Behavioral/Cognitive

Learning from Ingroup Experiences Changes Intergroup Impressions

Yuqing Zhou,[1*] Björn Lindström,[2*] Alexander Soutschek,[3] Pyungwon Kang,[4] Philippe N. Tobler,[4†] and Grit Hein[1†]

[1]Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, Würzburg 97080, Germany, [2]Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, 1081 HV, The Netherlands, [3]Department of Psychology, Ludwig Maximilian University, Munich 80802, Germany, and [4]Department of Economics and Laboratory for Social and Neural Systems Research, University of Zurich and Neuroscience Center Zurich, University of Zurich and Swiss Federal Institute of Technology Zurich, Zurich, CH-8006, Switzerland

Humans form impressions toward individuals of their own social groups (ingroup members) and of different social groups (outgroup members). Outgroup-focused theories predict that intergroup impressions are mainly shaped by experiences with outgroup individuals, while ingroup-focused theories predict that ingroup experiences play a dominant role. Here we test predictions from these two psychological theories by estimating how intergroup impressions are dynamically shaped when people learn from both ingroup and outgroup experiences. While undergoing fMRI, male participants had identical experiences with different ingroup or outgroup members and rated their social closeness and impressions toward the ingroup and the outgroup. Behavioral results showed an initial ingroup bias in impression ratings which was significantly reduced over the course of learning, with larger effects in individuals with stronger ingroup identification. Computational learning models revealed that these changes in intergroup impressions were predicted by the weight given to ingroup prediction errors. Neurally, the individual weight for ingroup prediction errors was related to the coupling between the left inferior parietal lobule and the left anterior insula, which, in turn, predicted learning-related changes in intergroup impressions. Our findings provide computational and neural evidence for ingroup-focused theories, highlighting the importance of ingroup experiences in shaping social impressions in intergroup settings.

Key words: **fMRI; impression formation; ingroup; learning; outgroup; prediction error**

Significance Statement

Living in multicultural societies, humans interact with individuals of their own social groups (ingroup members) and of different social groups (outgroup members). However, little is known about how people learn from the mixture of ingroup and outgroup interactions, the most natural experiences in current societies. Here, participants had identical, intermixed experiences with different ingroup and outgroup individuals and rated their closeness and impressions toward the ingroup and the outgroup. Combining computational models and fMRI, we find that the weight given to ingroup experiences (ingroup prediction errors) is the main source of intergroup impression change, captured by changes in connectivity between the parietal lobe and insula. These findings highlight the importance of ingroup experiences in shaping intergroup impressions in complex social environments.

Received Jan. 5, 2022; revised June 29, 2022; accepted July 11, 2022.

Author contributions: Y.Z. and B.L. analyzed data; Y.Z., B.L., and G.H. wrote the first draft of the paper; Y.Z., B.L., A.S., P.K., P.N.T., and G.H. edited the paper; Y.Z., B.L., A.S., P.K., P.N.T., and G.H. wrote the paper; B.L., P.N.T., and G.H. designed research; B.L., A.S., and P.K. performed research.

This work was supported by German Research Foundation HE 4566/5-1 and HE 4566/3-1 to G.H.; German Research Foundation Emmy Noether fellowship SO 1636/2-1 to A.S.; and Swiss National Science Foundation Grants 100014_165884, 100019_176016, and IZKSZ3_162109 to P.N.T.

*Y.Z. and B.L. contributed equally to this work.

†P.N.T. and G.H. contributed equally to this work as senior authors.

The authors declare no competing financial interests.

Correspondence should be addressed to Grit Hein at Hein_G@ukw.de or Yuqing Zhou at yuqingzhou0514@gmail.com.

https://doi.org/10.1523/JNEUROSCI.0027-22.2022

Introduction

In multicultural societies, people encounter individuals from their own social group (ingroup members) and from different social groups (outgroup members), and form impressions toward these ingroup and outgroup individuals. As impressions predict behaviors (Tajfel et al., 1971; Amodio and Cikara, 2021), it is important to understand the mechanisms that shape impressions in intergroup contexts.

One basic mechanism that shapes impressions is learning (Tobler et al., 2006; Burke et al., 2010; Lindström et al., 2014; Hein et al., 2016a; Siegel et al., 2018; Kang et al., 2021).

According to learning theory, positive experiences with a person establish positive associations toward this individual, while negative experiences have the opposite effect, driven by unpredicted positive or negative experiences that elicit prediction errors (Pettigrew and Tropp, 2006; Barlow et al., 2012; Lindström and Tobler, 2018; Hackel et al., 2019). The updating of impressions regarding others corresponds to gradually reducing prediction errors by adjusting predictions about them (Hackel and Amodio, 2018; Mende-Siedlecki, 2018). Importantly, in complex social environments, individuals learn from intermixed experiences with both ingroup and outgroup members. However, it remains unclear how such ingroup and outgroup experiences dynamically shape neural learning processes and learning-related changes in impressions toward ingroup and outgroups.

One class of influential social science theories proposes that intergroup impressions are mainly shaped by experiences with outgroup individuals ("outgroup-focused theories"). Emphasizing the importance of outgroup-related information, there is evidence that outgroup experiences change various dimensions of intergroup relationships, including ingroup-outgroup boundaries (Bettencourt et al., 1992; Marcus-Newhall et al., 1993; Dovidio et al., 2003), outgroup prejudice (Chu and Griffey, 1985; Pettigrew and Tropp, 2006; Schlueter and Scheepers, 2010), and empathy for outgroup members (Malhotra and Liyanage, 2005; Paluck, 2009; Hein et al., 2016a).

Alternatively, another class of social science theories emphasizes the importance of ingroup-related information ("ingroup-focused theories"). Ingroup-focused theories predict that changes of intergroup impressions are mainly driven by ingroup experiences ("ingroup-focused theories") because they are of greater relevance for people's social identity than outgroup-related information (Tajfel et al., 1971; Tajfel and Turner, 1979). This is particularly true for individuals who strongly identify with their ingroup (Marques et al., 1988; Branscombe et al., 1993; Mendoza et al., 2014) because a positive perception and evaluation of the ingroup are linked to a positive self-concept (Tajfel et al., 1971; Tajfel and Turner, 1979).

Outside the intergroup context, learning-related updates of impressions have been associated with changes in a neural network, including the mPFC, the inferior parietal lobule (IPL), the ventrolateral PFC (vlPFC), posterior cingulate cortex, the striatum, and the temporoparietal junction (TPJ) (Hackel et al., 2015; Ligneul et al., 2016; Hughes et al., 2017; Mende-Siedlecki, 2018; Park et al., 2021), but it remains unclear whether and how learning from intermixed experiences with ingroup and outgroup individuals change intergroup impressions and underlying neural circuitries.

To address this question, we designed an experiment in which identical experiences with the ingroup and outgroup could dynamically affect the impressions toward both groups. The participants inside the scanner underwent a modified version of a Pavlovian learning experiment. In each learning trial, an individual from the ingroup or the outgroup ostensibly could give up money to save the participant from pain. In reality, participants were relieved from pain in 75% both from ingroup and outgroup individuals. Participants rated their closeness toward the ingroup and the outgroup in each trial, and their ingroup and outgroup impression before and after the experiment (for details, see Fig. 1; Materials and Methods). This setup allowed us to model how participants learn from the same ingroup and outgroup experiences and how these learning processes shape intergroup impressions on the behavioral and neural level.

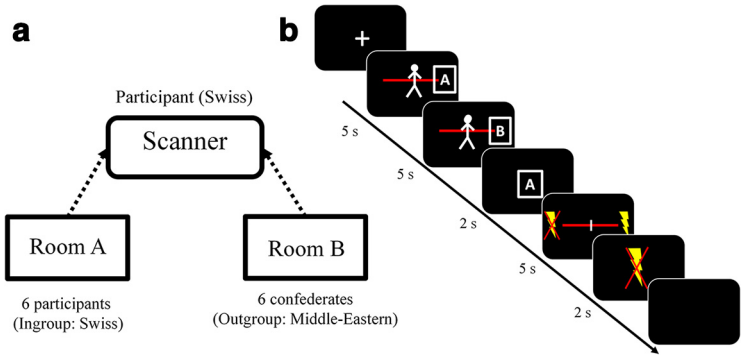

Figure 1. Experimental task. ***a***, Experimental procedure. The Swiss participant in the scanner interacted with both ingroup members (6 Swiss participants, in this example seated in Room A) and outgroup members (6 Middle Eastern confederates, in this example seated in Room B) during the experiment. The assignment of room type (A or B) to social groups (ingroup or outgroup) was counterbalanced across participants. ***b***, Example trial sequence. At the beginning of each trial, the participants rated their closeness toward the ingroup and outgroup on separate rating scales that were presented in randomized order. To do so, they moved a manikin representing themselves toward or away from the room in which ingroup or outgroup members were seated. The next screen revealed the room in which the person that could influence the participant's experience of electric shock was located (e.g., Room "A," which in this example represented the ingroup). Next, participants rated their expectancy of receiving shock, followed by the symbol that represented the ostensible decision of the other person (a crossed-out lightning bolt representing the decision to give up money to save the participant from shocks and a lightning bolt representing the decision to keep the money, resulting in pain for the participant). The experienced outcome (painful shock or no shock) was delivered during the final blank screen.

Based on previous evidence (Tajfel et al., 1971; Hein et al., 2016a), we hypothesized that initially participants would show more positive impressions toward the ingroup compared with the outgroup. Based on the implications of learning theory (Rescorla and Wagner, 1972; Tobler et al., 2006), we anticipated that equal experiences with the ingroup and outgroup reduce the initial ingroup versus outgroup difference in impressions. According to outgroup-focused theories, the effect should be driven primarily by outgroup prediction errors. In this case, the influence of outgroup prediction errors, captured by trial-by-trial changes in closeness ratings, should predict the ingroup versus outgroup impression difference after learning. In contrast, ingroup-focused theories predict that changes in intergroup impressions are primarily linked to ingroup prediction errors, with stronger effects in individuals who identify more strongly with the ingroup. In this case, the ingroup versus outgroup impression difference after learning should be predicted by ingroup prediction errors, captured by respective changes in trial-by-trial closeness ratings. On the neural level, the respective learning processes should be related to activations in the brain regions that have been associated with the updating of impressions, including the mPFC, IPL, vlPFC, posterior cingulate cortex, striatum, and the TPJ (Hackel et al., 2015; Ligneul et al., 2016; Hughes et al., 2017; Mende-Siedlecki, 2018). The resulting change in the ingroup versus outgroup difference in impression may be tracked by the anterior insula (AI), given that this region has been linked to increased outgroup positivity (Hein et al., 2016a).

Materials and Methods

Participants

Thirty healthy males (mean age ± SD = 22.7 ± 2.8 years) volunteered to take part in this study. We chose an all-male instead of a gender-mixed

group of participants so that we could also use all-male confederates and avoid the complications of gender-mixed pairing of participants and confederates. All participants were born and raised in Switzerland and had normal or corrected-to-normal vision. The ingroup consisted of 6 additional male Swiss participants who eventually completed an independent study. The outgroup comprised 6 male confederates whose families originated from the Middle East. The individuals present in any given session did not know each other. One participant completed the fMRI study, but because of technical problems, only the behavioral results but not the imaging data were properly recorded. We report the behavioral and computational modeling results of all 30 participants in the main manuscript and the corresponding behavioral results of 29 participants (mean age ± SD = 22.8 ± 2.8 years). The behavioral and computational modeling findings were robust to the exclusion of the one participant. The imaging results are based on the available 29 imaging datasets. A sensitivity analysis using G*Power 3.1 indicates that given α (5%) and considering four predictors in the regression model, the current sample size ($N = 30$) has 80% power to detect a moderate effect size of $f^2 = 0.28$ and critical $t = 2.06$. The study was approved by the Research Ethics Committee of the Canton (state) of Zurich. All participants provided written informed consent after the experimental procedure had been fully explained. Participants were reminded of their right to withdraw at any time during the study.

Experimental design

Prescanning procedure. At the beginning of the testing session, an experimenter welcomed both the Swiss participants and the Middle Eastern confederates and informed them that the aim of the experiment was to measure cultural differences in decision-making. After the introduction, all participants were assigned a number ranging from 1 to 13 which determined their place in the testing laboratory. The testing laboratory consisted of two rooms that could be separated by a sliding door. Care was taken to ensure that all Swiss participants and all Middle Eastern confederates were placed in different partitions. The separated rooms were named "Room A" and "Room B," and the relationships between the room name and group membership were counterbalanced across participants.

After the room assignment, a well-established priming procedure (Dijksterhuis and van Knippenberg, 1998) served to enhance the social group manipulation. Specifically, participants were asked to write down five attributes that are typical for Middle Eastern males and for Swiss males. This procedure is commonly used to activate group-related stereotypes (Hein et al., 2016a), in our case to activate the stereotype of Middle Eastern males in our Swiss participants.

