The social transmission of empathy relies on observational reinforcement learning

Yuqing Zhou[1 2*], Shihui Han[3], Pyungwon Kang[4], Philippe N. Tobler[4] & Grit Hein[2*]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences.
[2] Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg.

[3] School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, PKU-IDG/McGovern Institute for Brain Research, Peking University.

[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.

Address correspondence to:

Prof. Dr. Grit Hein,

Translational Social Neuroscience Lab, University Hospital of Würzburg

Margarete-Höppel-Platz 1, 97080 Würzburg, Germany

Email: Hein_G@ukw.de

Dr. Yuqing Zhou,

CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing 100101, China

Email: zhouyq@psych.ac.cn

Abstract

Theories of moral development propose that empathy is transmitted across individuals, yet the mechanism through which empathy is socially transmitted remains unclear. We conducted three studies to investigate whether, and if so, how observing empathic responses in others affects the empathy of the observer. Our results show that observing empathic or non-empathic responses generates learning signals that respectively increases or decreases empathy ratings of the observer and alters empathy-related responses in the anterior insula (AI), i.e., the same region that correlated with empathy baseline ratings, as well as its functional connectivity with the temporal-parietal junction (TPJ). Together, our findings provide a neurocomputational mechanism for the social transmission of empathy that accounts for changes in individual empathic responses in empathic and non-empathic social environments.

Teaser:

Observing empathic and non-empathic reactions elicits learning that changes the subjective and neural empathy of the observer.

Introduction

Empathy – the ability to share the feelings and thoughts of others – can spread across individuals (*1*). Supporting this notion, there is evidence that self-reported empathy increases if empathy is highly valued by others (*2*, *3*), and when watching empathic responses of others (*4*). However, these proofs of principle were unable to elucidate the mechanisms through which empathy is socially transmitted.

An influential but untested theory suggests that the social transmission of empathy is based on a learning process that is triggered by observing the empathic reactions of others ("empathic conditioning"(*5*). According to observational learning theory (*6*), individuals learn from the differences between empathic responses they observe in others and the empathic response they expected to see in others. The mismatch between the observed and expected responses generates so called observational prediction errors that are known to drive learning-related changes in the actions of the observer (*7–9*). Here, we investigate whether humans can learn to increase or decrease empathy by observing that others show more or less empathy than predicted.

Neurally, observational learning signals have been associated with activation of the mirror neuron system, including the dorsolateral prefrontal cortex (dlPFC), and premotor cortex, as well as the mentalizing network, including the temporal parietal junction (TPJ), dorsal medial prefrontal cortex (dmPFC), and anterior temporal lobe (ATL) (*7–11*). Using brain stimulation and computational modeling, a recent study

suggested that disruption of the left TPJ weakens participants' choice adjustment when confronted with dissenting information from others (*12*). A possible interpretation of this finding is that reduced TPJ activation results in reduced social influence on learning.

So far, learning based on observational prediction errors has been associated with the social transmission of fear (*13–15*), value-based decision making (*7*, *8*, *16*) and the propensity to take or avoid risks (*9*). However, it remained unclear whether, and if so, how observing empathic reactions to the pain of others affects learning of empathic responses in the observer.

To address this question, we developed an observational-learning-of-empathy paradigm, which we combined with functional magnetic resonance imaging (fMRI) and computational modelling (Study 1). The behavioral results of Study 1 were substantiated by the results of a behavioral control study (Study 2) and replicated in an independent behavioral study (Study 3).

All studies consisted of three parts: a baseline session in which we assessed participants' empathy ratings independently of any experimental manipulation, an observational empathy learning session, and a generalization session that aimed to test whether potential learning-related changes in empathy ratings generalize to individuals that were not part of the learning session (**Figure 1A**). In the baseline and the generalization session, participants rated their empathy when observing videos showing painful or non-painful stimulation in others (**Figure 1B**). In the observational

empathy learning session, participants witnessed the reactions of a demonstrator to the pain of a recipient and were randomly assigned to two groups, a high and a low empathy group. In the high empathy group, participants observed strong empathic reactions whereas in the low empathy group, participants observed weak empathic reactions to the same pain inflicted on the recipient. In the high empathy group, the demonstrator's ratings of the recipient's pain were consistently higher than the participant's baseline ratings, indicating a stronger empathic reaction than the participant's empathy baseline. In the low empathy group, the demonstrators' ratings of the recipient's pain were consistently lower than the participant's baseline ratings, indicating a weaker empathic reaction compared to the participant (see Methods for details). In two of the studies, the observed ratings reflected the reactions of a human demonstrator (Studies 1 and 3), whereas in a third, control study, the observed ratings were from a computer (Study 2). After observing high or low empathic reactions, participants rated how they themselves felt when watching pain in the recipient (**Figure 1C**).

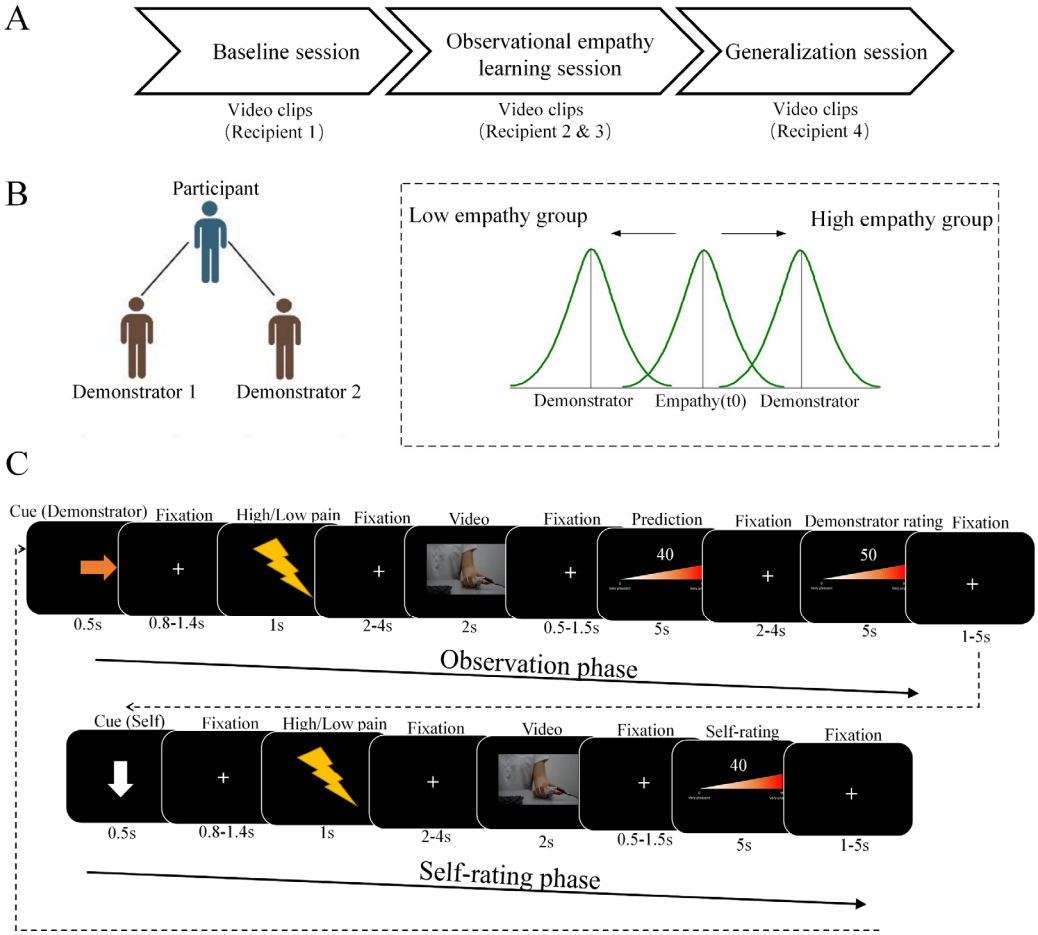

Figure 1. Experimental setup. (A) The main experiment consisted of the baseline session, observational empathy learning session and the generalization session. In each session, participants viewed video clips of different recipients receiving electrical stimulation. The mean pain intensity ratings were comparable across recipients indicated by a pilot stimulus validation study (Figure S2). (B) During the observational learning session, participants observed ratings of two demonstrators (Study 1 and Study 3: human demonstrators; Study 2: computer demonstrators). The ratings of these demonstrators were generated by a pre-defined algorithm, based on the participant empathy ratings in the baseline session (Empathy(t0)) as well as the experimental group the participant was assigned to (i.e., high empathy or low empathy group). **(C)** Example trial of the observational-learning-of-empathy task. Each trial started with an observation phase, followed by a self-rating phase. In the observation phase, the participants observed the ratings of another person (demonstrator) who watched and reacted to the painful stimulation inflicted on a recipient. The observation phase started with an arrow, followed by a lightning bolt cue that indicated the intensity of the recipient's pain (bright color indicating painful, dark color indicating non-painful stimulation) and a video showing the recipient receiving the respective stimulation. Participants were asked to predict the ratings of the demonstrator for this specific video on a scale from zero (predicting that the demonstrator would feel nothing when seeing the other in pain) to hundred (predicting

that the demonstrator would feel extremely bad when seeing the other in pain). At the end of the observation phase, the actual rating of the demonstrator was shown. The self-rating phase started with an arrow pointing to the participant. Next, they viewed the cue indicating the intensity of the recipient's stimulation, watched the video showing the stimulation of the same recipient as in the observation phase, and rated how they felt after seeing the stimulation of the recipient (from zero – not feeling anything, to hundred – feeling extremely bad). The trial structure of the control study (Study 2) was identical, except that we presented computer-generated ratings in the observation phase.

