqua, C., Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. *Journal of Cognitive Neuroscience*, **25**(2), 258–72. [10.1162/jocn_a_00324](https://doi.org/10.1162/jocn_a_00324).

- Koban, L., Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and Biobehavioral Reviews*, **46**, 71–84. [10.1016/j.neubiorev.2014.02.014](https://doi.org/10.1016/j.neubiorev.2014.02.014).
- Krall, S.C., Rottschy, C., Oberwelland, E., et al. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure & Function*, **220**(2), 587–604. [10.1007/s00429-014-0803-z](https://doi.org/10.1007/s00429-014-0803-z).
- Li, P., Jia, S., Feng, T., Liu, Q., Suo, T., Li, H. (2010). The influence of the diffusion of responsibility effect on outcome evaluations: electrophysiological evidence from an ERP study. *Neuroimage*, **52**(4), 1727–33. [10.1016/j.neuroimage.2010.04.275](https://doi.org/10.1016/j.neuroimage.2010.04.275).
- Li, P., Shen, Y., Sui, X., et al. (2013). The neural basis of responsibility attribution in decision-making. *PLoS One*, **8**(11), e80389. [10.1371/journal.pone.0080389](https://doi.org/10.1371/journal.pone.0080389).
- Li, Z., Yu, H., Zhou, Y., Kalenscher, T., Zhou, X. (2020). Guilty by association: how group-based (collective) guilt arises in the brain. *Neuroimage*, **209**, 116488. [10.1016/j.neuroimage.2019.116488](https://doi.org/10.1016/j.neuroimage.2019.116488).
- Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, **10**, 647–56. [10.1038/nrn1890](https://doi.org/10.1038/nrn1890).
- McClure, S.M., Berns, G.S., Montague, P.R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, **38**(2), 339–46. [10.1016/S0896-6273\(03\)00154-5](https://doi.org/10.1016/S0896-6273(03)00154-5).
- Mclaren, D.G., Ries, M.L., Xu, G., Johnson, S.C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage*, **61**(4), 1277–86. [10.1016/j.neuroimage.2012.03.068](https://doi.org/10.1016/j.neuroimage.2012.03.068).
- Montague, P.R., Dayan, P., Sejnowski, T.J. (1996). A framework for mesen-cephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, **16**, 1936–47. [10.1523/JNEUROSCI.16-05-01936.1996](https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996).
- O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**(5669), 452–4. [10.1126/science.1094285](https://doi.org/10.1126/science.1094285).
- Pagnoni, G., Zink, C.F., Montague, P.R., Berns, G.S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, **5**(2), 97. [10.1038/nrn802](https://doi.org/10.1038/nrn802).
- Preusschoff, K., Bossaerts, P., Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, **51**(3), 381–90. [10.1016/j.neuron.2006.06.024](https://doi.org/10.1016/j.neuron.2006.06.024).
- Radke, S., De Lange, F.P., Ullsperger, M., De Bruijn, E.R.A. (2011). Mistakes that affect others: an fMRI study on processing of own errors in a social context. *Experimental Brain Research*, **211**(3–4), 405–13. [10.1007/s00221-011-2677-0](https://doi.org/10.1007/s00221-011-2677-0).
- Ruff, C.C., Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, **15**(8), 549–62. [10.1038/nrn3776](https://doi.org/10.1038/nrn3776).
- Saxe, R., Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in ‘theory of mind’. *Neuroimage*, **19**(4), 1835–42. [10.1016/S1053-8119\(03\)00230-1](https://doi.org/10.1016/S1053-8119(03)00230-1).
- Schrouff, J., Rosa, M.J., Rondina, J.M., et al. (2013). PRoNTTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, **11**(3), 319–37. [10.1007/s12021-013-9178-1](https://doi.org/10.1007/s12021-013-9178-1).
- Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593–9. [10.1126/science.275.5306.1593](https://doi.org/10.1126/science.275.5306.1593).
- Seeley, W.W., Menon, V., Schatzberg, A.F., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, **27**(9), 2349–56.
- Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, **303**, 1157–62. [10.1126/science.1093535](https://doi.org/10.1126/science.1093535).
- Sridharan, D., Levitin, D.J., Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, **105**(34), 12569–74.
- Tobler, P.N., Christopoulos, G.I., O’Doherty, J.P., Dolan, R.J., Schultz, W. (2008). Neuronal distortions of reward probability without choice. *Journal of Neuroscience*, **28**(45), 11703–11. [10.1523/JNEUROSCI.2870-08.2008](https://doi.org/10.1523/JNEUROSCI.2870-08.2008).
- Tracy, J.L., Robins, R.W. (2004). Putting the self into self-conscious emotions: a theoretical model. *Psychological Inquiry*, **15**(2), 103–25. [10.1207/s15327965pli1502_01](https://doi.org/10.1207/s15327965pli1502_01).
- Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**(3), 829–58. [10.1002/hbm.20547](https://doi.org/10.1002/hbm.20547).
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, **92**(4), 548. [10.1037/0033-295X.92.4.548](https://doi.org/10.1037/0033-295X.92.4.548).
- Weiner, B. (2010). The development of an attribution-based theory of motivation: a history of ideas. *Educational Psychologist*, **45**(1), 28–36. [10.1080/00461520903433596](https://doi.org/10.1080/00461520903433596).
- Yu, H., Hu, J., Hu, L., Zhou, X. (2014). The voice of conscience: neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, **9**(8), 1150–8. [10.1093/scan/nst090](https://doi.org/10.1093/scan/nst090).
- Yu, H., Koban, L., Chang, L.J., et al. (2020). A generalizable multivariate brain pattern for interpersonal guilt. *Cerebral Cortex*, **30**(6), 3558–72. [10.1093/cercor/bhz326](https://doi.org/10.1093/cercor/bhz326).
- Zhu, R., Feng, C., Zhang, S., Mai, X., Liu, C. (2019). Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses. *Neuroimage*, **186**, 476–86. [10.1016/j.neuroimage.2018.11.012](https://doi.org/10.1016/j.neuroimage.2018.11.012).