The 6 Middle Eastern confederates and 6 of the Swiss participants remained in Room A and Room B. The seventh Swiss participant, who had signed up for the fMRI study, was taken to the control room of the fMRI scanner. Here, a pain electrode was attached to the back of the left hand of the participant, and the individual pain threshold was determined based on a standard procedure (Hein et al., 2016a, b). In particular, the participant was asked to rate the intensity of the electrical stimulation he received from 1 (not painful) to 10 (extremely painful). We used the intensity corresponding to a subjective value of 8 as the intensity inside the scanner. Electrical stimulation (bipolar, monophasic; maximum duration: 1000 ms; input range, 5 V; output range, 50 mA) was delivered with a single-current stimulator (DS5; Digitimer). Before entering the fMRI scanner, the participant was informed that he might receive painful stimulation and that in each trial, one of the individuals in Room A or Room B could prevent the upcoming painful stimulation for the participant by giving up money he would otherwise earn.

To avoid possible reputation effects that might influence the behavior of our participants, all ratings were kept anonymous, and we ensured that the participants did not meet the confederates after the experiment. Thus, all ratings and decisions were personal and could not be observed by the other participants. Everyone was also clearly informed that they would not meet after the experiment because the scanned participant needed to stay longer for an anatomic scan.

Scanning procedure. During the fMRI scans, each trial began with a fixation cross presented for 1-7 s. Next, the first closeness rating scale was shown for 5 s, followed by the second closeness rating scale, also presented for 5 s. For both ratings, participants moved a manikin that represented themselves on a 10 step scale toward or away from Room A or Room B to indicate how close they felt toward the individuals in the respective room. To avoid movement preparation and to keep the participants engaged, the manikin was placed at a random location on the scale. The order of closeness ratings for Rooms A and B was randomized across trials.

After the closeness ratings, a letter was shown for 2 s, indicating the room in which the interaction partner (i.e., the person that could ostensibly save the participant from painful shock) was located. Next, the 10 step rating scale was presented for 5 s and participants rated their expectancy of receiving painful shock in the current trial from not likely at all (0, indicated by a crossed-out lightning bolt) to very likely (1, indicated by a lightning bolt). To facilitate the interpretation of the results, the coding of the expectancy ratings was inverted such that 0 corresponds to strongest possible expectation of receiving shocks. To maintain the attention of the participant and reduce motion preparation in each trial, the starting point of the rating scales occurred at a random position. Thereafter, the decision of the other individual was shown for 2 s. A crossed-out lightning bolt symbol indicated that one of the individuals from the respective room had given up 5 CHF to save the participant from painful shock, whereas a lightning bolt symbol indicated upcoming painful stimulation. Based on the decision, participants then either did or did not receive shock during the final blank screen.

Participants were relieved from painful shock in 75% of all trials. Because this ratio was the same in the ingroup and the outgroup condition, the objective experiences with the two groups were equalized, and primarily positive. The learning experiment consisted of three blocks, with 24 trials each, resulting in 72 trials in total.

Questionnaires

Before and after learning, we used an impression scale to assess participants' impressions toward ingroup (Swiss) and outgroup (Middle Eastern) individuals. Participants rated their impressions toward individuals in Room A and Room B separately before and after the learning experiment. The impression scale included questions regarding perceived group membership ("What would be the chance that you would use the word 'we' to describe yourself and these people"; "If you were to meet these people by chance, what are the chances you would become friends?"), regarding perceived similarity ("How similar are these people to you?"; "How much do you have in common with these people?"), and likability ("How likable do you find these persons?"; "How comfortable do you find the idea of meeting these people in the future?") (Hein et al., 2010, 2016a, 2018). We first tested the learning-induced impression changes separately for these three dimensions. The analyses showed pre- to post-learning changes in perceived likability and group membership, indicated by significant time × group interactions ($\chi^2_{(1)} = 5.92$, $p = 0.015$) and ($\chi^2_{(1)} = 4.20$, $p = 0.041$), respectively. In contrast, there were no significant pre- to-post changes in perceived similarity ($\chi^2_{(1)} = 0.24$, $p = 0.63$; see Fig. 2). Based on these results, the individual impression scores before and after the experiment were calculated using the average of the likability and perceived group membership scores (i.e., the two dimensions that showed significant learning-induced changes) (see Fig. 3). Using the average of all items of the impression scale revealed converging results (Table 1). Moreover, before learning, participants completed the ingroup identification questionnaire (Leach et al., 2008), which measures identification and satisfaction with the ingroup in general. Finally, before learning, participants indicated their tendency for outgroup discrimination using the Modern Racism Scale, modified to the Middle Eastern context (McConahay, 1986).

Statistical analyses

Analyses of impression, closeness, and expectancy ratings. We performed linear mixed models (LMM, lme4) (Bates et al., 2015) in R version 4.1.1 for the behavioral analyses on impression, closeness, and expectancy ratings. In particular, we conducted LMMs with group (ingroup/outgroup), time (before/after-learning) and group × time as predictors, and the impression, closeness, and expectancy rating scores as the dependent variable. For the impression ratings, the time variable reflected participants' impression scores before and after the study; for the closeness and expectancy ratings, it reflected ratings in the first and the last trial of the experiment. We used participants as random intercepts and slopes for the fixed effects of group and time. Likelihood ratio tests were applied to assess the significance of the fixed effects. The resulting χ^2 values indicate how much more likely the data are under the assumption of a more complex model (i.e., a model including a particular parameter) than under the assumption of a simpler model (i.e., a model not including this particular parameter). The closeness ratings of 3 participants contained null values, which were replaced by the average closeness rating of all other participants at the corresponding trial.

Computational modeling

Computational modeling of expectancy ratings. First, we modeled participants' trial-by-trial expectancy ratings using a standard Rescorla-Wagner (Rescorla and Wagner, 1972) reinforcement learning (RL) algorithm. The RL model assumes that participants change their expectancy of being saved from shocks when new information reveals that the experienced outcome is different from the expected outcome as follows:

$$V_i(t+1) = V_i(t) + \alpha_i \delta_i \quad (1)$$

$$\delta_i(t) = R(t) - V_i(t) \quad (2)$$

$$Expectancy\,Rating_i = \beta_i \times V_i(t) \quad (3)$$

Therefore, on each trial t, the value of the (future) expectation $V_i(t+1)$ for group i (ingroup or outgroup) is a function of current value $V_i(t)$ and the prediction error δ_i (Eq. 1), which corresponds to the difference between the experienced outcome $R(t)$ at trial t and the current value V_i (Eq. 2). Finally, the current value is multiplied by a response parameter β ($\beta > 0$), which linearly maps the current value to expectancy ratings (Eq. 3). The learning rate α ($0 \leq \alpha \leq 1$) controls the extent to which the current expectancy rating is updated by new information. Consequently, a low learning rate corresponds to a slow integration of prediction errors into current values.

We tested different learning models that assumed the same learning rate and response parameter, the same learning rate and different response parameters, different learning rates and the same response parameter, or different learning rates and response parameters for the ingroup and the outgroup. Moreover, to assess the potential influence of initial expectations and to capture individual variability in initial expectations, we defined RL models in which the initial expectancies ($V1$) were set as free parameter(s). Thus, for every model, we examined a variant with the same or different initial expectancies (i.e., $V1$) toward the ingroup and outgroup, and eight models were tested in total. We identified the maximum likelihood estimate using least squares to fit the model. The winning model was selected based on the Bayesian Information Criterion (BIC) (for details, see Parameter estimation and model comparisons). The winning learning model was the model that included the same learning rate, different response parameters, and same initial expectancy ratings for the ingroup and the outgroup (for plots of expectancy ratings for ingroup and outgroup separately, see Figs. 4*a*, 5*a*).

Computational modeling of closeness ratings. Next, we fitted the trial-by-trial closeness ratings regarding the ingroup or outgroup as a linear function of previous prediction errors (i.e., based on the winning model determined above). As the closeness ratings tended to be auto-correlated between trials, we used the trial-wise difference in closeness ratings to reduce this correlation (a standard method in time series analysis) (Seth, 2010). The change of

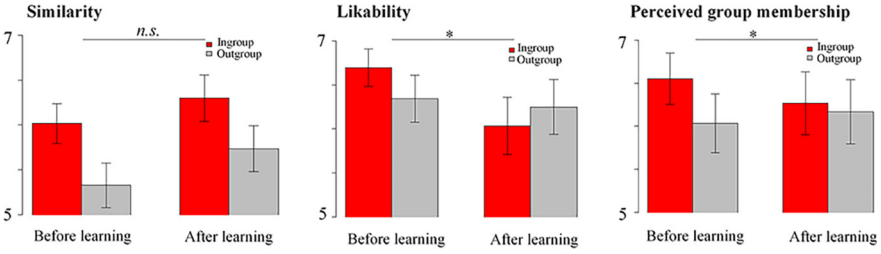

Figure 2. Average ratings before and after learning separately for the three dimensions of the impression scale. Ratings of perceived similarity (left), perceived likability (middle), and perceived group membership (right) for the ingroup (red) and the outgroup (gray) before and after the learning experiment. Larger values indicate higher perceived similarity, perceived likability, and perceived group membership. Investigating the effects on perceived similarity, likability, and perceived group membership separately showed no significant time × group interactions in similarity ($\chi^2_{(1)} = 0.24$, $p = 0.63$), but significant time × group interactions in likability ($\chi^2_{(1)} = 5.92$, $p = 0.015$) and perceived group membership ($\chi^2_{(1)} = 4.20$, $p = 0.041$). These results showed that the learning experiment mainly changed the impressions with regard to perceived likability and group membership. *$p < 0.05$.

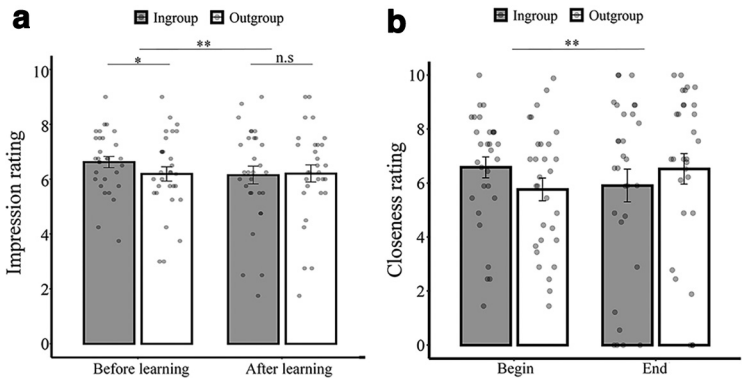

Figure 3. Impression and closeness ratings before and after learning. ***a***, Impression ratings before and after the learning experiment. Larger values represent more positive impressions. ***b***, Closeness ratings at the beginning and end of the experiment. For both types of ratings, the initial ingroup bias was reduced after learning. Error bars indicate SEM. Each small dot represents an individual participant. **$p < 0.01$; *$p < 0.05$.

Table 1. Impression rating change averaged across all three dimensions of the impression scale[a]

		Regressor			Separate analyses (group)	
Rating	Statistic value	Group	Time	Group × Time	Before	After
Impression rating	χ^2	6.16	0.03	4.44	8.02	1.24
	p	0.013	0.86	0.035	0.005	0.26

[a]The time variable indicates the individual impression scores before and after the experiment.

closeness on each trial (t) corresponds to the difference between the closeness ratings on the current trial (t) and on the previous trial (t-1). In all models, we assumed that changes in closeness to group i (i.e., the ingroup or the outgroup) were a linear function of the time-discounted sum of previous prediction errors to outcomes from i (as originating from the RL model, Eqs. 1-3) as follows:

$$\Delta\, Closeness_i(t) = W_i \sum_{j=1}^{t} \gamma^{t-j} \delta_{ij} \tag{4}$$

$$\Delta Closeness_i(t) = W_{pos} \sum_{j=1}^{t} \gamma^{t-j} \delta pos_{ij} + W_{neg} \sum_{j=1}^{t} \gamma^{t-j} \delta neg_{ij} \tag{5}$$

$$\Delta Closeness_i(t) = W_{ingroup} \sum_{j=1}^{t} \gamma^{t-j} \delta_{j(ingroup)} + W_{outgroup} \sum_{j=1}^{t} \gamma^{t-j} \delta_{j(outgroup)} \tag{6}$$

$$\Delta Closeness_i(t) = W_i \sum_{j=1}^{t} \gamma^{t-j} R_{ij} \tag{7}$$

The winning model (Eq. 4, Closeness Model 1) assumes closeness updates to group i to result from previous prediction errors to outcomes arising from i, and not to spill over between groups (e.g., receiving a shock from the outgroup did not predict a subsequent change in ingroup closeness). If trial t entails a decision from an outgroup individual, the closeness updates for the ingroup at that trial result from the time-discounted sum of the previous ingroup related prediction errors plus an additional prediction error of zero. This model has two parameters. The first parameter W captures the magnitude (weight) of influence of prediction errors on changes in closeness. The W parameter ranges from −10 to 10, because 10 represents the maximum of the closeness ratings. Larger W corresponds to stronger influence of prediction errors on closeness updates. The second parameter is the discount parameter γ ($0 \leq \gamma \leq 1$), capturing an exponential decay of the influence of previous prediction errors over time, such that more recent prediction errors have greater impact on the changes in closeness than earlier prediction errors. If γ is close to 1, all preceding prediction errors receive the same weight; and if it is close to 0, only the last prediction error leads to subsequent changes in closeness.