We hypothesized that observing others would increase the observer's empathy (as measured by ratings) in the high empathy group and decrease it in the low empathy group. The change in empathy ratings should be driven by learning signals, specifically, observational prediction errors, referring to the discrepancy between the predicted and observed empathy ratings. If the change in empathy is specific to the observational learning from another human, the learning-related changes in empathy ratings should be stronger after observing empathy ratings generated by human demonstrators (Studies 1 and 3) compared to computer-generated ratings (Study 2).

On the neural level, learning others' empathy responses might be related to activations in the brain regions that were shown to be involved in observational learning and the processing of social influence, including the dlPFC and dmPFC, the premotor cortex, (*7*, *8*, *10*, *11*), and the TPJ (*12*, *16*). Inspired by previous evidence showing that learning from own experiences about others changes empathy-related responses in the anterior insula (AI) (*17*), we further hypothesized that the learning-related changes in empathy may alter the interaction between regions encoding the observational learning signals and the AI.

Results

1. Manipulation checks across studies

One would expect that participants' emotion ratings when observing the pain of others would relate to trait empathy. To test this, we used a regression model with the average ratings in the baseline rating session for painful videos of all three studies as dependent variable and participants score on the Interpersonal Reactivity Index (IRI, (*18*)), study (studies 1, 2, and 3), and study x IRI score as predictors. This analysis revealed a significant effect of IRI score (β = 0.23, t = 2.87, p = 0.005), but no significant effects of study (β =*-0.03,* t = -0.34, p = 0.74) and study x IRI score (β = 0.08, t = 0.87, p = 0.38), indicating that across studies, the emotion ratings elicited by watching the painful stimulation of recipients were similarly related to trait empathy.

As a second manipulation check, we assessed the expectation that observing high and low empathic responses should differently change participants' impressions of the demonstrator. To test this, we used the pre- and post-learning impression scores (*17*, *19*) of Studies 1 and 3 (i.e., the studies including human demonstrators) as dependent variable and study (study1, study3), group (high, low empathy) and time (pre-, post-learning experiment) as predictors. We found a significant group x time interaction (β = 0.84,t = 2.42, p = 0.02), which occurred similarly for Study 1 and Study 3 (β = -0.08,t = -0.16, p = 0.87). While participants' impressions towards the demonstrators did not differ between the high and low empathy groups before the experiment (β = 0.07,t = 0.36, p = 0.72), their impression ratings towards the demonstrators were

more positive in the high compared to the low empathy group after the experiment (β = 0.91,t = 3.20, p = 0.002). This demonstrates that our experimental manipulation (observing empathic vs non-empathic responses) had an influence on how participants perceived the demonstrators, thus, validating the social manipulation.

2. Results of the fMRI study

2.1 Regression (model-independent) analyses of behavior

In the observation phase, participants predicted the empathy ratings of the demonstrator. Entering these prediction ratings as dependent variable in a linear mixed model (LMM) with group (high empathy, low empathy), trial number and group × trial number as predictors revealed a significant group × trial number interaction (χ2(1) = 26.04, p < 0.001), indicating that participants expected increasing empathy ratings of the demonstrators in the high (χ2(1) = 3.88, p = 0.05) and decreasing empathy ratings in the low empathy (χ2(1) = 27.58, p < 0.001) group (**Figure 2A**).

Next, we analyzed participants' own empathy ratings from the self-rating phase. An LMM with group (high empathy, low empathy), trial number and group × trial number as predictors revealed a significant group × trial number interaction (χ2(1) = 39.57, p < 0.001), indicating an increase in participants' empathy ratings in the high (χ2(1) = 5.44, p = 0.02) and a decrease in empathy ratings in the low empathy group (χ2(1) = 41.60, p < 0.001). Similarly, a LMM with group (high empathy, low empathy), session (baseline, observational empathy learning (1-4) and generalization,

197 coded as 0-5 respectively), and group × session as predictors and the average of
198 participants' empathy ratings in the respective session as the dependent variable
199 showed a significant group × session interaction ($\chi 2(1) = 116.5$, $p < 0.001$, **Figure**
200 **2B)**. There was a trend towards a difference in baseline empathy ratings between the
201 low and high empathy group ($t(50) = 1.87$, $p = 0.067$). Separate analyses then showed
202 a significant increase in empathy ratings across sessions in the high empathy group
203 ($\chi 2(1) = 101.3$, $p < 0.001$, **Figure 2B**), and a significant decrease in empathy ratings
204 across sessions in the low empathy group ($\chi 2(1) = 40.40$, $p < 0.001$, **Figure 2B**).

205 The observed changes in participants' empathy ratings might be driven by social
206 desirability and the wish to conform with the ratings of the demonstrators, and
207 influenced by empathy baseline ratings. To evaluate the influence of social
208 desirability and conformity on the change in empathy ratings during learning, we
209 calculated the individual scores measured from social desirability (SDS-17; (*20*)) and
210 conformity scales (*21*) for the high and low empathy group separately. We then
211 conducted a regression analysis with the change in empathy ratings between baseline
212 and generalization sessions as dependent variables, and the individual scores on social
213 desirability and conformity scales as predictors. We also included the averaged
214 baseline empathy ratings as predictor to check whether individual differences in
215 empathy baseline ratings account for group differences in subsequent empathy
216 changes. The analyses revealed no significant effects (**Table S1**, $ps > 0.31$), rendering
217 the possibility unlikely that the individual changes in empathy ratings were driven by

individual differences in social desirability and conformity or by baseline differences in empathy. Together, these results show that participants shifted their empathy ratings towards the ratings of the demonstrators, that these changes could not be explained by social desirability and that they were preserved even when participants were no longer presented the demonstrators' ratings (i.e., generalization session).

Participants also reported how much pain they thought the person in the video clip was experiencing and how much time they were willing to spend in order to help the pain recipient before and after the experiment. Consistent with the change in empathy ratings, participants in the high empathy group evaluated the intensity of the pain experienced by the recipient as significantly stronger (M = 7.1 vs. 6.4, $t(25)$ = 2.71 , p = 0.01, **Figure 2C**) and were willing to spend more time to help the recipient after learning (M = 26.1 min vs. 20.8 min, $t(25)$ = 4.16, $p < 0.001$, **Figure 2C**) compared to before learning. In contrast, there were no such learning-related changes in the low empathy group (pain intensity: M = 6.8 vs. 6.7, $t(25)$ = 0.46, p = 0.65; prosocial tendency: M = 26.5 min vs. 24.7 min, $t(25)$ = 0.7, p = 0.49, **Figure 2C**). Finally, in both groups, the difference in empathy ratings between the baseline and the generalization session predicted the individual pre-to-post difference in the willingness to spend time in order to help the recipient (high and low empathy group combined: *rho* = 0.345, p = 0.012; high empathy group only, *rho* = 0.407, p = 0.039; low empathy group only, *rho* = 0.392, p = 0.048, **Figure 2D**). Together, these results suggest that observing the empathic reactions of the demonstrators changed the

predictions and empathy ratings of the observer. Moreover, changes in empathy ratings influenced participants' willingness to invest time in order to help the recipient.

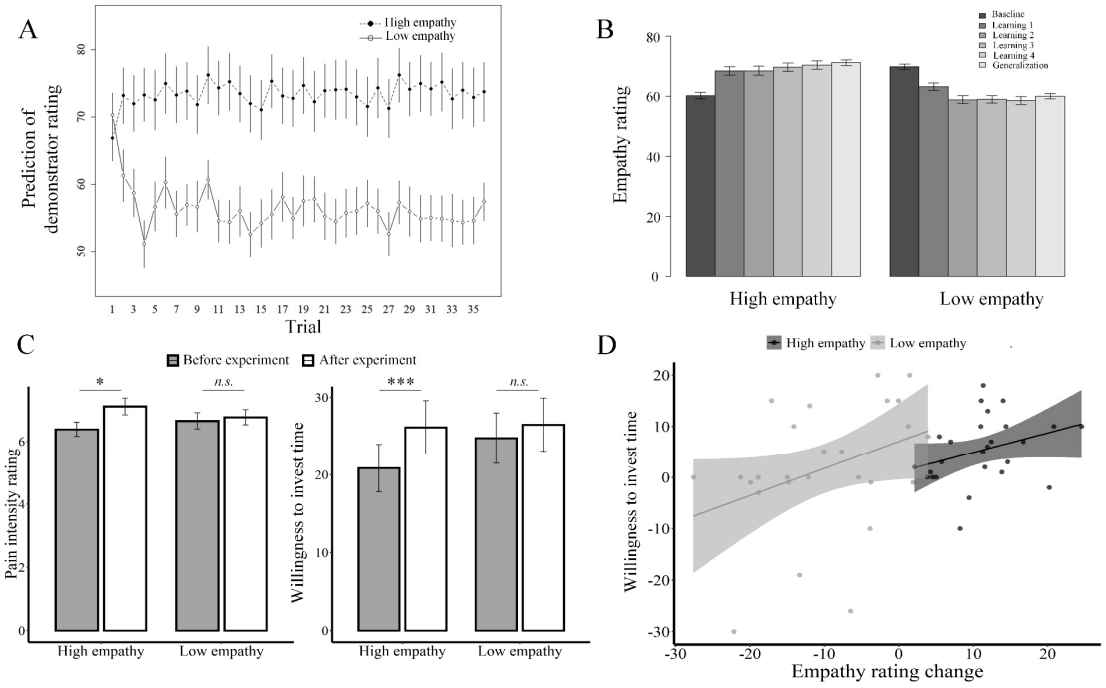

Figure 2. Observation-induced changes in predictions, empathy ratings, and willingness to spend time to help after learning. (A) Predictions of demonstrator ratings in the observation phase (mean across participants) increased in high empathy (black dots) but decreased in low empathy (white dots) groups across trials. (B) Averaged empathy ratings in each session of the experiment show an increase in the high empathy group, and a decrease in the low empathy group. (C) Average pain intensity ratings and willingness to spend time to help the recipient increased in the high empathy group after learning. (D) In both groups, the change in empathy rating from the baseline to the generalization session was related to participants' willingness to spend time to help the recipient.