Closeness Model 2 (Eq. 5) is similar to Closeness Model 1, except that we divided the prediction errors by experienced outcome and added them up separately. Accordingly, the model has three free parameters: weight for positive prediction errors from the corresponding group (W_{pos}), weight for negative prediction errors from the corresponding group (W_{neg}), and discounting parameter (γ). Closeness Model 3 (Eq. 6) assumes that unexpected outcomes for the ingroup or the outgroup will lead to a change of closeness to both groups (e.g., receiving a shock from the outgroup could predict a subsequent change in ingroup closeness). This model has also three free parameters: weight for the ingroup prediction errors ($W_{ingroup}$), weight for the outgroup prediction errors ($W_{outgroup}$), and discounting parameter (γ). Finally, we fitted Closeness Model 4 (Eq. 7), in which the change of closeness was determined solely by experienced (actual) outcome (R) to verify that the closeness update depends not only on the experience of shock, but on prediction errors. Equation 7 is similar to Equation 4, except that the sum of previous prediction errors has been replaced with the sum of previous outcome values.

Parameter estimation and model comparisons. To fit the parameters of the different computational models, we identified the maximum likelihood estimate using least squares, which determines the set of parameters that maximize the probability of trial-by-trial expectancy ratings or closeness updates given the specific model. Both the Learning Models and Closeness Models were estimated based on a linear likelihood function. Parameters were independently fitted to each participant using the Broyden-Fletcher-Goldfarb-Shanno optimization method. To avoid local minima in parameter fitting, optimization was initiated with five randomly selected start values. Model implementations and parameter fitting were done in R version 4.1.1. To examine the degree to which the model explained the experimental data, we also calculated the mean squared error over the expectancy and closeness ratings.

We compared models with the BIC, which penalizes the model evidence with model complexity as follows: BIC = −2ln(L) + ln(n)k, where −ln(L) is the negative log-likelihood, n is the number of responses used to compute the likelihood, and k is the number of model parameters. BIC measures are summed across all participants. A smaller BIC indicates a better model fit.

Model identification

We used simulations to assess that our experiment allowed us to dissociate models of interest, as well as identifying parameters of interest within the winning model. We performed a model identifiability analysis by simulating data from 30 participants and fitting this simulated dataset to all the candidate models (Cutler et al., 2021). We applied the same estimation methods that were used for analyzing the experimental dataset and repeated the whole procedure 10 times. For the learning model (8 candidate models in total), α was varied from 0 to 1, and the β parameters were varied from 0 to 3 (the range shown by our participants). For the closeness model (4 candidate models in total), we set the W parameters from −10 to 10, and the γ parameters from 0 to 1. By plotting the confusion matrices of how many times each model won (based on BIC measures, see Fig. 4*e*,*f*), we show that the models are identifiable with our model comparison process.

Parameter recovery

To assess the reliability of the winning model and the interpretability of the free parameters, we also performed parameter recovery on simulated data as recommended for modeling analyses (Palminteri et al., 2017; Zhang et al., 2020; Cutler et al., 2021; Lockwood and Klein-Flügge, 2021). To do this, we computed the Pearson correlation between parameter values that were used for generating simulated data and the values recovered by the model fitting procedure. A high correlation indicates a reliable parameter fitting procedure.

In more detail, we simulated the data 1000 times and applied the same estimation methods that were used for the analyses of the experimental datasets to estimate the parameters. We performed the parameter recovery analyses on the winning models. Similar to the model identification analyses, for the learning model, we varied α from 0 to 1, and bounded the β parameters at 0 and 3 to reflect the range shown by our participants. For the closeness model, we set the W parameters from −10 to 10, and set the γ parameters from 0 to 1. For both RL models and Closeness models, we found high Pearson's correlations between the true simulated and estimated parameter values (learning model: $r > 0.98$; closeness model: $r = 1$), suggesting that our parameter fitting procedure was reliable (van Baar et al., 2019).

Linking parameter estimates to change in intergroup impressions

We performed multiple linear regression analyses in R version 4.1.1 to investigate whether the change in intergroup impressions was mainly driven by the unexpected experiences with outgroup members (as predicted by outgroup-focused theories), with the ingroup members (as predicted by ingroup-focused theories), or with both. The computational parameters (i.e., $W_{ingroup}$, $W_{outgroup}$, $\gamma_{ingroup}$, $\gamma_{outgroup}$) estimated from Closeness Model 1 (see details in the computational modeling section) were entered as predictors. The change of ingroup and outgroup impressions over the learning experiment was entered as the dependent variable (i.e., $(\text{ingroup-outgroup})_{\text{before}} - (\text{ingroup-outgroup})_{\text{after}}$). To confirm that $W_{ingroup}$ predicted intergroup impression change regardless of the initial expectancy, we conducted a regression model with the change of ingroup and outgroup impressions over the learning experiment as the dependent variable (i.e., $(\text{ingroup-outgroup})_{\text{before}} - (\text{ingroup-outgroup})_{\text{after}}$) and $W_{ingroup}$, $W_{outgroup}$, and the *V1* parameter (capturing the value of initial expectancies in learning model) as predictors. Both the predictors and the dependent variable were standardized.

Image acquisition and analyses

The experiment was conducted on a 3-T Philips whole-body MR scanner (Philips Medical Systems), equipped with an eight-channel Philips

SENSitivity Encoded (SENSE) head coil. Structural image acquisition consisted of T1-weighted images (voxel size 1 × 1 × 1 mm). For functional imaging, we used T2*-weighted EPI (number of slices: 40, TR: 2.38 s, voxel size: 3 × 3 × 3 mm, FOV: 240 × 240 mm, TE: 30 ms, flip angle: 90).

Imaging data were analyzed in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). We followed a standardized preprocessing procedure. Functional images were first slice-time corrected, realigned, and coregistered to the anatomic image of the participant. The anatomic image was processed using a unified segmentation procedure combining segmentation, bias correction, and spatial normalization to the MNI template (Ashburner and Friston, 2005); the same normalization parameters were then used to normalize the EPI images. Last, the functional images were spatially smoothed using an isotropic 8 mm FWHM Gaussian kernel.

Whole-brain analyses were conducted to examine the neural correlates of prediction errors. The first-level GLM analyses included the onsets and durations of fixation, closeness ratings, room revelation, outcome expectancy rating, feedback revelation, and shock delivery/omission. Parametric modulators modeled the prediction errors (derived from the RL model for the model-derived analysis and from the difference between the outcome and participants' expectancy ratings for the model-independent analysis) at the feedback revelation. The closeness ratings, room revelation, and feedback revelation were modeled separately for the onsets of ingroup and outgroup trials. These regressors were convolved with the canonical HRF and its time derivatives. Finally, the model contained six (three translation and three rotation) regressors to account for motion.

We assessed prediction error-related activity in a random effect model with one-sample *t* tests on the contrast images created by the parametric modulators. We first analyzed prediction error-related activation independently of group membership. To do so, we weighted both ingroup and outgroup prediction error regressors with a 1 on the first level and used the resulting contrast images to perform a one-sample *t* test against 0 on the second level. The regions encoding positive or negative prediction error were defined as the regions positively or negatively associated with prediction errors (i.e., significantly >0 or <0) in the second-level analysis (i.e., ingroup prediction errors + outgroup prediction errors). We also contrasted ingroup versus outgroup prediction error-related activity at the first level and then compared the resulting images against 0 in the second-level analysis. Imaging results were determined in whole-brain analyses, using a combined voxel-level threshold of $P_{uncorrected} < 0.001$ and a family-wise error (FWE) corrected cluster-level threshold of $p \leq 0.05$.

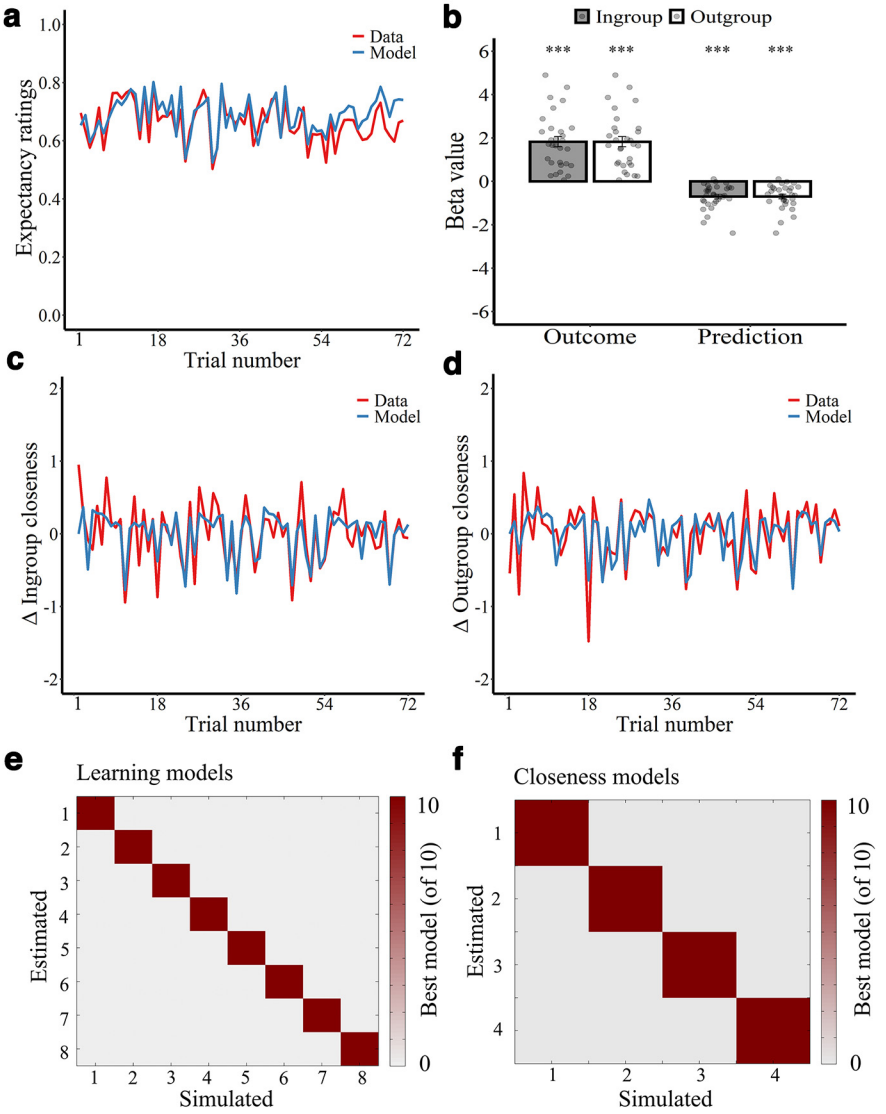

Figure 4. Computational models explain expectancy ratings and changes of ingroup and outgroup closeness ratings. ***a***, Predictions of being saved from shocks (red line) varied over the course of the experiment, and our learning model explained these predictions (blue line). The model estimates illustrate the best fitting model for expectancy ratings. ***b***, Results of LMM analyses. Both experienced and expected outcome influenced change in ingroup and outgroup closeness ratings. ***$p < 0.001$. Error bars indicate SEM. Each small dot represents an individual participant. Trial-by-trial changes of closeness ratings (red) and corresponding model estimates (blue) for the ingroup (***c***) and for the outgroup (***d***). The model estimates illustrate the best fitting Closeness model. ***a***, ***c***, ***d***, show averages over 30 participants. ***e***, ***f***, Model identification analyses for Learning models and Closeness models. Data were simulated 30 times with each candidate model. Then the BIC was applied to select the winning model. This procedure was repeated 10 times, and the confusion matrices showed how many times each model won. Strong diagonals represent reliable model identifiability.