2.2 Reinforcement learning model-based analyses of behavior

Having demonstrated that participants changed their empathy ratings after observing the ratings of others, we next sought to examine the computational mechanisms

supporting these changes. We first modelled participants' trial-by-trial predictions of demonstrator ratings using a Rescorla-Wagner reinforcement-learning model (*22*). The model fitted the data adequately for both the high and low empathy groups (r^2 (mean ± SD) = 0.22 ± 0.21 and 0.27 ± 0.21; **Figure 3A and 3B**, see Materials and methods for details). The estimated learning rate was comparable for high and low empathy groups (α: $t(50) = -0.179$, $p = 0.859$, 95% CI = [-0.08, 0.07], **Table S2**), suggesting that participants learned to predict the ratings of empathic and non-empathic demonstrators similarly well.

Next, we modeled trial-by-trial update of participants' empathy ratings as a linear function of the cumulative impact of observational prediction errors (*19*, *23*, *24*), as estimated by the reinforcement learning model. Bayesian model selection was used to identify the model that was most probable to generate the data, based on Laplace approximation (see Materials and methods for details). The winning model (Equation 5, Model 3, XP: 1) successfully captured dynamic changes of empathy ratings at the individual level for both high ($r^2 = 0.19 \pm 0.14$) and low ($r^2 = 0.24 \pm 0.12$) empathy groups (**Figure 3D**). The winning model (Equation 5, Model 3) assumes that the empathy ratings of participants for each trial *t* are driven by the time-discounted sum of previous observational prediction error. It considers the empathy ratings in the first half and second half of the learning session separately and adds up separately modeled positive and negative observational prediction errors.

In this model, two parameters (*W1* and *W2*) capture the magnitude (weight) of the influence of observational prediction errors on changes in participants' empathy ratings in the first and second half of the observational learning session. The weights are separated by the sign of the observational prediction errors (indicated by the pos/neg subscript). A larger *W* corresponds to a stronger influence of observational prediction errors on participants' empathy ratings. The discount parameter γ ($0 \leq \gamma \leq 1$) captures an exponentially decaying influence of previous observational prediction errors over time, such that more recent observational prediction errors have a greater impact on the changes in empathy ratings than earlier observational prediction errors. If γ is close to one, all preceding observational prediction errors receive the same weight, and if it is close to zero, only the last observational prediction error leads to subsequent changes in empathy ratings.

We fitted the empathy ratings separately for the high and the low empathy group. For the high empathy group, the weight parameters on positive observational prediction errors were significantly larger than zero ($W1_{pos}$: $t(25) = 6.79$, $p < 0.001$; $W2_{pos}$: $t(25) = 6.12$, $p < 0.001$, **Figure 3E**, **Table S2**), whereas the weight parameters on negative observational prediction errors were not different from zero (*ts* > -1.94, *ps* > 0.06, **Figure 3E , Table S2**). By contrast, the weight parameters on negative observational prediction errors were significantly larger than zero in the low empathy group ($W1_{neg}$: $t(25) = 3.23$, $p = 0.003$; $W2_{neg}$: $t(25) = 4.34$, $p < 0.001$, **Figure 3F,Table S2**), whereas the weight parameters on positive observational prediction

errors were not different from zero (*ts* > -1.87, *ps* > 0.07, **Figure 3F ,Table S2**). These results suggest that participants in the high and low empathy groups were predominantly influenced by the positive and negative observational prediction errors, respectively. We also checked the relationships between the weight parameters and the individual scores on social desirability and conformity scales. The analyses revealed no significant effects (*ps* > 0.16), providing little support for the notion that individual weights on observational prediction errors were influenced by individual differences in social desirability and conformity.

Next, we correlated the weight parameters with the change in empathy ratings across participants. The respective weight parameters in the first half of the learning session (i.e., $W1_{pos}$ for the high empathy group and $W1_{neg}$ for the low empathy group) were significantly associated with the increase in empathy rating in the high empathy group (*rho* = 0.39, *p* = 0.047, **Figure 3G**) and the decrease in empathy rating in the low empathy group (*rho* = -0.50, *p* = 0.009, **Figure 3H**), whereas the weight parameters in the second learning session were not (*ps* > 0.154). These results suggest that the weight of the observational prediction errors in the first half of the learning experiment majorly drives the overall changes in empathy ratings.

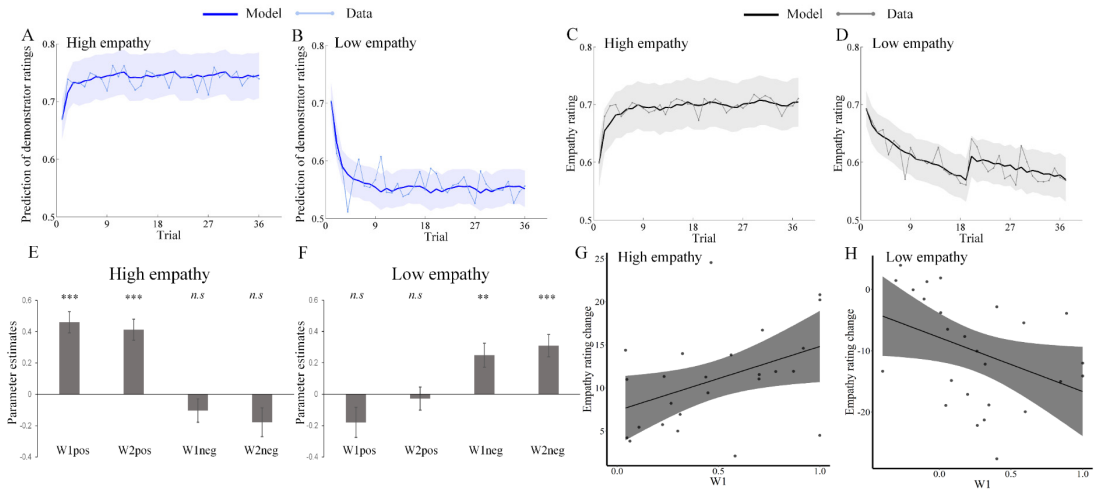

Figure 3. Computational models explain predictions and changes in empathy ratings. (A and B) Predictions of human demonstrator ratings (light blue line, shaded area represents the ±1 standard error) increased in the high empathy and decreased in the low empathy group, and our learning model explained these changes (dark blue line). (C and D) Trial-by-trial changes of empathy ratings (light grey, shaded area represents the ±1 standard error) and corresponding model estimates (dark grey) for the high and the low empathy groups. The model estimates illustrate the best-fitting model. (E and F) The value of the weight parameters for high (E) and low (F) empathy groups. The weight parameters for positive (negative) observational prediction errors were significantly larger than zero in high (low) empathy groups. (G and H) The weight parameters in the first half of the learning experiment (i.e., *W1*) significantly correlated with the change in empathy ratings across participants for both high and low empathy groups.

2.3 Neuroimaging results

Our fMRI data analyses focused on the neural mechanisms underlying the observational learning of empathy. As a manipulation check, we first examined the neural signals that significantly correlated with the trial-by-trial empathy rating when viewing the videos in the baseline session (i.e., before learning). Whole-brain analyses across all participants revealed activations in the dorsal medial cingulate (dMCC), bilateral anterior insula (AI), and the bilateral temporal parietal junction

(TPJ, **Figure 4A**), replicating the results of previous neuroimaging studies on the neural basis of empathy for pain (*25–30*).

Then, we investigated brain regions encoding observational prediction errors as identified by our computational model. Specifically, we regressed trial-by-trial observational prediction errors from the winning model as parametric modulator against neural activity when the ratings of demonstrators were revealed. In the high empathy group, the trial-wise observational prediction errors were related to activation in the dorsal medial prefrontal cortex (dmPFC) (**Figure 4B**; **Table 1A**), such that dmPFC activation was stronger when demonstrator ratings were higher than expected. In the low empathy group, we only observed significant neural responses with the inverse observational prediction errors, reflected by an increase of activation in the bilateral premotor cortex, the medial cingulate cortex (MCC), and the anterior insula (AI) when demonstrator ratings were lower than expected (**Figure 4C; Table 1B**).

Next, we compared the neural coding of observational prediction errors between the high and low empathy groups. The results revealed significant group differences in the TPJ, the MCC, premotor cortex, occipital cortex, and anterior insula (extending into the anterior temporal pole) (**Figure 4D, Table 1C**). These regions showed stronger activations when demonstrator ratings were higher than expected in the high empathy group ($ts > 2.78$, $ps < 0.010$). In contrast, in the low empathy group,

activations in these regions were stronger when demonstrator ratings were lower than expected ($ts < -3.56$, $ps < 0.002$).

Based on our modelling results, we further tested whether neural regions which differentially encoded observational prediction errors in the observation phase also showed group-related differential connectivities with regions encoding empathy-related activity in the self-rating phase.The strength of this functional coupling should depend on the individual weight given to the observational prediction errors (i.e., the *W1* parameter that accounted for the learning-related changes in empathy ratings). Given that the observational learning network contained several regions, we conducted a multi-region PPI analysis (*31*, *32*), which allows defining multiple seed regions and simultaneously assessing the respective connectivity changes depending on a given variable (here *W1*). We defined the seed regions by the brain regions that showed the strongest differential coding of observation prediction errors between groups (**Table 1C**). We calculated the connectivity strength between each of these seeds and 264 target regions that were defined with an established template (*33*), and assessed which of these connectivities was modulated by the *W1* parameter. Visualization of the suprathreshold edges revealed that the left TPJ showed the largest number of connectivities that were influenced by the magnitude of *W1*. This result held when we used different threshold values (ranging from 0.001 to 0.05) to identify significant connectivities, indicating the robustness of our results (**Figure S1**).

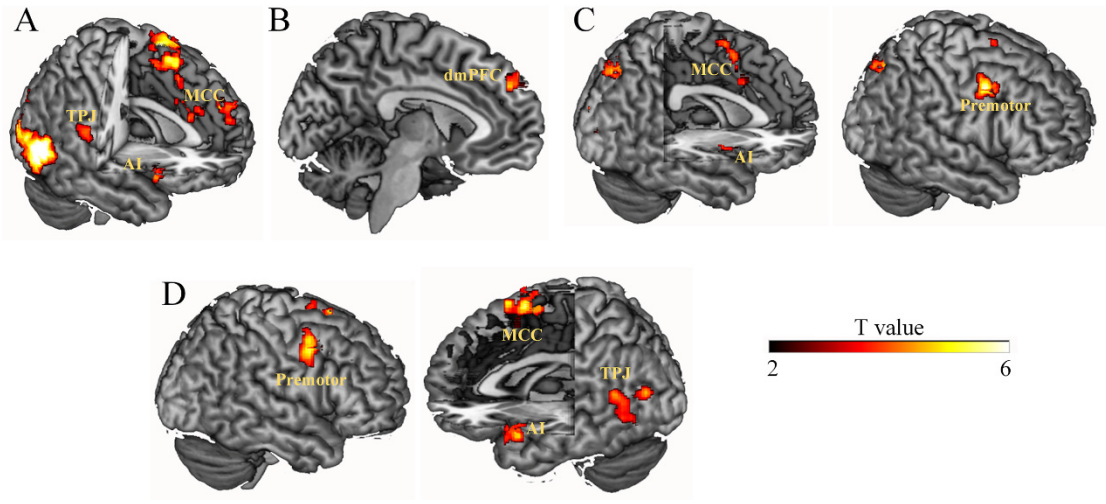

Figure 4. Neuroimaging results. (A) Neural responses associated with trial-by-trial empathy ratings in the baseline session. (B) Neural representation of observational prediction errors in the high empathy group. (C) Neural representation of observational prediction errors in the low empathy group. (D) Regions encoding observational prediction errors differently between high and low empathy groups. Significant clusters were identified by combining a voxel-level threshold of $p < .001$ (uncorrected) and a cluster-level threshold of $p < 0.05$, *FWE corrected* across the whole brain. Display threshold at $p_{uncorrected} < 0.001$; AI = anterior insula, MCC = mid cingulate cortex, TPJ = temporal parietal junction; dmPFC = dorsal medial prefrontal cortex; premotor = premotor cortex; SMA = supplementary motor area.

Table 1: Brain regions correlating with trial-by-trial observational prediction errors for the high and the low empathy group separately and across both groups.

Region	Cluster Size	MNI Coordinates			Peak z
		X	Y	Z	
A) High empathy group					
dmPFC	80	-6	48	34	3.90
B) Low empathy group					
R_premotor	122	52	4	48	4.93
L_fusiform	173	-44	-66	-24	4.68
Cerebellum	125	-8	-60	-12	4.68
L_premotor	82	-42	6	46	4.66
Precuneus	672	22	-74	44	4.50

Cerebellum	120	0	-70	-32	4.43
R_Lingual	86	10	-64	-12	4.41
dMCC/SMA	310	6	2	64	4.37
R_AI	82	46	2	2	4.27
Cuneus	551	2	-80	18	4.22
C) High vs. low empathy groups					
dMCC	316	4	4	64	5.01
R_Premotor	284	56	2	44	4.73
L_AI	280	-52	8	-12	4.73
L_Occipital	96	-46	-78	18	4.60
L_TPJ	192	-56	-58	20	4.45

dmPFC: dorsal medial prefrontal cortex. dMCC: dorsal medial cingulate cortex. SMA: Supplementary motor area. AI: anterior insula. TPJ: temporoparietal conjunction. L: Left, R: Right. Significant clusters were identified by combining a voxel-level threshold of $p < .001$ (uncorrected) and a cluster-level threshold of $p < .05$, *FWE corrected*.

Based on the results of the multi-region PPI, we chose the left TPJ as a seed to estimate the connectivity strength between the left TPJ (**Figure 4D**) and other brain regions when viewing others in pain in the self-rating phase in the first-level analysis. We then conducted a second-level analysis with the individual *W1* parameter ($W1_{pos}$ for the high empathy group and $W1_{neg}$ for the low empathy group) as a covariate.

Whole-brain analysis showed that the individual *W1* parameter modulated the connectivity of the left TPJ with the left AI (MNIxyz: -38/4/-10, Z_{stats}= 4.25, *p(cluster-FWE)* = 0.024), and with the vmPFC (MNIxyz: -8/50/-4, Z_{stats}= 4.63, *p(cluster-FWE)* < 0.001) differently in high and low empathy groups while participants watched painful videos during the self-rating phase (**Figure 5)**.

Specifically, the more strongly individuals weighted observational prediction errors (i.e., larger *W1* parameters), the weaker the left TPJ-vmPFC coupling in the low empathy group (r = -0.70, $p < 0.001$), and the stronger the left TPJ-vmPFC coupling in the high empathy group (r = 0.49, p = 0.010). Similarly, with increasing *W1* parameter, the coupling between the left TPJ and left AI increased in the high empathy group (r = 0.71, $p < 0.001$), and decreased in the low empathy group (r = -0.59, p = 0.001) (**Figure 5**). Importantly, the same AI-region that showed connectivity with the TPJ depending on the strength of the observational learning signal (*W1*) was also significantly correlated with the trial-by-trial empathy ratings in the baseline session ($t(51)$ = 2.31, p = 0.025), indicating that observational learning changed the communications of the TPJ with an AI region that is involved in the processing of empathy.

To test the specificity of these results, we performed a control analysis in which we estimated the connectivity strength between the left TPJ, and other brain regions when participants watched the painful videos in the observation phase (i.e., not the self-rating phase), and regressed this connectivity against the *W1* parameter in both groups. This analysis revealed no significant group differences in the impact of *W1* on TPJ connectivity even at a lenient threshold (i.e., $p < 0.05$, uncorrected). Thus, the group differentiating effect of the weight given to observational prediction errors on TPJ-AI as well as on TPJ-vmPFC connectivity was specific to the self-rating phase.

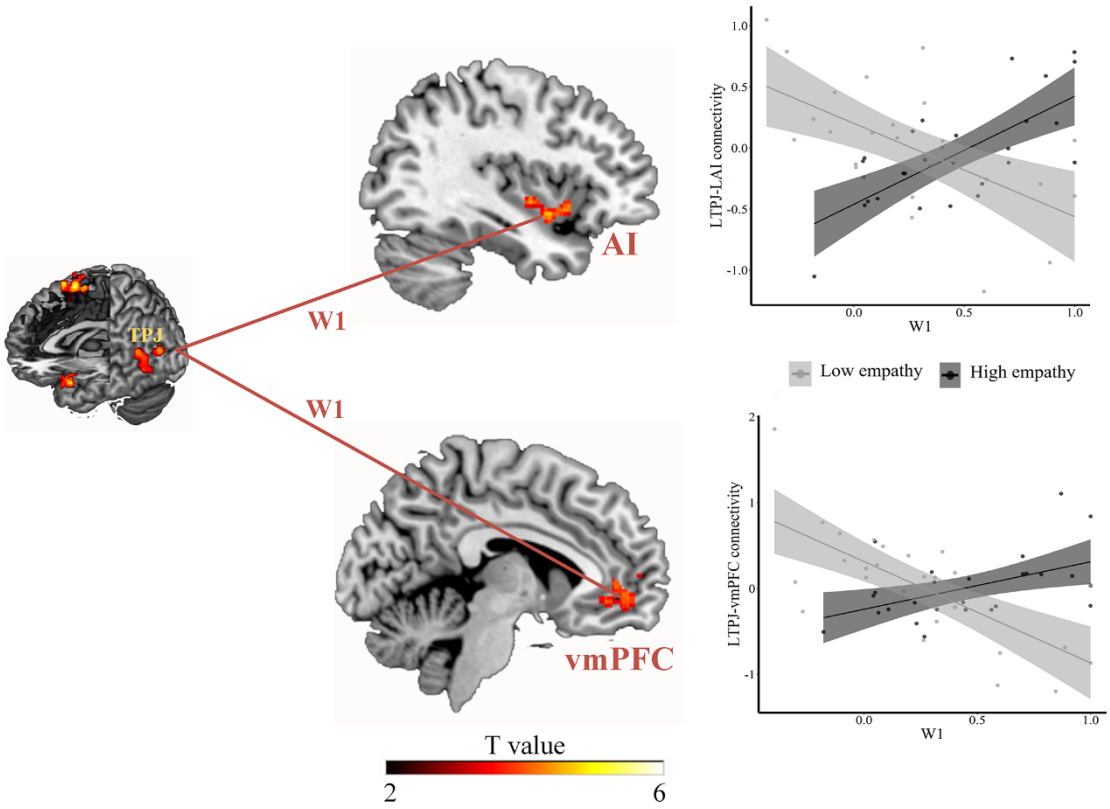

Figure 5. Group-specific impact of weight given to observational prediction errors on functional connectivity. The functional connectivity between the LTPJ and AI, and between the LTPJ and vmPFC during the self-rating phase correlated with the weights given to observational prediction errors across participants for both high and low empathy groups. Significant clusters were identified by combining a voxel-level threshold of $p < .001$ (uncorrected) and a cluster-level threshold of $p < 0.05$, *FWE corrected* across the whole brain. Display threshold at $p_{\text{uncorrected}} < 0.001$.