Psychophysiological interaction (PPI) analysis

To examine functional connectivity, we performed a PPI analysis (Friston et al., 1997; McLaren et al., 2012), using the generalized PPI toolbox (https://www.nitrc.org/projects/gppi), in which the inclusion of task regressors reduces the possibility of detecting functional connectivity simply caused by coactivation (McLaren et al., 2012). We extracted the time series of the left IPL cluster (i.e., 5 mm sphere centered on the peak voxel [*x*/*y*/*z* = −24/−55/53]), which showed differences between ingroup and outgroup prediction errors, as the physiological regressor. The condition onset times modeled in the analysis of activity (i.e., fixation, closeness rating, room revelation, outcome expectancy rating, feedback revelation, and shock delivery/omission) were separately convolved with a canonical HRF, creating the psychological regressors. Then the physiological regressor was multiplied with the psychological regressors to obtain the PPI terms (PPI regressor). The physiological, psychological, and PPI regressors as well as six motion parameters were entered into the GLM. We first used this GLM model to determine regions in which connectivity strength with the left IPL was modulated by revealing the ingroup or outgroup outcome (vs the implicit baseline) at the first-level analyses. Thus, we put a weight of 1 on the PPI regressor in which the corresponding psychological regressor was the onset time of ingroup/

outgroup outcome revelation, and a weight of 0 on all other regressors at the first level. Next, we determined regions whose connectivity strength with the left IPL was modulated by the weight given to ingroup and outgroup prediction errors ($W_{ingroup}$/$W_{outgroup}$). To do so, we conducted second-level analyses with the $W_{ingroup}$ and $W_{outgroup}$ as the predictors that were regressed against the contrast images of the PPI regressors acquired from the first-level analyses.

To test the *a priori* assumption that the individual $W_{ingroup}$ parameter alters the functional connectivity between the IPL (i.e., the region that showed an ingroup bias for the processing of negative prediction errors), and regions that process negative prediction errors, we performed small-volume correction (SVC) in the regions encoding negative prediction errors (i.e., all regions where activity survived a combined voxel-level threshold of $P_{\text{uncorrected}} < 0.001$ and a FWE-corrected cluster-level threshold of $p < 0.05$ in the previous whole-brain level analysis). Significant SVC effects were reported using a combined voxel-level threshold of $P_{\text{uncorrected}} < 0.001$ and corrected cluster-level threshold of $p \leq 0.05$ based on a FWE correction.

To evaluate the correlation between the left IPL–left AI coupling and the impression change, we extracted the connectivity strength for the left AI regions identified by the PPI analysis (left AI: peak coordinate: −45/2/5). The β values of the ROI were extracted using MarsBaR (http://marsbar.sourceforge.net). We performed a regression analysis to test whether the association between the weight parameter (i.e., $W_{ingroup}$ or $W_{outgroup}$) and the left IPL–left AI coupling was significantly stronger when participants received feedback from the ingroup compared with the outgroup. Group (ingroup vs outgroup) was entered as a dichotomous predictor (0 representing ingroup and 1 representing outgroup). Weight parameters and group were then standardized and multiplied to generate the interaction term. The weight parameters, the group, and the interaction term were entered into the regression model to test the association with left IPL– left AI coupling (i.e., connectivity strength extracted above).

Bayes factor analyses

To interrogate null findings in our study, we conducted Bayes factor analyses. We calculated the Bayes factor in the program R version 4.1.1 (www.r-project.org) using the functions ttestBF and lmBF from the package BayesFactor (Morey et al., 2015). We based the Bayes factor analyses on the default priors for paired *t* test and regression designs (scale r on an effect size of 0.707). A Bayes factor indicates how much more likely the alternative hypothesis is, compared with the null.

Model-independent analysis testing the effect of positive and negative ingroup and outgroup prediction errors

We performed a model-independent analysis to investigate how positive and negative ingroup and outgroup experiences influenced intergroup impression change. To this end, we calculated the change of closeness (ΔCloseness) in each trial. The change of closeness on each trial (*t*) corresponds to the difference between the closeness ratings on the current trial (*t*) and on the previous trial (*t*-1). We then divided them by the size of prediction errors of the previous trial (i.e., Index = ΔCloseness/Prediction errors). This is because participants always perform the closeness rating at the start of the trial (Fig. 1), so the prediction error of the previous trial is the most recent prediction error that drives the closeness change. We then averaged the Index separately for positive ingroup (Index pos-in), negative ingroup (Index neg-in), positive outgroup (Index pos-out), and negative outgroup (Index neg-out) experiences for each individual. Larger values of this index reflect a greater change in closeness because of unexpected events. Next, we tested a linear regression model in which the Index neg-in, Index neg-out, Index pos-in, and Index pos-out were entered as independent variables to predict the intergroup impression change and the left IPL–left AI connectivity strength. As two outliers were identified for the Index value (i.e., >3 SDs), we also performed robust linear regression (rlm function in R package MASS) to control the influence of outliers. The robust linear regression returned similar results.

Data and code availability

Behavioral data and analysis code to reproduce the figures are available at: https://osf.io/za29y/.

Results

Behavioral impression and closeness change

To assess how learning affected ingroup and outgroup impressions, participants completed a standardized impression scale for the ingroup and outgroup before and after the learning experiment (for details, see Materials and Methods). The individual impression scores before and after the experiment were calculated using the average of the likability and perceived group membership scores, that is, the two dimensions that showed significant learning-induced changes (for details of the impression scale, see Fig. 2; Materials and Methods). We conducted a LMM with group (ingroup/outgroup), time (before/after-learning), and group × time as predictors, and the rating scores on the impression scale as dependent variable. The results revealed a significant group × time interaction ($\chi^2_{(1)} = 7.34$, $p = 0.007$, $\beta = -0.49$, SE = 0.18; for converging results using the average of all items of the impression scale, see Fig. 3*a*; Table 1). Clarifying the interaction effect, separate analyses showed significantly more positive impressions toward ingroup compared with outgroup members before learning ($\chi^2_{(1)} = 5.78$, $p = 0.02$, $\beta = 0.44$, SE = 0.18), but no significant effect of group after learning ($\chi^2_{(1)} = 0.20$, $p = 0.66$, $\beta = 0.06$, SE = 0.13, Bayes factor = 0.21). We also determined participants' closeness ratings toward the ingroup and outgroup at the beginning (the first trial) and the end (the last trial) of the experiment. We observed a significant group × time interaction in closeness ratings ($\chi^2_{(1)} = 6.87$, $p = 0.009$, $\beta = 1.44$, SE = 0.55, Fig. 3*b*). Elucidating the interaction effect, separate analyses revealed higher closeness ratings for the ingroup compared with the outgroup before learning ($\chi^2_{(1)} = 3.91$, $p = 0.048$, $\beta = -0.82$, SE = 0.41), but no significant group difference after learning ($\chi^2_{(1)} = 2.86$, $p = 0.091$, $\beta = 0.62$, SE = 0.37, Bayes factor = 0.69). Thus, both impression and closeness ratings showed group-related changes induced by the experiences in the experiment.

Behavioral results of expectancy ratings

We next tested whether participants showed differential expectancy ratings toward the ingroup and the outgroup. To this end, we averaged expectancy ratings across the tasks separately for the ingroup and the outgroup. Overall, participants underestimated the chance of being saved from shocks (real probability = 0.75; average expectancy = 0.66, $t_{(29)} = -3.1$, $p = 0.004$, 95% CI = [0.61, 0.72]), with no differences between the ingroup and the outgroup condition (paired *t* test: $t_{(29)} = -0.015$, $p = 0.88$, 95% CI = [−0.05, 0.05], Bayes factor = 0.20). To estimate participants' initial expectation regarding the ingroup and the outgroup, we analyzed the first expectancy ratings. The initial expectation of being saved was lower than the real probability (i.e., 0.75) for both the ingroup (mean = 0.64, $t_{(29)} = -2.26$, $p = 0.03$, 95% CI = [0.53, 0.74]) and the outgroup (mean = 0.61, $t_{(29)} = -2.55$, $p = 0.02$, 95% CI = [0.49, 0.72]), with no significant differences between groups (paired *t* test: $t_{(29)} = 0.54$, $p = 0.60$, 95% CI = [−0.08, 0.14], Bayes factor = 0.22). There was no significant difference in the SDs of the expectation ratings across trials between the ingroup and the outgroup ($t_{(29)} = -0.27$, $p = 0.79$, 95% CI = [−0.02, 0.02], Bayes

factor = 0.20), indicating comparable variation in the expectation of being saved from shocks by the ingroup and the outgroup.

To examine the change of expectancy rating in the course of the task, we used a LMM that included group (ingroup/outgroup), time (first trial/last trial), and group × time as predictors, and the expectancy ratings as the dependent variable. The results revealed a significant group × time interaction ($\chi^2_{(1)}$ = 5.32, p = 0.021, β = 0.11, SE = 0.05), reflecting a significantly larger ingroup versus outgroup difference before learning (0.64 vs 0.61, i.e., at the beginning of the experiment) than after learning (0.60 vs 0.68, i.e., at the end of the experiment). The mean values point to a learning-related decrease of expectations to receive help from the ingroup (pre/post: 0.64/0.60), and increased expectation of receiving help from the outgroup (pre/post: 0.61/0.68). However, the pre- to-post changes in expectancy ratings within each group were not significant (ingroup pre vs post: $\chi^2_{(1)}$ = 0.37, p = 0.54, β = 0.04, SE = 0.06; outgroup pre vs post: $\chi^2_{(1)}$ = 1.46, p = 0.23, β = −0.08, SE = 0.06).

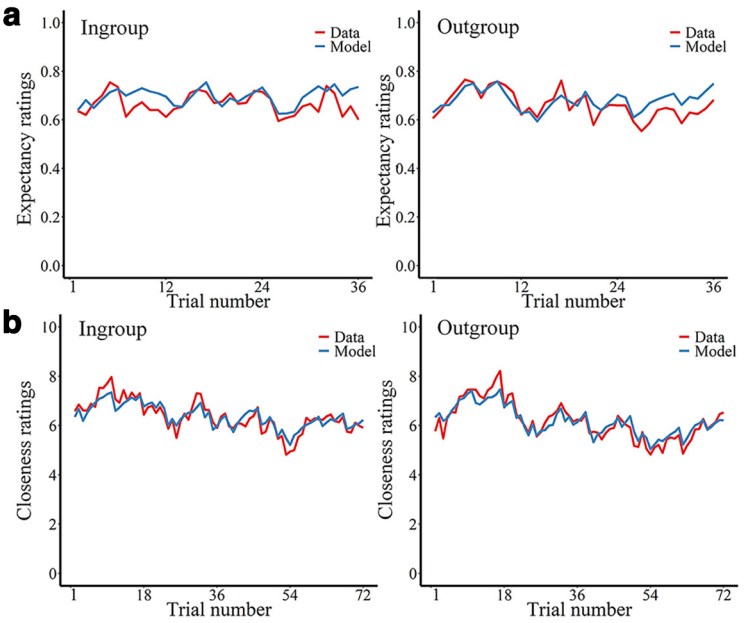

Figure 5. Computational models explain expectancy and closeness ratings for both ingroup and outgroup. ***a***, Predictions of being saved from shocks (red line) varied over the course of the experiment, and our learning model explained these predictions (blue line) for the ingroup (left) and for the outgroup (right). A larger value indicated a higher expectation of being saved from shocks. The model estimates illustrate the best fitting model for expectancy ratings. ***b***, Trial-by-trial closeness ratings (red line) and corresponding model estimates (blue line) for the ingroup (left) and for the outgroup (right). The figure shows averages over 30 participants.

Computational processes underlying dynamic changes in expectancy ratings and ingroup and outgroup closeness ratings

To understand how the learning experience changed ingroup and outgroup impressions, we examined the computational processes underlying the dynamic changes in ratings. First, we modeled participants' trial-by-trial expectancy ratings using a Rescorla-Wagner RL model (Rescorla and Wagner, 1972) to test for differences in basic learning mechanisms between the ingroup and outgroup conditions. To further control the influence of initial expectations and to capture individual variability in initial expectations, we defined RL models in which the initial expectancies were entered as free parameters. We first verified good model recovery with respect to our main model space, which allowed us to meaningfully compare the different learning models (Fig. 4*e*). Model comparison revealed that expectancy ratings concerning the ingroup and outgroup were best characterized by a model with the same learning rate, different response parameters, and the same initial expectancy for the ingroup and outgroup (r^2 = 0.37 ± 0.27; mean ± SD; Fig. 4*a*, see Fig. 5*a* for the expectancy ratings for ingroup and outgroup separately, see Table 2 and Materials and Methods for details of model comparisons). Independent parameter recovery analyses using the winning learning model with simulated data showed good recovery (r > 0.98) over a wide parameter space.