To specify the results of the PPI analysis, we tested whether the observed AI region is associated with changes in empathy during the self-rating phase. To do so, we regressed the individual *W1* parameters ($W1_{pos}$ for the high empathy group and $W1_{neg}$ for the low empathy group) against the neural activity to the painful videos in the self-rating phase, and calculated the contrast between the high and the low empathy groups. The results showed significant activation in the AI ($p = 0.021$, SVC-FWE corrected). Post hoc comparisons revealed that an increase in *W1* resulted in an

increase in AI activation in the high empathy group (r = 0.379, p = 0.056, **Figure 6A**), and in a decrease of AI activation in the low empathy group (r = -0.514, p = 0.007, **Figure 6A**). As our behavioral results indicated that the learning effects were preserved even when participants were no longer presented with the demonstrators' ratings (i.e., generalization session), we further compared the neural activations of left AI before (i.e., baseline session) and after (i.e., generalization session) learning between high and low empathy groups. The results showed a significant group (high empathy, low empathy) × session (baseline session, generalization session) interaction (peak = -36/-2/-6, p = 0.030, SVC-FWE corrected for the left AI cluster identified in the PPI analysis). More specifically, left AI responses were increased in participants in the high empathy group after learning ($t(25)$ = 2.18, p = 0.039, **Figure 6B**) and decreased in the low empathy group after learning ($t(25)$ = -2.52, p = 0.018, **Figure 6B**). The same analyses in vmPFC did not reveal any significant results (SVC-FWE correction, ps > 0.289). Together, these results suggest that the observational learning signals alter empathy-related responses at the neural level.

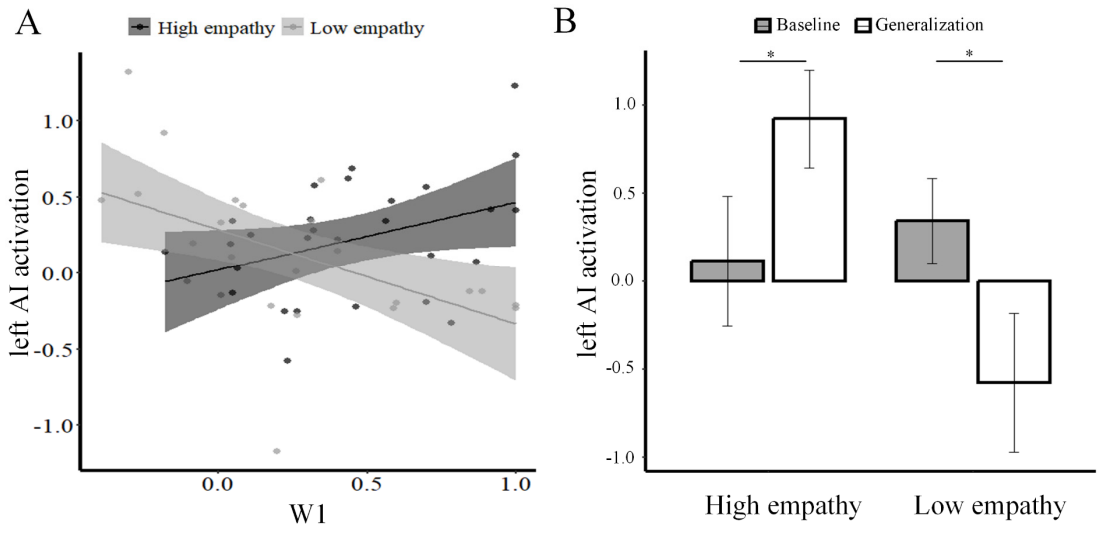

Figure 6. Neural responses in the AI region identified by the PPI analysis (Figure 5, upper panel). (A) The activation of AI correlated with the weights given to observational prediction errors across participants for both high and low empathy groups. (B) Participants in the high empathy group showed increased AI activation in generalization compared to baseline session whereas participants in the low empathy group showed the reverse pattern.

3. Results of the non-social control study

The results of our fMRI study demonstrated significant changes in empathy ratings in both high and low empathy groups. The computational model further linked the changes in empathy ratings to the weight given to observational prediction errors. However, it is possible that participants provided higher ratings in the high empathy group compared to the low empathy group only because they were shown larger numbers. Viewing larger or smaller numbers could anchor participants' responses on these values, thereby creating systematic biases (*34*). To examine this possibility as well as the extent to which the observational learning effect depends on observing the behavior of human vs. nonhuman computer demonstrators, we investigated

observational learning from non-social demonstrators (i.e., from computer-generated ratings) in a control study.

We first tested whether participants would learn to predict the observed empathy ratings of the computer demonstrators similarly to participants who learned to predict the observed empathy ratings of human demonstrators. To this end, we conducted an LMM with experiment (fMRI, non-social control), group (high empathy, low empathy), trial number and group × trial number as predictors, and participants' predictions of demonstrators' ratings as the dependent variable. The results revealed a significant group × trial number interaction ($\chi2(1) = 8.31$, $p = 0.004$). The experiment × group × trial number interaction was not significant ($\chi2(1) = 0.22$, $p = 0.64$, **Figure 7A**), indicating that participants paid attention to the computer-generated ratings and learned to predict them.

Next, we tested whether participants' empathy ratings were similarly influenced by human and computer demonstrators. As in the analysis of the fMRI study, we first fitted the participants' predictions of (computer) demonstrators' ratings using a Rescorla-Wagner reinforcement-learning model (*22*). The reinforcement learning model fitted the predictions of computer demonstrators' ratings adequately ($r^2 = 0.17 \pm 0.19$ for the high empathy group, $r^2 = 0.29 \pm 0.22$ for the low empathy group), and did not differ from the fMRI study ($t(50) = 0.996$, $p = 0.324$ for the high empathy group; $t(54) = -0.369$, $p = 0.713$ for the low empathy group). We then extracted the trial-wise observational prediction errors from both the fMRI study and the non-social

control study and fitted an LMM to directly test the association between trial-wise observational prediction errors and changes in empathy ratings. If participants change their empathy ratings based on the observational prediction errors, we would expect a positive association between trial-wise observational prediction errors and changes in empathy ratings. The LMM included experiment (fMRI, non-social control), empathy group (high empathy, low empathy), and trial-wise observational prediction errors, as well as their interactions as fixed effects predicting trial-wise changes of empathy ratings. The analysis revealed a significant experiment × prediction errors interaction ($\chi 2(1) = 5.34$, $p = 0.021$, **Table S3** for full statistical results) indicating a stronger relationship between trial-wise prediction errors and changes in empathy ratings in the fMRI study, compared to the control study (**Figure 7B**). Thus, although participants predicted the ratings of human demonstrators and computer demonstrators similarly well, the observations of the computer influenced the empathy ratings of participants to a lesser extent than the observations of the human demonstrator.

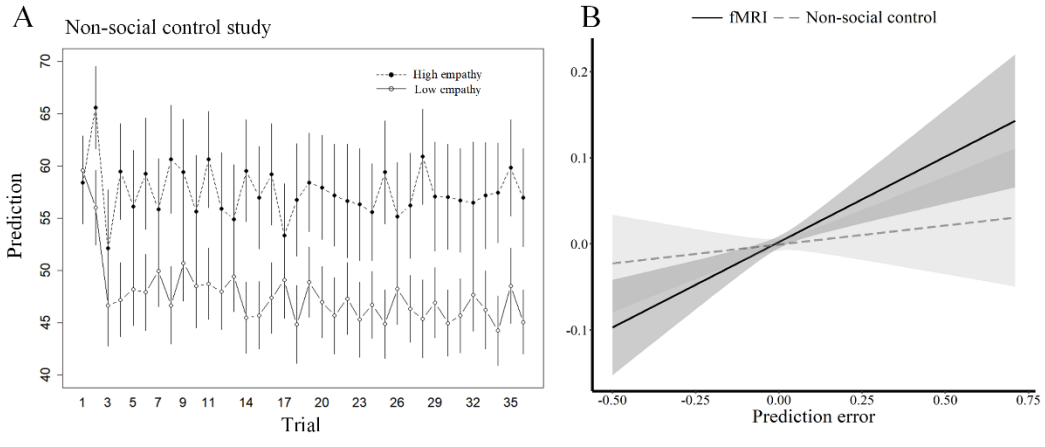

Figure 7. **Participants learn, but to a lesser degree, from computer compared to human demonstrators.** (A) Participants' predictions diverged in the high and low empathy groups of the non-social control study. (B) Experiment × prediction error

516 interaction. The observational prediction errors shifted participants' empathy ratings
517 more strongly in the fMRI study compared to the non-social control study.
518
519 **4. Results of the behavioral replication study**
520 To test the reproducibility of the learning effects observed in the fMRI study, we
521 conducted a behavioral study with the identical paradigm on an independent sample.
522 In addition, participants were seated alone in the behavioral experimental rooms in
523 which their ratings were unobserved and they would not interact with the
524 experimenter to reduce the effect of social desirability. We first analyzed the
525 predictions from the observational learning phase. To this end, we conducted an
526 LMM with group (high empathy, low empathy), trial number and group × trial
527 number as predictors, and participants' predictions of ratings as the dependent
528 variable. The results revealed a significant group × trial number interaction ($\chi2(1) =$
529 67.4, $p < 0.001$, **Figure 8A**), indicating that participants expected increasing empathy
530 ratings of the demonstrators in the high ($\chi2(1) = 14.37$, $p < 0.001$) and decreasing
531 empathy ratings in the low empathy ($\chi2(1) = 62.4$, $p < 0.001$) group also in this
532 independent sample.

533 We then tested the participants' own empathy ratings in the self-rating phase of
534 the observational learning session. The LMM with group (high empathy, low
535 empathy), trial number and group × trial number as predictors, and participants'
536 empathy ratings as the dependent variable revealed a significant group × trial number
537 interaction ($\chi2(1) = 18.56$, $p < 0.001$). Replicating the results of the fMRI study,
538 participants showed increasing empathy ratings in the high empathy group ($\chi2(1) =$

4.46, $p = 0.03$), and decreasing empathy ratings in the low empathy group ($\chi 2(1) = 16.82$, $p < 0.001$) over the course of learning. We also analyzed the empathy ratings of our participants over the whole experiment (i.e., from the baseline session to the generalization session). To this end,we conducted an LMM with group (high empathy, low empathy), session (baseline session, observational empathy learning session 1-4 and generalization session, coded as 0-5 respectively) and group × session as predictors, and participants' empathy ratings as the dependent variable. Similar to the results of the fMRI study, we found a significant group × session interaction ($\chi 2(1) = 34.4$, $p < 0.001$), with an increase in ratings across sessions in the high empathy group ($\chi 2(1) = 9.13$, $p = 0.003$) and a decrease in ratings across sessions in the low empathy group ($\chi 2(1) = 37.6$, $p < 0.001$). In summary, the prediction data and participants' own empathy ratings in the behavioral replication study resembled those of the fMRI study.

To test if these changes in empathy ratings were associated with the observational learning mechanism revealed in Study 1, we first fitted the participants' predictions using a Rescorla-Wagner reinforcement-learning model (*22*). The reinforcement learning model fitted the predictions well ($r^2 = 0.22 \pm 0.19$ for the high empathy group, $r^2 = 0.28 \pm 0.19$ for the low empathy group), and did not differ from the fMRI study ($t(49) = 0.001$, $p = 0.999$ for the high empathy group; $t(51) = -0.212$, $p = 0.833$ for the low empathy group). We then extracted the trial-wise observational prediction errors and associated them with trial-wise changes in empathy ratings in an LMM.

The results revealed that trial-wise observational prediction errors positively predicted the trial-wise changes of empathy ratings ($\chi 2(1) = 25.75$, $p < 0.001$, **Figure 8B**), and similarly well in the high and the low empathy group ($\chi 2(1) = 0.225$, $p = 0.61$).

To compare studies more thoroughly, we also integrated an additional LMM with experiment (fMRI, behavioral replication), empathy group (high empathy, low empathy), and trial-wise observational prediction errors, as well as their interaction to predict the trial-wise changes of empathy ratings. The analysis showed that the experiment × prediction errors interaction effect was not significant ($\chi 2(1) = 0.55$, $p = 0.46$, **Figure 8B**, **Table S3** for full statistical results), compatible with the notion that participants' empathy ratings were similarly influenced by the observational prediction errors in the fMRI study and the behavioral replication study. In summary, the behavioral replication study resulted in similar behavior as the fMRI study.

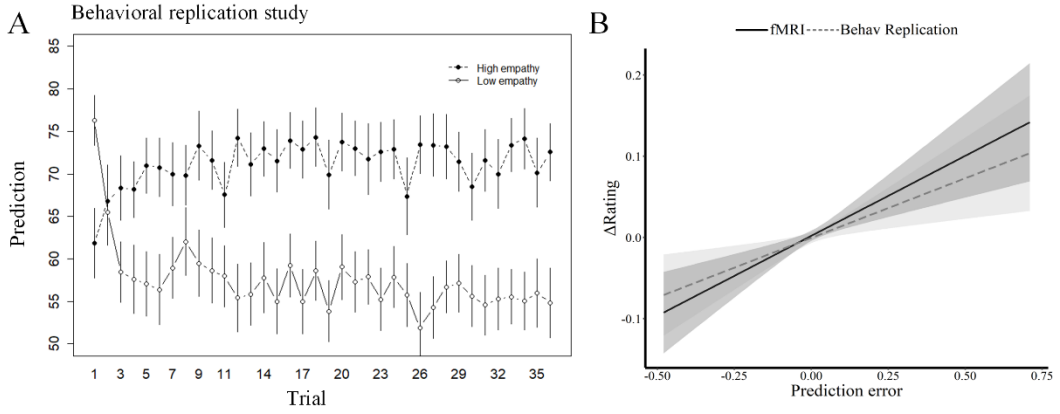

Figure 8. **Replication of behavior in the fMRI study.** (A) Trial-by trial prediction in the behavioral replication study. The results showed differential effects in the high and low empathy groups. (B) Effect of prediction error on changes in empathy ratings in the behavioral replication study (red) and the fMRI study (blue). The interaction between experiment × prediction error was not significant, indicating comparable observational learning of empathy in both studies.

Discussion

The assumption that empathy can be transmitted between individuals forms the basis of influential theories of moral development (*1*). Here, we provide mechanistic insights into the social transmission of empathy. Confirmed in two independent studies and substantiated by a control study, our results showed that empathy is transmitted by learning from observed empathic reactions of others. The observational learning of empathy can increase or decrease empathy in the observer, depending on the role model the participants learn from. Notably, the learning-related changes in empathy were elicited by observing empathic responses of an unknown, random individual, and expressed themselves on the subjective (empathy ratings) and neural level (connectivity between TPJ and an AI region that correlated with trial-by-trial empathy ratings as well as the neural activity of AI region). This indicates that the social transmission of empathy occurs in 'random' social interactions and changes the neural responses to the misfortune of others, here their pain.

The finding that observing empathic responses in others changes empathic responses in the observer is important, because empathy is commonly related to an increase in prosocial behavior (*35*, *36*). In line with these findings, the learning-related increase in empathy ratings was related to an increase in participants' willingness to invest time to help another person. From a policy point of view, these results suggest that creating a highly empathic environment may enhance prosocial tendencies. On the flipside, our findings also show that the presence of non-empathic individuals can undermine empathy and prosocial motivation.

It has been shown before that empathy ratings of a group can shift individual empathic feelings and influence donations to a homeless shelter (*4*). Going beyond these previous results, our study reveals a mechanism through which empathy is transmitted across individuals. We show that the extent to which people change their subjective and neural responses to the pain of others is predicted by the weight they give to the prediction-error signal generated by the discrepancy between expected and observed empathy ratings of others. Specifically, our results show that participants generate positive observational prediction errors if human demonstrators display a stronger empathic reaction than expected, and, as a result, increase their empathy ratings. In contrast, being confronted with individuals who show less empathy than expected results in negative prediction errors and a decrease in empathy ratings of the observer.

It is well established that observational learning parameters can predict differences in socially relevant phenomena such as the social transmission of fear (*13–15*), and the social modulation of risk (*9*) and choice preferences (*7*, *8*). In influential theoretical models, observational learning has long been assumed to constitute a mechanism for the social transmission of empathy (*5*). Providing the first empirical evidence for this notion, we show that an observational learning model can predict the extent to which empathy is transmitted from one individual (i.e., the demonstrator) to another (i.e., the observer) and applied by the observer to third parties uninvolved in the learning process (generalization).

We find that learning from observing other's empathic reactions does not only change participants' empathy ratings, but also their neural responses to other's pain. Specifically, the weight participants assigned observational prediction errors modulated connectivity between regions associated with observational learning, such as the TPJ (*12*, *16*, *37*), and regions associated with the processing of other's pain, such as the AI (*17*, *25–27*, *29*, *30*). Taking an individual difference perspective, the more strongly a person weighted the observational prediction errors, the stronger the coupling of left TPJ-AI in the high empathy group, and the weaker the TPJ-AI coupling in the low empathy group. Apart from this, the individual differences in the magnitude of observational learning (i.e., weight parameter) also modulated the neural activations in the AI. Thus, the empathy shown by the role model modulated the way in which observational prediction error weights affected brain connectivity.

The finding of the processing of observational prediction errors in the left TPJ is in line with recent evidence linking this region to social influence on reward learning (*12*, *16*) and prosocial decision making (*38*). Extending these previous results, our findings show that learning by observing high and low empathic individuals modulates the connectivity between the left TPJ and the AI as well as the vmPFC. Importantly, the AI region that was modulated by learning was also active when participants observed another person in pain in the baseline session. Therefore, observational learning indeed changed the processing of other's pain in the AI, i.e., a region that forms a central part of the empathy network (*17*, *25–27*, *29*, *30*).

643 Neural responses in the vmPFC have been related to value computation in general
644 (*39*), and in particular, to the computation of the value of pain (*40*). Given the present
645 findings, it is possible that observing empathic responses of others changes
646 participants' valuation of the pain of others to justify an increase or decrease in their
647 own empathy ratings. Together, our neural findings uncover a neural mechanism for
648 the social transmission of empathy that can explain the plasticity of empathic
649 responses in different social environments.

650 Although we show a change in empathy ratings and neural responses to the pain
651 of others that is closely predicted by learning parameters, alternative explanations to
652 observational learning have to be considered. First, the observed changes in subjective
653 and neural empathy responses may reflect mere imitation of motor responses. The
654 results of the non-social control study argue against this alternative explanation.
655 Although participants paid attention to, and learned to predict, the computer-generated
656 ratings equally well as those of the human demonstrators, they did not use the learned
657 information to update their own empathy as much as with human demonstrators.
658 Second, participants may have changed their ratings to conform with the ratings of the
659 demonstrator. Testing this assumption, we found no significant relationship between
660 participants' ratings on a well-established conformity scale (*41*) and their changes of
661 empathy ratings in the observational-learning-of-empathy task. Third, and related to
662 conformity, participants may have shown higher empathy ratings in the high-empathy
663 group to please the demonstrator or the experimenter. We assessed individual

differences in social desirability (based on a social desirability scale, (*19*) and found no significant relationship with the observed changes in empathy ratings. In addition, in the behavioral replication study, participants were seated alone during the experiment, such that they were unobserved and could not interact with the experimenter. Although this setting minimized the influence of social desirability, the findings still replicated the learning-related changes in empathy ratings observed in the fMRI study. Based on this evidence, and given that the estimates from our observational learning model fitted the changes in empathy ratings and neural responses to other's pain, observational learning is likely to contribute to the social transmission of empathy.

That said, we acknowledge that our study was based purely on female participants, which allowed us to control for unspecific gender effects (e.g., induced by gender-mixed pairings of participants and confederates), but limits the generalizability of our results. Future studies should test the effect of observational learning of empathy in males. Moreover, although our results show that learning from observing the empathic reactions of a demonstrator changes the willingness of observers to invest time to help a recipient of pain, it would be important to investigate changes in actual prosocial behavior in real life.