The average response parameters estimated from the winning learning model did not differ between the ingroup and the outgroup ($t_{(29)}$ = −0.47, p = 0.645, 95% CI = [−0.22, 0.14], Bayes factor = 0.21). To examine individual differences in response parameters between the ingroup and outgroup regardless of direction, we subtracted the response parameters of the ingroup from those of the outgroup and examined the absolute difference. We found that the absolute differences were significantly >0 ($t_{(29)}$ = 3.51, p = 0.001, 95% CI = [0.11, 0.40]), indicating that it is necessary to use different response parameters at the individual level for modeling and that individuals show group-related variation in response parameters.

Table 2. Comparisons of model fits for expectancy ratings[a]

Computational models	*K*	Mean r^2	BIC	BIC[1b]
Expectancy ratings				
1 learning rate, 1 response parameter, 1 initial expectancy	3	0.29	−7915	−7582
1 learning rate, 1 response parameter, 2 initial expectancies	4	0.36	−8060	−7727
1 learning rate, 2 response parameters, 1 initial expectancy[c]	4	0.37	−8136	−7794
1 learning rate, 2 response parameters, 2 initial expectancies	5	0.39	−8116	−7779
2 learning rates, 1 response parameter, 1 initial expectancy	4	0.36	−7965	−7630
2 learning rates, 1 response parameter, 2 initial expectancies	5	0.38	−7952	−7621
2 learning rates, 2 response parameters, 1 initial expectancy	5	0.39	−8018	−7677
2 learning rates, 2 response parameters, 2 initial expectancies	6	0.40	−7964	−7628

[a]BIC measures are summed across all participants. Lower BIC values indicate better model fit. Mean squared error over expectancy ratings indicates goodness of fit. *K* is the number of free parameters in each model.
[b]BIC[1] indicates the BIC values based on 29 participants.
[c]The winning model.

Next, we tested how ingroup and outgroup prediction errors (i.e., the difference between received feedback and expected feedback) affected trial-by-trial changes in ingroup and outgroup closeness ratings. As the ratings tended to be auto-correlated across trials, we used the trial-wise difference in closeness ratings to reduce the correlation (a standard method in time series analysis) (Seth, 2010).

We first tested, in a model-independent manner, the hypothesis that dynamic changes in closeness ratings depend both on the nature of the outcome and the expectations about the outcome. If prediction errors explain changes in closeness better than the outcome alone, the trial-wise updates of closeness should not only be positively correlated with the nature of the outcome (a positive outcome corresponds to positive experiences related to not receiving shock) but also negatively correlated with the expected outcome (Rutledge et al., 2014; Will et al., 2017). To

Table 3. Comparisons of model fits for closeness ratings[a]

Computational models	*K*	Mean r^2	BIC	BIC[1b]
Ingroup closeness ratings				
Closeness Model 1: Group-specific prediction error weights[c]	2	0.20	194	122
Closeness Model 2: Outcome-specific prediction error weights	3	0.22	280	204
Closeness Model 3: Prediction error weights from both groups	3	0.21	286	211
Closeness Model 4: Experienced outcome only	2	0.19	308	429
Outgroup closeness ratings				
Closeness Model 1: Group-specific prediction error weights[c]	2	0.20	86	37
Closeness Model 2: Outcome-specific prediction error weights	3	0.22	185	135
Closeness Model 3: Prediction error weights from both groups	3	0.21	207	152
Closeness Model 4: Experienced outcome only	2	0.17	267	271

[a]BIC measures are summed across all participants. Lower BIC values indicate better model fit. Mean squared error over change of closeness ratings to ingroup or outgroup indicates goodness of fit. *K* is the number of free parameters in each model.
[b]BIC[1] indicates the BIC values based on 29 participants.
[c]The winning models.

this end, we first fitted an LMM to test the influence of experienced and expected outcome on changes of closeness. As hypothesized, for changes of closeness toward both the ingroup and outgroup, we found a significant positive effect with experienced outcome (Fig. 4*b*, ingroup: $\chi^2_{(1)} = 49.57$, $p < 1 \times 10^{-12}$, $\beta = 1.83$, SE = 0.26; outgroup: $\chi^2_{(1)} = 37.98$, $p < 1 \times 10^{-10}$, $\beta = 1.65$, SE = 0.27), but a negative effect with expected outcome (Fig. 4*b*, ingroup: $\chi^2_{(1)} = 10.55$, $p = 0.001$, $\beta = -0.70$, SE = 0.22; outgroup: $\chi^2_{(1)} = 18.89$, $p < 1 \times 10^{-5,}$ $\beta = -0.89$, SE = 0.20). Thus, the two main constituents of prediction errors (i.e., experienced and expected outcomes) explain trial-by-trial changes in closeness ratings.

Next, we formally modeled the trial-by-trial update of ingroup and outgroup closeness ratings as a linear function of the cumulative impact of prediction errors, which we estimated with the best fitting RL model. We initially verified good model recovery with respect to our main model space, which allows us to meaningfully compare the different closeness models (Fig. 4*f*). Our winning closeness model (Eq. 4, Closeness Model 1) successfully captured dynamic changes of closeness ratings at the individual level for both the ingroup ($r^2 = 0.20 \pm 0.17$; mean ± SD; see Fig. 4*c*) and outgroup ($r^2 = 0.20 \pm 0.18$; mean ± SD; see Fig. 4*d*). The winning model assumes that the changes in closeness to group *i* (i.e., the ingroup or the outgroup) for each trial *t* are driven by the time-discounted sum of previous prediction errors to outcomes from the corresponding group *i*. We chose this model because it outperformed a range of alternatives, including a model that solely considered experienced outcome, but not outcome expectations (for details of model comparisons, see Table 3) as follows:

$$\Delta Closeness_i(t) = W_i \sum_{j=1}^{t} \gamma^{t-j} \delta_{ij} \quad (4)$$

The winning model (Eq. 4, Closeness Model 1) has two free parameters to capture the changes in closeness to each group: the weight W ($-10 \leq W \leq 10$, bounded by the rating scales), which determines the average influence of prediction errors (δ_{ij}) arising from group *i* on changes in closeness toward group *i*, and the discount parameter γ ($0 \leq \gamma \leq 1$), which determines how steeply older prediction errors are discounted (for additional details, see Materials and Methods). If γ is close to 1, all preceding prediction errors receive the same weight; and if it is close to 0, only the most recent prediction error leads to subsequent changes in closeness. Analyses on simulated data showed good parameter recovery ($r = 1$) for the winning closeness model over a wide parameter space.

Table 4. Computational model parameters of the winning closeness model[a]

Computational parameters	Mean (SD)	Mean (SD)[1b]
Weight on prediction error from ingroup ($W_{ingroup}$)	1.68 (1.34)	1.71 (1.35)
Weight on prediction error from outgroup ($W_{outgroup}$)	1.61 (1.42)	1.63 (1.45)
Discounting factor for ingroup ($\gamma_{ingroup}$)	0.16 (0.31)	0.16 (0.32)
Discounting factor for outgroup ($\gamma_{outgroup}$)	0.10 (0.24)	0.10 (0.25)
Learning rate (α)	0.13 (0.10)	0.13 (0.10)
Response parameter for ingroup ($\beta_{ingroup}$)	1.27 (0.70)	1.31 (0.68)
Response parameter for outgroup ($\beta_{outgroup}$)	1.31 (0.76)	1.35 (0.75)
Initial expectancy value (V1)	0.35 (0.24)	0.33 (0.22)

[a]Initial expectancy values are coded such that zero corresponds to the strongest possible expectation of being saved from pain.
[b]Mean (SD)[1] indicates the model parameters of the winning model based on 29 participants.

We then fitted the changes in ingroup and outgroup closeness ratings separately and estimated group-specific parameters ($W_{ingroup}$, $W_{outgroup}$, $\gamma_{ingroup}$, $\gamma_{outgroup}$). The parameters estimated by the winning model (the weight parameter W and the discount parameter γ) did not differ between the ingroup and outgroup (Table 4, W: $t_{(29)} = 0.43$, $p = 0.699$, 95% CI = [−0.26, 0.40], Bayes factor = 0.21; γ: $t_{(29)} = 0.849$, $p = 0.403$, 95% CI = [−0.08, 0.19], Bayes factor = 0.27). We also confirmed that prediction errors, rather than the experience of shock per se, gave the best account of the data (Table 3, Ingroup: ΔBIC = 114, Outgroup: ΔBIC = 181). Together, these results demonstrate that ingroup and outgroup closeness is dynamically updated based on prediction errors that are elicited by the behavior of ingroup and outgroup members.

We also modeled the trial-by-trial raw closeness ratings based on the winning model (Closeness Model 1) by adding a free parameter *W0* for the baseline closeness ratings for the ingroup or the outgroup. We then extracted the weight parameters estimated in this raw closeness model and correlated them with the weight parameters identified in the original differential closeness model. The weight parameters were highly correlated for both the ingroup ($r = 0.939$, $p < 0.001$) and the outgroup ($r = 0.808$, $p < 0.001$), suggesting that the individual variance in weight parameters is similar for differential closeness ratings (capturing closeness change) and raw closeness ratings (for plots of raw closeness rating, see Fig. 5*b*).

Comparing the influence of ingroup and outgroup prediction errors

So far, our results indicate that ingroup and outgroup experiences both generated prediction errors that resulted in learning-related updates of ingroup and outgroup closeness. Next, we investigated how the learning from ingroup and outgroup experiences relates to the changes in intergroup impressions shown in Figure 3*a*. To this end, we tested whether the observed update in impressions was mainly driven by the experiences with the outgroup members (as predicted by outgroup-focused theories), by experiences with the ingroup members (as predicted by ingroup-focused theories), or by both.

To address this question, we conducted a multiple linear regression analysis with the changes of ingroup and outgroup impressions before and after learning as dependent variable, that is, $(\text{ingroup-outgroup})_{\text{before}} - (\text{ingroup-outgroup})_{\text{after}}$. As independent variables, we used the $W_{ingroup}$ and $W_{outgroup}$ parameters (reflecting the weight with which ingroup and outgroup prediction errors affect the updates in ingroup and outgroup closeness), as well as $\gamma_{ingroup}$ and $\gamma_{outgroup}$ parameters (reflecting how steeply older prediction errors are discounted). The variance inflation factor for all variables in the regression was < 3.1,

Table 5. Predicted change in intergroup impressions from model parameters

Parameter	β (SE)	t	p	p^{1a}
Intergroup impression change (average of perceived likability and group membership)				
$W_{ingroup}$[b]	0.74 (0.29)	2.56	0.017	0.012
$W_{outgroup}$	−0.20 (0.29)	−0.68	0.501	0.507
$\gamma_{ingroup}$	0.10 (0.20)	0.48	0.635	0.500
$\gamma_{outgroup}$	0.07 (0.18)	0.39	0.703	0.638
Intergroup impression change (average of all three dimensions of the impression scale)				
$W_{ingroup}$[b]	0.65 (0.24)	2.25	0.034	0.027
$W_{outgroup}$	−0.01 (0.29)	−0.03	0.974	0.983
$\gamma_{ingroup}$	0.18 (0.20)	0.90	0.377	0.299
$\gamma_{outgroup}$	0.09 (0.18)	0.52	0.608	0.558

[a] p^1 indicates p value based on the impression scores from 29 participants.
[b] The $W_{ingroup}$ parameter uniquely predicted the intergroup impression change.

indicating that multicollinearity was not a concern. The results revealed that the reduction of the ingroup bias in impression ratings was uniquely predicted by the $W_{ingroup}$ parameter (β = 0.74, SE = 0.29, t = 2.56, p = 0.017, 95% CI [0.14, 1.34]; Table 5; Fig. 6*a*). Importantly, the more strongly an individual weighted ingroup prediction errors, the more pronounced the reduction of ingroup bias in impression ratings was (Fig. 6*b*). By contrast, the $W_{outgroup}$ parameter did not significantly predict the decline of ingroup favoritism (β = −0.20, SE = 0.29, t = −0.68, p = 0.501, 95% CI [−0.81, 0.40], Bayes factor = 0.42, Fig. 6*a*). These results show that a decrease of ingroup closeness because of a violation of ingroup predictions can result in a decline of ingroup favoritism. This finding was replicated if all items of the impression scale were included (Table 5).

To ascertain the specificity of these results, we also examined whether the computational parameters estimated from the RL model associate with the changes of impressions. None of the parameters in the RL model [i.e., α (learning rate), β (response parameters), and *V1* (initial expectancy)] correlated with the change of impressions ($r < 0.351$, $p > 0.057$, Bayes factor < 1.46; $\rho < 0.29$, $p > 0.12$). To further confirm that $W_{ingroup}$ predicted intergroup impression change regardless of the initial expectancy, we conducted a regression model in which the $W_{ingroup}$, $W_{outgroup}$ and *V1* parameters (capturing the value of initial expectancies in learning model) were entered as independent variables to predict the impression change. The results showed a significant effect of $W_{ingroup}$ (β = 0.79, SE = 0.27, t = 2.92, p = 0.007, 95% CI [0.23, 1.34]), whereas the $W_{outgroup}$ (β = −0.33, SE = 0.27, t = −1.22, p = 0.23, 95% CI [−0.89, 0.23], Bayes factor = 0.63) and *V1* (β = −0.19, SE = 0.16, t = −1.13, p = 0.27, 95% CI [−0.53, 0.15], Bayes factor = 0.58) parameters were not related to the change of impressions.