In sum, our study shows how empathy spreads in random social interactions and provides a computational and neural mechanism for the social transmission of empathy across societies.

Materials and methods

Participants

We recruited three independent samples that are described below.

Study 1 - fMRI study. 55 healthy females (mean age ± SD = 21 ± 2.1 years) participated in the fMRI study as paid volunteers. We chose an all-female instead of a gender-mixed group of participants so that we could also use all-female confederates and avoid the complications of the gender-mixed pairing of participants and confederates. Three participants were excluded from further analyses due to excessive head movements (> 3 mm) during scanning. The analyses included data from 52 participants (mean age ± SD = 21 ± 2.1 years; 26 in the high empathy group).

Study 2 - Non-social control study. 57 healthy females (mean age ± SD = 20.8 ± 2.4 years) participated in the non-social control study as paid volunteers. One participant was excluded because of technical issues during the experiment. Data from 56 participants were analyzed (26 in the high empathy group). Of these, one participant did not fill in the questionnaire (see below).

Study 3- Behavioral replication study. 56 healthy females (mean age ± SD = 20.8 ± 1.9 years) participated in the behavioral replication study as paid volunteers. Four participants were excluded because of technical issues during the experiment. Data from 52 participants were analyzed (25 in the high empathy group).

The experimental procedures were approved by the local Research Ethics Committee (No. 2018-01-04). All participants had normal or corrected-to-normal

vision, no history of psychological or neurological disorders, and 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. The sample size for the studies was determined by an *a priori* power analysis using G*Power 3.1(*42*) for a within-between interaction in a repeated-measures analysis of variances (ANOVA) design with two groups (groups: high empathy, low empathy) and two measurements (time: before learning, after learning). A total sample size of 46 participants (23 participants per group) was required for each study to detect a medium effect size of $f = 0.25$ at $\alpha = 0.05$ (two-tailed) with a power of 90%. We recruited more than 46 participants in all studies to account for possible data loss.

Questionnaires

In Studies 1 and 3 (i.e., the studies with human demonstrators), participants rated their impression of the demonstrator before and after the experiment (*17*, *19*, *26*, *43*). In addition, participants rated the perceived empathy of the demonstrator ("How empathic do you find this person?") from 1 (not empathic at all) to 9 (extremely empathic). In Study 1, participants also rated the the perceived pain intensity of the recipient ("How much pain did the person in the video clip experience?") from 1 (none at all) to 9 (extreme) and indicated how much time they would like to spend comforting the recipient (0-60 min in 1 min increments), an item that was used to measure prosocial tendencies in previous studies (*44*).

Participants of all three studies completed the social desirability scale (SDS-17, (*20*) as well as the conformity scale (*41*) to measure their propensity to respond in a socially desirable manner and their tendency to conform to others. We used the Interpersonal Reactivity Index (IRI, *17*) and the subscales measuring empathy and behavioral contagion from the Empathy Index (*45*) to measure the trait empathy. There were no differences in these trait measures across the three studies ($ps > 0.166$, **Table S4**).We also compared the trait measures between groups (i.e., high and low empathy group) within each study, and the results revealed no significant difference in these trait measures between the high and low empathy group for all studies ($ps > 0.068$, **Table S5**).

Preparation and validation of the stimulus set

For the purpose of this study, we recorded videos of four different females receiving painful and non-painful stimulation. In each video clip, two pain electrodes were visibly attached to the recipient's right hand. The recipient reacted to the shocks by twitching her hand and arm when receiving a painful electrical stimulation and acted calmly when receiving a non-painful electrical stimulation. For each recipient, we recorded at least 10 video clips showing painful stimulation and 4 video clips showing non-painful stimulation with a duration of 2 s each. We then selected 25 out of the 40 video clips showing painful stimulations for further stimulus validation.

To validate the video clips, we conducted an online study with 37 female participants (mean age ± SD = 21.9 ± 4.4 years).The rating task was completed

electronically via a Qualtrics link (https://www.qualtrics.com/). Participants were instructed to watch the 25 video clips and to rate the pain intensity felt by the recipient ("How painful do you think the model feels?") on a 7-point Likert scale (1 = not painful at all, 7 = extremely painful). The order of the presentation of the video clips was randomized. Based on these ratings, we selected four video clips showing painful stimulation for each recipient (16 video clips in total). We then averaged the pain intensity ratings for each recipient and conducted further statistical tests. The mean pain intensity ratings were comparable across recipients ($F(3,34) = 0.473$, $p = 0.703$, $\eta^2_p, = 0.040$, **Figure S2**).

Experimental design and procedure

Study 1 - fMRI study

Prescanning procedure

Before the experiment, participants briefly met two other individuals (confederates who were not known by the participant) who were trained to act as the demonstrators during the observational learning task. Participants and confederates were instructed together. They were told that the current study was part of a project on pain perceptions and that they would be randomly assigned to one of two groups; a 'recipient' group that would receive painful or non-painful electrical stimulations, or an 'observer' group that would watch the stimulation of the recipients and rate their feelings. The participants and the two confederates were ostensibly assigned to the 'observer' group.

768 Next, the individual pain thresholds of the participants and confederates were
769 determined by a standard procedure (*17*, *46*, *47*) to provide a first-hand experience of
770 the stimulation they would observe in recipients. To do so, participants and
771 confederates entered into a private room successively in which another experimenter
772 performed the pain threshold assessment. More specifically, two pain electrodes were
773 attached to the back of the left or right hand. Using a Digitimer DS7 electrical
774 stimulator, a low-voltage electric shock (0.5 mA) was delivered and increased in
775 increments of 0.5 mA. Participants and confederates were asked to rate the intensity
776 of the respective electrical stimulation from 0 (not painful at all) to 10 (extremely
777 painful). Participants and confederates were informed that the recipients would
778 receive pain stimulation with the intensity they rated as "8" and non-painful
779 stimulation with the intensity they rated as "1" in the pain thresholding procedure.

780 After measuring individual pain thresholds, the experimenter introduced the
781 empathy rating scale. Participants and confederates were told that they would be
782 asked to indicate how they felt when watching a video clip of a recipient on a scale
783 from 0 (did not feel anything) to 100 (feeling extremely bad). Next, the participants
784 received instructions for the observational empathy learning task in the preparation
785 room while the two confederates were seated outside. They were then instructed that
786 apart from reporting their feelings when watching the video clips, their task would be
787 to predict the ratings of the demonstrators (i.e., the two confederates) as accurately as
788 possible. To help with their predictions, the participants would see the rating of the

789 demonstrator in real time after their prediction. We made clear to the participants that
790 their own ratings were personal and could not be observed by others.

791 *Scanning Procedure*

792 The fMRI scanning session consisted of a baseline session, an observational empathy
793 learning session, and a generalization session.

794 In the baseline session, the participant in the scanner watched the video clips of a
795 person receiving either painful (18 trials) or non-painful (12 trials) stimulations. Each
796 trial started with a lightning bolt symbol (1000 ms) indicating the pain intensity the
797 recipient was about to receive (bright = painful; dark = non-painful). After a fixation
798 period (500 – 1500 ms), the video showed the hand of the recipient undergoing
799 stimulation for 2000 ms. Participants were then asked to report their current feelings
800 from 0 (felt nothing at all) to 100 (felt extremely bad) in 5000 ms.

801 The observational empathy learning session was adapted from an observational
802 learning paradigm we used previously (*7*, *8*). In each trial, an observation phase (i.e.,
803 observing the demonstrator's empathy ratings) was followed by a self-rating phase
804 (i.e., making empathy ratings oneself; **Figure 1**). To distinguish the two phases and
805 the different demonstrators, the beginning of each phase was marked with arrows in
806 different colors (500 ms) pointing away (observation phase) or towards (self-rating
807 phase) the participant. During the observation phase, the lightning bolt symbol (1000
808 ms) was shown followed by the presentation of a video clip (2000 ms). Participants
809 were told that the demonstrator had watched this video and rated her feelings. Then,

participants had 5000 ms to predict the demonstrator's ratings. After that, the rating of the demonstrator was presented (2000 ms). Next, an arrow pointing to the participant indicated the start of the self-rating phase (500 ms). After the presentation of the lightning bolt (1000 ms) and the video clip (2000 ms), participants were asked to rate how they felt when watching the video on a scale from zero (not feeling anything) to hundred (feeling extremely bad) (5000 ms). The videos used in the observation and the self-rating phase showed the same recipient receiving the same type of stimulation (i.e., either depicting painful or non-painful stimulation).

The observational empathy learning session consisted of four blocks, with 12 trials in each block, resulting in 48 trials in total (36 trials of painful and 12 trials of non-painful videos). To prevent habituation, participants saw the video clips of two different recipients (one recipient for two blocks) in the observational empathy learning session.

Unbeknownst to the participants, the ratings of all demonstrators were generated by a pre-defined algorithm, based on the participant empathy ratings in the baseline session. In the high empathy group, the observed ratings for pain videos were drawn from a normal distribution in which the mean equaled the participant mean in the baseline session plus three standard deviations (SD = 5). In the low empathy group, they were drawn from a normal distribution in which the mean equaled the participant mean in the baseline session minus three standard deviations (SD = 5). As a result, in the high empathy group, participants observed empathy ratings that were consistently

831 higher, and in the low empathy group they observed ratings that were consistently
832 lower than their baseline ratings for the painful videos. The observed ratings for non-
833 painful videos were sampled from a normal distribution in which the mean of the
834 distribution was the individual mean in the baseline session (SD = 5).

835 The generalization session was identical to the baseline session, except that the
836 participants provided emotion ratings when observing a new recipient, i.e., video clips
837 that were not part of the baseline or the observational empathy learning session. The
838 participant and confederates were informed that they would not meet after the study
839 and had separate visual displays to keep empathy ratings anonymous.