So far, the results of the regression analysis are in line with the ingroup-focused theories. According to one central prediction of the ingroup-focused theories, learning from ingroup experiences should be stronger in individuals with stronger ingroup identification because they react more strongly to violations of ingroup expectations (Branscombe et al., 1993; Biernat et al., 1999; Mendoza et al., 2014). Correlating the individual scores of the ingroup identification scale (Leach et al., 2008) with the individual $W_{ingroup}$, $W_{outgroup}$ parameters and the intergroup impression change revealed a significant relation with $W_{ingroup}$ (r = 0.47, $p_{(\text{Bonferroni-corrected})}$ = 0.024, $p_{(\text{uncorrected})}$ = 0.008, 95% CI = [0.136, 0.712]; Fig. 6*c*), and intergroup impression change (r = 0.44, $p_{(\text{Bonferroni-corrected})}$ = 0.042, $p_{(\text{uncorrected})}$ = 0.014, 95% CI = [0.100, 0.694]; Fig. 6*d*), but not with $W_{outgroup}$ (r = 0.25, $p_{(\text{Bonferroni-corrected})}$ = 0.543, $p_{(\text{uncorrected})}$ = 0.181, 95% CI = [−0.120, 0.560], Bayes factor = 0.69). In contrast, individual differences in outgroup dislike (assessed with a modified version of the modern racism scale, McConahay,1986) were not related to either $W_{ingroup}$ (r = −0.07, $p_{(\text{uncorrected})}$= 0.704, $p_{(\text{Bonferroni-corrected})}$ = 1, 95% CI [−0.421, 0.296], Bayes factor = 0.36) or $W_{outgroup}$ (r = 0.21, $p_{(\text{uncorrected})}$ = 0.270, $p_{(\text{Bonferroni-corrected})}$ = 0.540, 95% CI [−0.164, 0.528], Bayes factor = 0.56).

Neural mechanisms underlying the ingroup-focused theories

Having established that our computational model of dynamic changes in group identification can explain trial-by-trial changes in closeness and predict change in intergroup impressions, we investigated the neural implementation of the model, by identifying brain activity corresponding to the estimated mechanisms and parameters. First, we investigated neural regions that encoded positive or negative prediction errors in general (i.e., regardless of the group membership manipulation). To do so, we regressed the prediction error estimates against neural activity when the decision of the ingroup or outgroup individual was revealed. Negative prediction errors (i.e., shock prevention expected but shock delivered) were related to activity in dorsal anterior cingulate cortex (dACC) and AI. The more negative the prediction errors, the larger the neural responses in these regions (Fig. 7*a*; Table 6). Positive prediction errors (i.e., shock expected but prevented) correlated with activity in striatum, dorsomedial PFC (dmPFC), ventromedial PFC (vmPFC), and ventrolateral PFC (vlPFC) (Fig. 7*b*; Table 6). These classical learning regions showed stronger neural responses the more positive the prediction error.

Next, we contrasted the encoding of prediction errors between group conditions. We found a significant difference in the left IPL, indicating a stronger association with negative prediction errors for the ingroup compared with the outgroup (MNIxyz: −24/−55/53, Z_{stats}= 4.38, $P_{\text{FWE whole-brain corrected}}$ = 0.012) (Fig. 7*c*,*d*). The reverse contrast did not show any significant results even at the very liberal threshold of $P_{\text{uncorrected}} < 0.05$ (whole-brain). A model-independent analysis, defining prediction errors as the difference between the outcome and participants' expectancy ratings, revealed similar regions as those identified with model-derived prediction errors (Fig. 8).

Our behavioral results suggest that the individual differences in the weight of the ingroup prediction error signal ($W_{ingroup}$) predict the change in intergroup impressions. We reasoned that the weight given to the ingroup prediction errors could modulate the neural response in the IPL region that encoded ingroup prediction errors more strongly than outgroup prediction errors. However, a second-level regression of the neural response elicited by ingroup prediction errors on the individual $W_{ingroup}$ parameters revealed no significant effects even at the very liberal threshold of $P_{\text{uncorrected}} < 0.05$ (whole-brain). Thus, ingroup prediction errors appear to be encoded independently of the individual weight given to the ingroup prediction error.

Alternatively, it is possible that the individual $W_{ingroup}$ parameter alters the functional connectivity between the IPL (i.e., the region that showed an ingroup bias for the processing of negative prediction errors), and regions that process negative ingroup and outgroup prediction errors. We defined an ROI based on the entire neural network that was involved in the processing of negative prediction errors (Table 6; Fig. 7*a*) and conducted a generalized PPI analysis (Friston et al., 1997; McLaren et al., 2012) to test whether the connectivity between the IPL and regions

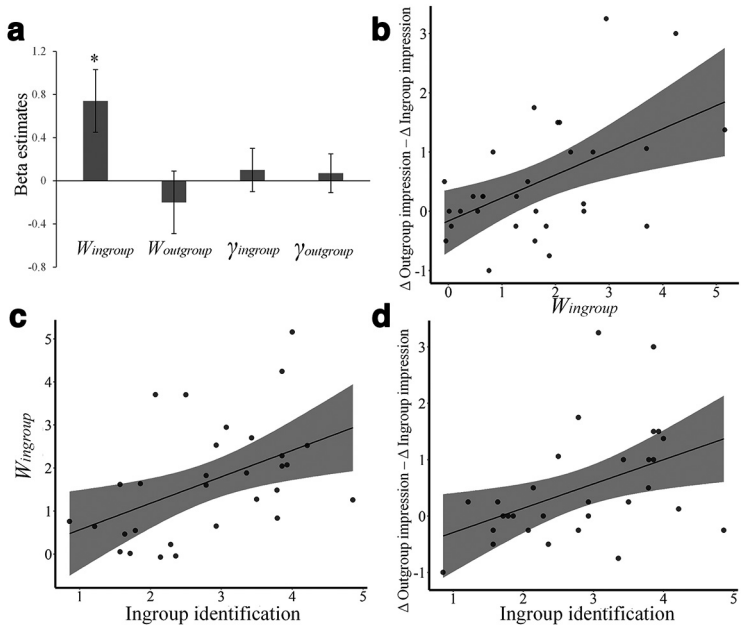

Figure 6. Behavioral results support ingroup-focused theories of intergroup impression change. ***a***, Linear regression predicting impression change (Δoutgroup impression – Δingroup impression) with computational model parameters. The regression coefficient for $W_{ingroup}$ was significant, but that for $W_{outgroup}$ was not. Error bars indicate SEM. ***b***, Correlation between $W_{ingroup}$ parameter and decline of ingroup favoritism as measured by impression rating. ***c***, Correlation between ingroup identification and the $W_{ingroup}$ parameter. ***d***, Correlation between ingroup identification and decline of ingroup favoritism as measured by impression ratings. *$p < 0.05$.

encoding negative prediction errors is influenced by the individual $W_{ingroup}$ parameter. We estimated the connectivity strength between the left IPL seed (physiological regressor) and regions processing negative prediction errors when participants observed the ingroup feedback (psychological regressor, i.e., when they found out whether the ingroup individual was willing to help or not) in the first-level analysis. We then conducted a second-level regression analysis to test connections that were specifically associated with individual differences in $W_{ingroup}$ parameters after regressing out the influence of $W_{outgroup}$ parameters. Only connectivity between the left IPL and the left AI (MNIxyz: −45/2/5, Z_{stats}= 3.88, p (SVC-FWE) = 0.016) was modulated by $W_{ingroup}$ (Fig. 9*a*). Specifically, individuals who weighted ingroup prediction errors more strongly when they updated their subjective closeness toward the ingroup (i.e., larger $W_{ingroup}$ parameters) also showed increased left IPL–left AI coupling when receiving feedback from the ingroup.

To examine whether this effect was specific for ingroup experience, we also conducted whole-brain level regression analyses with $W_{outgroup}$ as predictor (regressing out the influence of $W_{ingroup}$ parameters) on the connectivity strength between left IPL and the rest of the brain at the onset of outgroup feedback revelation, and found no significant regions even at lenient thresholds. To further validate this difference, we performed a regression analysis with the weight parameters ($W_{ingroup}$, $W_{outgroup}$), group (ingroup/outgroup), and weight parameter × group as predictors and IPL–AI connectivity strength as dependent variable. The results revealed a significant weight parameter × group interaction (β = 0.37, SE = 0.12, t = 3.08, p = 0.003), reflecting a stronger effect of $W_{ingroup}$ compared with $W_{outgroup}$ on the strength of IPL–AI connectivity (Fig. 9*b*). According to these results, the association between weight parameters and the strength of the connectivity between left IPL and left AI was significantly stronger when participants received ingroup compared with outgroup feedback. These results corroborate our behavioral regression analysis, by demonstrating that the $W_{ingroup}$ parameters uniquely modulated the neural connectivity between left IPL and left AI during ingroup feedback revelation and support the ingroup-focused theories at the neural level.

Finally, as people who showed larger $W_{ingroup}$ also displayed greater decline of ingroup favoritism after learning, we tested whether the individual strength of the left IPL to left AI connectivity during ingroup feedback revelation is related to the decrease of ingroup favoritism in impression rating. Correlation analyses showed that the connectivity strength between left IPL and left AI was positively associated with the reduction of ingroup favoritism (r = 0.39, p = 0.037, 95% CI = [0.025, 0.661]; Fig. 9*c*). Moreover, there was no significant correlation between the left IPL–left AI coupling and impression change when participants received feedback from the outgroup (r = 0.10, p = 0.614, 95% CI = [−0.279, 0.448], Bayes factor = 0.38). These results suggest that the neural connectivity between left IPL and left AI during the ingroup feedback was preferentially related to the intergroup impression change.

Investigating the effect of positive and negative ingroup and outgroup prediction errors

So far, we demonstrated that the weight given to ingroup experiences (ingroup prediction errors) was the main source of intergroup impression change, captured by changes in connectivity between the left IPL and left AI. To investigate whether and how the valence of experiences influence the intergroup impression change, we conducted an additional analysis that explicitly assessed the weight of positive and negative ingroup and outgroup prediction errors on the update of closeness rating. To do so, we computed an index (ΔCloseness/Prediction errors) for each trial, and averaged separately for positive ingroup prediction errors (Index pos-in), negative ingroup prediction errors (Index neg-in), positive outgroup prediction errors (Index pos-out), and negative outgroup prediction errors (Index neg-out) for each individual. Larger values of this index reflect a greater change in closeness because of unexpected positive and negative ingroup and outgroup experiences.

Next, we tested a linear regression model in which the Index neg-in, Index neg-out, Index pos-in, and Index pos-out were entered as independent variables to predict the intergroup impression change and the left IPL–left AI connectivity strength. According to the results, only the Index neg-in was significantly associated with the decrease of ingroup favoritism (β = 1.09, SE = 0.318, t = 3.43, p = 0.002, 95% CI [0.44, 1.75]), and the increase of left IPL–left AI coupling when receiving feedback from the ingroup (β = 1.13, SE = 0.31, t = 3.64, p = 0.001, 95% CI [0.49, 1.77]). Together, these results suggest that particularly the negative and unexpected experiences with the ingroup drive the impression change, reflected by the connectivity strength of left IPL and left AI.

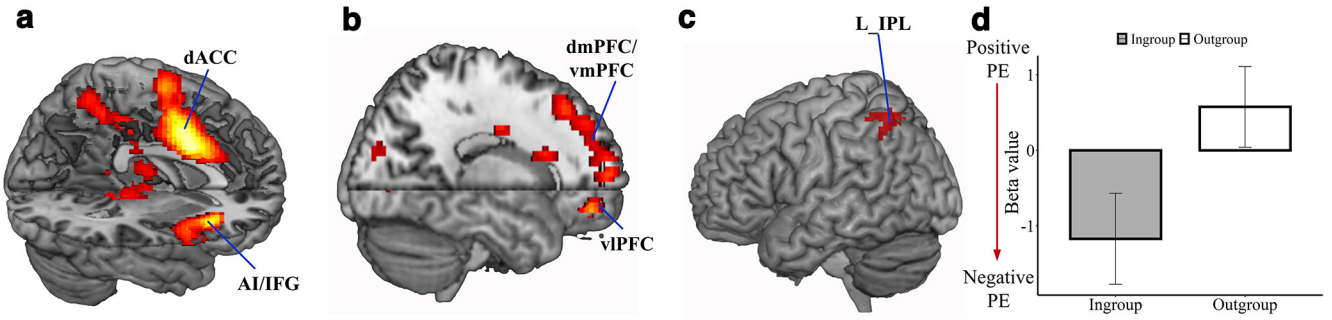

Figure 7. Neural activation related to group-independent and group-dependent processing of model-derived prediction errors. ***a***, Negative prediction errors correlated with activity in dACC and AI independent of group. ***b***, Positive prediction errors correlated with activity in dmPFC, vmPFC, lateral PFC (lPFC), and striatum (caudate) independent of group. ***c***, ***d***, Left IPL activity was more strongly associated with negative prediction errors for the ingroup compared with the outgroup. Significant clusters were identified by combining a voxel-level threshold of $p < 0.001$ (uncorrected) and a cluster-level threshold of $p \leq 0.05$, FWE-corrected across the whole brain. Display threshold at $p_{uncorrected} < 0.001$. Error bars indicate SEM.

Table 6. Brain regions correlating with trial-wise prediction errors estimated from the winning learning models[a]

		MNI coordinates			
Region	Cluster size	*x*	*y*	*z*	Peak *z*
Negative prediction errors					
dACC	1774	6	11	32	6.04
R_AI/IFG	318	36	20	5	5.41
L_AI/IFG	497	−48	2	8	5.35
L_TPJ	269	−57	−28	20	5.17
R_precentral	125	45	−1	47	5.00
R_TPJ	309	60	−31	23	4.86
Thalamus	403	9	−25	−10	4.30
Positive prediction errors					
R_caudate	226	27	−19	32	4.96
dmPFC	140	−12	35	47	4.89
R_occipital	191	36	−88	8	4.78
R_lPFC	107	42	44	−13	4.73
L_fusiform	298	−24	−73	−7	4.53
R_fusiform	107	27	−82	−10	4.29
vmPFC	214	15	59	8	4.09

[a]IFG, Inferior frontal gyrus; L, left; R, right. Significant clusters were identified by combining a voxel-level threshold of $p < 0.001$ (uncorrected) and a cluster-level threshold of $p \leq 0.05$, FWE-corrected.

Discussion

We investigated how intermixed experiences with several ingroup and outgroup individuals dynamically change subjective closeness with, and impressions of, the ingroup and outgroup. Our results show that experiences with the ingroup can reduce ingroup favoritism (Fig. 3), and the reduction is stronger when initial identification with the ingroup is strong. It is noteworthy that participants had identical experiences with the ingroup and outgroup, and fully processed the outgroup feedback (Fig. 4), as indicated by comparable weights and discounting parameters for ingroup and outgroup prediction errors.

Importantly, the reduction in ingroup favoritism, reflected by intergroup impressions change, was predicted by the weights on ingroup prediction errors ($W_{ingroup}$, Fig. 6), that is, the weight that individuals assign to ingroup experiences, which, in turn changed ingroup closeness. The positive association between $W_{ingroup}$ and intergroup impression change holds even after individual differences in initial expectancy ratings were taken into account. This indicates that participants changed intergroup impressions based on the importance they assigned to ingroup experiences, independently of the initial expectancy ratings.

Specifying these findings, we show that negative ingroup experiences are particularly relevant for changing intergroup impressions. Although our participants initially had low expectations with regard to the ingroup and the outgroup, and, in the course of the study, mainly made positive experiences with both groups, it was the weight given to negative ingroup prediction errors that drove the overall impression change. Previous studies showed that people were more influenced by negative information than positive information (e.g., Reeder and Brewer, 1979; Fiske, 1980; Baumeister et al., 2001). Our findings extend this conclusion to an intergroup setting and suggest that the negative ingroup experiences may serve as an efficient strategy to reduce ingroup favoritism.

The finding that ingroup experiences outweighed outgroup experiences with regard to impression formation supports ingroup- rather than outgroup-focused theories of group impression formation. In line with previous evidence showing an ingroup bias in information processing (Howard and Rothbart, 1980; Foddy et al., 2009; Hughes et al., 2017), our results indicate that learning from ingroup individuals can be more relevant for the formation of intergroup impressions compared with learning from outgroup individuals. Going beyond the existing evidence on ingroup biases in information processing (Howard and Rothbart, 1980; Foddy et al., 2009; Hughes et al., 2017), we reveal a mechanism through which ingroup experiences can affect intergroup impression formation. We demonstrate that intergroup impressions are shaped by learning-related changes in ingroup associations (closeness), caused by errors in the predictions of ingroup behavior. More generally, our findings show that models of impression formation can be enriched by learning processes originally established in the individual, nonsocial domain (Rescorla and Wagner, 1972), extending recent studies that investigated impression formation regardless of intergroup processes (Hackel et al., 2015; Ligneul et al., 2016; Mende-Siedlecki, 2018).

Our participants showed a reduction in ingroup favoritism in impression ratings, although they entered the experiment with comparable expectations toward the ingroup and the outgroup. Given the ingroup favoritism in impression ratings before the experiment, the comparable expectancy ratings may seem surprising. However, being all naive, the participants of our study could not know from prior experience whether anybody (ingroup or outgroup member) would give up money to save them from shocks and thus entered the experiment with relatively low expectations for help in both conditions. Supporting this assumption, there was no significant relationship between the individual ingroup bias in impression ratings and the intergroup differences in the

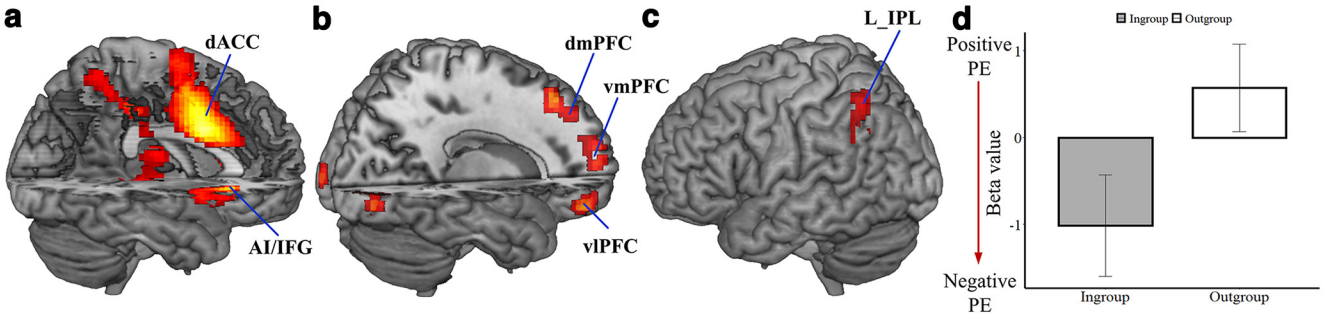

Figure 8. Neural activation related to group-independent and group-dependent processing of model-independent prediction errors. Prediction errors were defined by the difference between the outcome and participants' expectancy ratings. ***a***, Negative prediction errors correlated with activity in dACC and AI independent of group. ***b***, Positive prediction errors correlated with activity in dmPFC, vmPFC, IPFC, and striatum (caudate) independent of group. ***c***, ***d***, Left IPL activity was more strongly associated with negative prediction errors in the ingroup compared with the outgroup condition. Display threshold at $p < 0.001$, cluster size > 60 voxels.

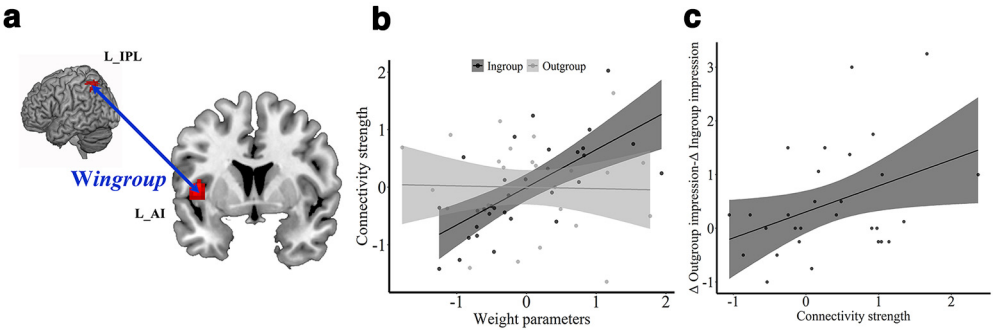

Figure 9. Left IPL-left AI connectivity influenced by weights given to ingroup prediction errors and related to impression change. ***a***, The $W_{ingroup}$ parameter modulated the coupling between left IPL and left AI during ingroup feedback revelation. Significant clusters were identified by SVC-FWE in the entire neural network that was involved in the processing of negative prediction errors shown in Figure 7*a*. ***b***, The association between weight parameters and left IPL to left AI coupling was significantly stronger when receiving feedback from the ingroup compared with the outgroup. ***c***, The connectivity strength between left IPL and left AI when receiving ingroup feedback correlated with the decline in ingroup favoritism in impression ratings. Display threshold at $p < 0.001$. Error bars indicate SEM.

initial expectancy ratings ($r = 0.063$, $p = 0.742$, Bayes factor = 0.36), suggesting that the initial expectancy ratings did not reflect the individual ingroup bias in other dimensions, such as impression.

On the neural level, the only region that showed a significant group difference (ingroup bias) in the processing of learning signals was the left IPL (Fig. 7*c*). The IPL has been associated with the computation of "saliency maps" (Molenberghs et al., 2007; Vandenberghe and Gillebert, 2009; Kahnt et al., 2014), giving rise to importance-weighted internal representations of the external world (Treue, 2003). Moreover, there is evidence that the activity of the IPL is sensitive to group membership. For example, left IPL activity was stronger when participants perceived actions performed by ingroup compared with outgroup members, and this activation difference was further related to the individual ingroup bias in the judgment of action (Molenberghs et al., 2013). Extending previous findings, our results indicate that the IPL is also modulated by group membership in intergroup impression formation processes indexed by learning signals. Given its role in the computation of "saliency maps," it is conceivable that the IPL increases the salience (i.e., the weight) of ingroup learning signals compared with outgroup learning signals. As such, intergroup impressions are more strongly formed by ingroup experiences, although participants learned from the ingroup and the outgroup feedback indicating that they paid attention to information from both social groups.

The individual extent of intergroup impression change was predicted by the individual extent of functional coupling between the IPL and the AI. The more strongly a person weighted the ingroup prediction errors, the stronger the coupling between the AI and the IPL, and the stronger the reduction in the ingroup bias with regard to impression ratings. The observed functional coupling between the IPL and the AI is in line with previous findings of close structural connections between the AI and the parietal, as well frontal, occipital, and limbic regions (Ghaziri et al., 2017). In particular, the IPL and AI are coactivated during motivationally consistent judgments (Leong et al., 2019) and learning of negative information generated by the ingroup (Hughes et al., 2017). Other studies highlighted the role of the AI in group categorization (Lieberman et al., 2005; Cikara et al., 2017), and in intergroup impression processing (Beer et al., 2008; Hein et al., 2016a; Hughes et al., 2017). For example, the AI has been shown to predict implicit negative attitudes toward members of a different race (Beer et al., 2008), and to contribute to learning-related impression change in intergroup contexts (Hein et al., 2016a; Hughes et al., 2017). Together, the current results suggest that the role of the AI in forming affective impressions is not limited to contexts with outgroup learning (Hein et al., 2016a), but extends to, and is facilitated by, connectivity with IPL in contexts with ingroup learning.

In addition to its role in integrating social motivation and learning (Hein et al., 2016a), the AI is also a part of the "salience network," which is particularly involved in the detection of motivationally salient stimuli, and is frequently associated with social emotion (Sridharan et al., 2008; Craig, 2009; Cauda et al., 2012; Amodio, 2014). The IPL in turn computes "saliency maps"

(Molenberghs et al., 2007; Vandenberghe and Gillebert, 2009; Kahnt et al., 2014). Building on these findings, the present findings uncovered a connectivity between the more dorsal (IPL) and more ventral (AI) regions that might serve the selection of salient events (i.e., ingroup experiences) for enhanced processing of ingroup-specific information and subsequent adjustment of intergroup impressions.

In our study, participants had intermixed experiences with ingroup and outgroup individuals, and could therefore directly compare ingroup and outgroup experiences. In the context of direct group comparisons, group membership becomes more salient (Turner et al., 1987; Van Knippenberg and Wilke, 1988; Zhou et al., 2020). Because of the high salience of the ingroup category, participants may have perceived themselves as representatives of their ingroup, rather than as individuals (Paolini et al., 2010), assigning the ingroup greater relevance for their own social identity (Tajfel et al., 1971), which in turn may have made ingroup experiences more relevant than outgroup experiences.

Although our results emphasize the importance of ingroup experiences and ingroup associations for shaping intergroup impressions, they do not disregard the relevance of (positive) outgroup experiences. In line with previous studies (Hein et al., 2016a), our results highlight that experiences with outgroup individuals elicit learning-related signals (i.e., prediction errors) that change outgroup closeness. The outgroup-related learning processes had less impact on intergroup impressions than ingroup-related learning processes, but it is conceivable that they may have had a stronger effect in the absence of the more salient ingroup experiences. Apart from that, previous research suggests that outgroup members may be perceived as more homogeneous compared with ingroup members, and this outgroup homogeneity effect may influence multiple social cognitive processes, including impression formation (Song et al., 2018). It is thus plausible that the relevance of outgroup prediction errors on intergroup impression change could be modulated by the individual level of perceived outgroup homogeneity, a question that should be investigated in future studies.

In conclusion, our results demonstrated that intergroup impressions are mainly shaped by ingroup experiences, and revealed the underlying neural mechanisms. If their predictions about ingroup behavior are not confirmed by experience, people perceive outgroup individuals as positively as ingroup individuals. Based on this finding, portraying the ingroup in a more realistic light is a promising strategy to reduce ingroup favoritism, an insight that could have practical implications for the improvement of intergroup relations.

References

Amodio DM (2014) The neuroscience of prejudice and stereotyping. Nat Rev Neurosci 15:670–682.

Amodio DM, Cikara M (2021) The social neuroscience of prejudice. Annu Rev Psychol 72:439–469.

Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26:839–851.

Barlow FK, Paolini S, Pedersen A, Hornsey MJ, Radke HR, Harwood J, Rubin M, Sibley CG (2012) The contact caveat: negative contact predicts increased prejudice more than positive contact predicts reduced prejudice. Pers Soc Psychol Bull 38:1629–1643.

Bates D, Kliegl R, Vasishth S, Baayen H (2015) Parsimonious mixed models. arXiv 1506.04967.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. Rev Gen Psychol 5:323–370.

Beer JS, Stallen M, Lombardo MV, Gonsalkorale K, Cunningham WA, Sherman JW (2008) The Quadruple Process model approach to examining the neural underpinnings of prejudice. Neuroimage 43:775–783.

Bettencourt BA, Brewer MB, Croak MR, Miller N (1992) Cooperation and the reduction of intergroup bias: the role of reward structure and social orientation. J Exp Soc Psychol 28:301–319.

Biernat M, Vescio TK, Billings LS (1999) Black sheep and expectancy violation: integrating two models of social judgment. Eur J Soc Psychol 29:523–542.

Branscombe NR, Wann DL, Noel JG, Coleman J (1993) In-group or out-group extremity: importance of the threatened social identity. Pers Soc Psychol Bull 19:381–388.

Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. Proc Natl Acad Sci USA 107:14431–14436.

Cauda F, Costa T, Torta DM, Sacco K, D'Agata F, Duca S, Geminiani G, Fox PT, Vercelli A (2012) Meta-analytic clustering of the insular cortex: characterizing the meta-analytic connectivity of the insula when involved in active tasks. Neuroimage 62:343–355.

Chu D, Griffey D (1985) The contact theory of racial integration: the case of sport. Sociol Sport J 2:323–333.

Cikara M, Van Bavel JJ, Ingbretsen ZA, Lau T (2017) Decoding 'us' and 'them': neural representations of generalized group concepts. J Exp Psychol Gen 146:621–631.

Craig AD (2009) How do you feel–now? The anterior insula and human awareness. Nat Rev Neurosci 10:59–70.

Cutler J, Wittmann MK, Abdurahman A, Hargitai LD, Drew D, Husain M, Lockwood PL (2021) Ageing is associated with disrupted reinforcement learning whilst learning to help others is preserved. Nat Commun 12:1–13.

Dijksterhuis A, Van Knippenberg A (1998) The relation between perception and behavior, or how to win a game of trivial pursuit. J Pers Soc Psychol 74:865–877.

Dovidio JF, Gaertner SL, Kawakami K (2003) Intergroup contact: the past, present, and the future. Group Process Intergroup Relat 6:5–21.

Fiske ST (1980) Attention and weight in person perception: the impact of negative and extreme behavior. J Pers Soc Psychol 38:889–906.

Foddy M, Platow MJ, Yamagishi T (2009) Group-based trust in strangers: the role of stereotypes and expectations. Psychol Sci 20:419–422.

Friston K, Buechel C, Fink G, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Ghaziri J, Tucholka A, Girard G, Houde JC, Boucher O, Gilbert G, Descoteaux M, Lippé S, Rainville P, Nguyen DK (2017) The corticocortical structural connectivity of the human insula. Cereb Cortex 27:1216–1228.

Hackel LM, Amodio DM (2018) Computational neuroscience approaches to social cognition. Curr Opin Psychol 24:92–97.

Hackel LM, Doll BB, Amodio DM (2015) Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. Nat Neurosci. 18:1233–1235.

Hackel LM, Berg JJ, Lindström BR, Amodio DM (2019) Model-based and model-free social cognition: investigating the role of habit in social attitude formation and choice. Front Psychol 10:2592.

Hein G, Silani G, Preuschoff K, Batson CD, Singer T (2010) Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron 68:149–160.

Hein G, Engelmann JB, Vollberg MC, Tobler PN (2016a) How learning shapes the empathic brain. Proc Natl Acad Sci USA 113:80–85.

Hein G, Morishima Y, Leiberg S, Sul S, Fehr E (2016b) The brain's functional network architecture reveals human motives. Science 351:1074–1078.

Hein G, Engelmann JB, Tobler PN (2018) Pain relief provided by an outgroup member enhances analgesia. Proc R Soc Lond B Biol Sci 285:20180501.

Howard JW, Rothbart M (1980) Social categorization and memory for in-group and out-group behavior. J Pers Soc Psychol 38:301–310.

Hughes BL, Zaki J, Ambady N (2017) Motivation alters impression formation and related neural systems. Soc Cogn Affect Neurosci 12:49–60.

Kahnt T, Park SQ, Haynes JD, Tobler PN (2014) Disentangling neural representations of value and salience in the human brain. Proc Natl Acad Sci USA 111:5000–5005.

Kang P, Burke CJ, Tobler PN, Hein G (2021) Why we learn less from observing outgroups. J Neurosci 41:144–152.

Leach CW, Van Zomeren M, Zebel S, Vliek ML, Pennekamp SF, Doosje B, Ouwerkerk JW, Spears R (2008) Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. J Pers Soc Psychol 95:144–165.

Leong YC, Hughes BL, Wang Y, Zaki J (2019) Neurocomputational mechanisms underlying motivated seeing. Nat Hum Behav 3:962–973.
Lieberman MD, Hariri A, Jarcho JM, Eisenberger NI, Bookheimer SY (2005) An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. Nat Neurosci 8:720–722.
Ligneul R, Obeso I, Ruff CC, Dreher JC (2016) Dynamical representation of dominance relationships in the human rostromedial prefrontal cortex. Curr Biol 26:3107–3115.
Lindström B, Tobler PN (2018) Incidental ostracism emerges from simple learning mechanisms. Nat Hum Behav 2:405–414.
Lindström B, Selbing I, Molapour T, Olsson A (2014) Racial bias shapes social reinforcement learning. Psychol Sci 25:711–719.
Lockwood PL, Klein-Flügge MC (2021) Computational modelling of social cognition and behaviour: a reinforcement learning primer. Soc Cogn Affect Neurosci 16:761–771.
Malhotra D, Liyanage S (2005) Long-term effects of peace workshops in protracted conflicts. J Confl Resolut 49:908–924.
Marcus-Newhall A, Miller N, Holtz R, Brewer MB (1993) Cross-cutting category membership with role assignment: a means of reducing intergroup bias. Br J Soc Psychol 32:125–146.
Marques JM, Yzerbyt VY, Leyens JP (1988) The 'black sheep effect': extremity of judgments towards ingroup members as a function of group identification. Eur J Soc Psychol 18:1–16.
McConahay JB (1986) Modern racism, ambivalence, and the modern racism scale. In: Prejudice, discrimination, and racism (Dovidio JF, Gaertner SL, eds), pp 91–125. New York, NY: Academic Press.
McLaren DG, Ries ML, Xu G, Johnson SC (2012) A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage 61:1277–1286.
Mende-Siedlecki P (2018) Changing our minds: the neural bases of dynamic impression updating. Curr Opin Psychol 24:72–76.
Mendoza SA, Lane SP, Amodio DM (2014) For members only: ingroup punishment of fairness norm violations in the ultimatum game. Soc Psychol Pers Sci 5:662–670.
Molenberghs P, Halász V, Mattingley JB, Vanman EJ, Cunnington R (2013) Seeing is believing: neural mechanisms of action–perception are biased by team membership. Hum Brain Mapp 34:2055–2068.
Molenberghs P, Mesulam MM, Peeters R, Vandenberghe RR (2007) Remapping attentional priorities: differential contribution of superior parietal lobule and intraparietal sulcus. Cereb Cortex 17:2703–2712.
Morey RD, Rouder JN, Jamil T, Morey MR (2015) Package 'bayesfactor.' Available at https://cran.r-project.org/web/packages/BayesFactor/index.html.
Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ (2017) Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. PLoS Comput Biol 13:e1005684.
Paluck EL (2009) Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda. J Pers Soc Psychol 96:574–587.
Paolini S, Harwood J, Rubin M (2010) Negative intergroup contact makes group memberships salient: explaining why intergroup conflict endures. Pers Soc Psychol Bull 36:1723–1738.
Park B, Fareri D, Delgado M, Young L (2021) The role of right temporoparietal junction in processing social prediction error across relationship contexts. Soc Cogn Affect Neurosci 16:772–781.
Pettigrew TF, Tropp LR (2006) A meta-analytic test of intergroup contact theory. J Pers Soc Psychol 90:751–783.
Reeder GD, Brewer MB (1979) A schematic model of dispositional attribution in interpersonal perception. Psychol Rev 86:61–79.
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning: II. Current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.
Rutledge RB, Skandali N, Dayan P, Dolan RJ (2014) A computational and neural model of momentary subjective well-being. Proc Natl Acad Sci USA 111:12252–12257.
Schlueter E, Scheepers P (2010) The relationship between outgroup size and anti-outgroup attitudes: a theoretical synthesis and empirical test of group threat-and intergroup contact theory. Soc Sci Res 39:285–295.
Seth AK (2010) A MATLAB toolbox for Granger causal connectivity analysis. J Neurosci Methods 186:262–273.
Siegel JZ, Mathys C, Rutledge RB, Crockett MJ (2018) Beliefs about bad people are volatile. Nat Hum Behav 2:750–756.
Song H, Schuldt JP, McLeod PL, Crain RL, Dickinson JL (2018) Group norm violations in an online environmental social network: effects on impression formation and intergroup judgments. Group Process Intergroup Relat 21:422–437.
Sridharan D, Levitin DJ, Menon V (2008) A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc Natl Acad Sci USA 105:12569–12574.
Tajfel H, Turner JC (1979) An integrative theory of intergroup conflict. In: The social psychology of intergroup relations (Austin WG, Worchel S, eds), pp 33–47. Monterey, CA: Brooks/Cole.
Tajfel H, Billig MG, Bundy RP, Flament C (1971) Social categorization and intergroup behaviour. Eur J Soc Psychol 1:149–178.
Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. J Neurophysiol 95:301–310.
Treue S (2003) Visual attention: the where, what, how and why of saliency. Curr Opin Neurol 13:428–432.
Turner JC, Hogg MA, Oakes PJ, Reicher SD, Wetherell MS (1987) Rediscovering the social group: a self-categorization theory. Oxford: Basil Blackwell.
van Baar JM, Chang LJ, Sanfey AG (2019) The computational and neural substrates of moral strategies in social decision-making. Nat Commun 10:1–14.
Van Knippenberg A, Wilke H (1988) Social categorization and attitude change. Eur J Soc Psychol 18:395–406.
Vandenberghe R, Gillebert CR (2009) Parcellation of parietal cortex: convergence between lesion-symptom mapping and mapping of the intact functioning brain. Behav Brain Res 199:171–182.
Will GJ, Rutledge RB, Moutoussis M, Dolan RJ (2017) Neural and computational processes underlying dynamic changes in self-esteem. Elife 6:e28098.
Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C (2020) Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Soc Cogn Affect Neurosci 15:695–707.
Zhou Y, Gao T, Zhang T, Li W, Wu T, Han X, Han S (2020) Neural dynamics of racial categorization predicts racial bias in face recognition and altruism. Nat Hum Behav 4:69–87.