840 Study 2 - Non-social control study

841 The task of the control study was identical (i.e., instructions, number of sessions,
842 number of blocks, and number of trials) to the task of the fMRI study described
843 above, except that participants were told that they observed ratings generated by two
844 computers.

845 Study 3 - Behavioral replication study

846 To test the robustness of the learning effects observed in the fMRI study, we
847 conducted a behavioral study on an independent sample. The experimental procedure
848 was identical to the procedure of the fMRI study described above, except that the
849 demonstrators were represented by real participants instead of confederates. Care was
850 taken to ensure that the participants had neither met nor known each other before the
851 study. To further minimize a potential effect of reputation concerns on empathy

ratings, participants were seated alone in the laboratory, i.e., the experimenter was not present and did not interact with the participants during the experiment. Importantly, the ratings of the demonstrators in the observational-learning-of-empathy session were also generated with the pre-defined algorithm described above.

MRI Image acquisition

We acquired functional and anatomical images with a Siemens Trio 3.0 T MR scanner using a 12-channel phase-array head coil at the Center for MRI Research, Peking University. Multiband functional images were acquired with T2-weighted, gradient-echo, echo-planar imaging sequences sensitive to BOLD contrast (matrix = 112×112, 62 slices, 2×2×2 mm^3 voxel size, interslice gap = 0.3 mm, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, field of view (FOV) = 22.4 × 22.4 cm, flip angle (FA) = 90°, interleaved slice acquisition, multiband acceleration factor = 2). A high-resolution anatomical T1-weighted image was acquired for each participant (256×256mm matrix, 192 slices, 1×1×1.00 mm^3 voxel size; TR = 2530 ms, TE = 2.98 ms, inversion time (TI) =1100 ms, FOV = 25.6×25.6 cm, FA = 7°). Padded clamps were used to minimize head motion and earplugs attenuated scanner noise.

Data analyses

Regression analyses

We performed linear mixed models (LMM, 'lme4') in R v.4.1.1 (R Development Core Team, 2012) for the behavioral analyses on empathy ratings and prediction ratings as the dependent variables to investigate observational learning. In particular,

we conducted LMMs with empathy group (high empathy, low empathy), time and empathy group × time as predictors, and the empathy ratings or prediction ratings as the dependent variable. The time variable corresponds to the trial number (i.e., 1-36 trials) during the observational empathy learning session or the session number of the whole experiment (baseline, observational empathy learning (1-4) and generalization, coded as 0-5 respectively). We predicted significant empathy group × time interactions for both the empathy ratings and the prediction ratings. Specifically, we hypothesized that participants' prediction and empathy ratings would diverge between high and low empathy groups over the course of learning. We used participants as random intercepts.

In addition, we performed LMMs to compare observational learning effects as captured by computational models between studies. Specifically, experiment (fMRI, non-social control/ behavioral replication), empathy group (high empathy, low empathy), and trial-wise observational prediction errors (obtained in the reinforcement learning model) as well as their interactions were included as a fixed effect to predict the trial-wise changes of empathy ratings. We also used by-participant intercepts for all LMMs.

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).

Computational modeling

To investigate the mechanisms underlying changes in empathy on a trial-by-trial basis in the observational learning session, we employed a computational modeling approach (*19*, *23*, *24*). The results were based on the original (raw) ratings. Using normalized ratings revealed similar results. First, we modeled the predictions participants made regarding the ratings of the demonstrators using a standard Rescorla-Wagner (*22*) reinforcement learning (RL) algorithm in the observation phase. The RL model assumes that participants changed their predictions when the demonstrator ratings differed from the ratings expected by the participants.

$$V(t+1) = V(t) + \alpha \times \delta_i \quad [1]$$

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

Thus, on each trial *t*, the (future) predictions $V(t+1)$ of demonstrator ratings are a function of current predictions $V(t)$ and the prediction error δ (Equation 1), which corresponds to the difference between the actual demonstrator rating *R(t)* at trial *t* and the current prediction $V(t)$ (Equation 2). In our study, the demonstrator's rating can be higher or lower than expected. Observing higher ratings than expected generates a positive prediction error, while observing lower ratings than expected generates a negative prediction error. The learning rate α ($0 \leq \alpha \leq 1$) controls the

913 extent to which the current predictions of demonstrators' ratings are updated by new
914 information.

915 Next, we formally modelled the participants' empathy ratings in the self-rating
916 phase as a linear function of prediction errors elicited by demonstrator's empathy
917 ratings in the preceding observation phase. In all models, we assumed that
918 participants' ratings are a linear combination of the time-discounted sum of previous
919 observational prediction errors (as originating from the RL model, Equations 1-2) and
920 participants' baseline ratings ($Empathy(t0)$), which were defined as the individuals'
921 mean ratings towards painful videos in the baseline session when no social influence
922 was implemented.

923 We considered models which separated the first and second half of the
924 observational learning session as these two halves used different recipients of pain
925 stimulation in the videos. Moreover, we found that empathy group (high, low) and
926 session half (first, second) interacted for observational prediction errors ($\chi^2(1) =$
927 19.82, $p < 0.001$). Specifically, in the first half, the observational prediction errors
928 were mostly positive for the high empathy group and mostly negative for the low
929 empathy group, resulting in a group difference ($\chi^2(1) = 33.36$, $p < 0.001$). In contrast,
930 in the second half, the observational prediction errors were close to zero for both
931 groups, resulting in no difference between groups ($\chi^2(1) = 0.07$, $p = 0.79$). We also
932 considered models with common and separate weighting of positive and negative
933 prediction errors:

$$Empathy(t) = Empathy(t0) + W \sum_{j=1}^{t} \gamma^{t-j}\, \delta_j \quad [3]$$

$$Empathy(t) = Empathy(t0) + W_{pos} \sum_{j=1}^{t} \gamma^{t-j}\, \delta pos_j + W_{neg} \sum_{j=1}^{t} \gamma^{t-j}\, \delta neg_j \quad [4]$$

$$Empathy(t) = \begin{cases} Empathy(t0) + W1_{pos} \sum_{j=1}^{t} \gamma^{t-j}\, \delta pos_j + W1_{neg} \sum_{j=1}^{t} \gamma^{t-j}\, \delta neg_j, & t < 25 \\ Empathy(t0) + W2_{pos} \sum_{j=1}^{t} \gamma^{t-j}\, \delta pos_j + W2_{neg} \sum_{j=1}^{t} \gamma^{t-j}\, \delta neg_j, & t \geq 25 \end{cases} \quad [5]$$

$$Empathy(t) = Empathy(t0) + k \times R(t) \quad [6]$$

The winning model 3 (Equation 5) considered the empathy rating in the first and second half of the observational learning session separately, separated the prediction errors by sign, and added them up separately. This model included the parameters *W1* and *W2,* which capture the magnitude (weight) of the influence of observational prediction errors on changes in participants' empathy ratings in the first and second half of the observational learning session. The *W* parameter ranges from -1 to +1 because one represents the maximum of the empathy ratings after the transformation (i.e., divided by 100). A larger *W* corresponds to a stronger influence of observational prediction errors on participants' empathy ratings. The discount parameter γ ($0 \leq \gamma \leq 1$), captures an exponential decay of the influence of previous observational prediction errors over time, such that the more recent observational prediction errors have a greater impact on participants' empathy ratings than the earlier observational prediction errors. If γ is close to one, all preceding observational prediction errors receive the same weight, and if it is close to zero, only the last observational prediction error leads to subsequent changes in participants' empathy ratings.

We also tested less complex models in which positive and negative prediction errors were not modelled separately (Equation 3, Model 1) or the empathy ratings

were not fitted separately for the first and second half of the observational learning session (Equation 4, Model 2). Moreover, we tested an imitation model in which participants were allowed to differ in the extent to which they copied the demonstrators' ratings (Equation 6, Model 4). In this model, *k* represents the imitation parameter and *R(t)* is the actual demonstrator rating at trial *t*. We fitted all computational models to participants' ratings of the painful videos in both high and low empathy groups.

Parameter estimation

We optimized model parameters by minimizing the negative logarithm of the posterior probability (LPP) over the free parameters using MATLAB's fmincon function, initialized at multiple starting points of the parameter space.

$$\text{LPP} = -\log\big(P(\theta_M|D,M)\big) \propto -\log(P(D|M,\theta_M)) - \log(P(\theta_M|M))$$

Here, $P(D|M,\theta_M)$ is the likelihood of the data given the considered model M and parameter values θ_M, and $P(\theta_M|M)$ is the prior probability of the parameters. Following previous research (*48*), the prior probability distributions for the learning rate were defined as beta distributions (beta pdf(α,1.1,1.1)). For the weight parameters and forgetting parameters, the prior distributions were unknown and assumed to be uniform, such that every value in the parameter range had equal probability. Formally, this is equivalent to maximum likelihood estimation (*49*).

Model comparison

We computed the Laplace approximations to the model evidence (LAME) as criteria for model comparison, which measure the ability of each model to explain the experimental data, by trading-off their goodness-of-fit and complexity (*48*, *50*).

$$LAME = \log(P(D|M,\theta_M)) + \log(P(\theta_M|M)) + \frac{df}{2}\log 2\pi - \frac{1}{2}\log|H|$$

Where df is the number of model parameters, and |H| is the determinant of the Hessian.

The individual model comparison criteria (LAME) were then fed to the mbb-vb-toolbox (https://code.google.com/p/mbb-vb-toolbox/). For each model within a set of models, we estimated the exceedance probability (denoted XP), given the data gathered from all subjects. XP quantified the belief that the model was more likely than all the other models in the model space. An XP > 95% for one model within a set is typically considered as significant evidence in favor of this model being the most likely.

MRI Image analyses

Preprocessing

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 slice-time corrected, realigned, and coregistered to the anatomical image of the participant. The anatomical image was processed using a unified segmentation procedure combining segmentation, bias correction, and spatial normalization to the MNI template (*51*), the same normalization parameters were then used to normalize the EPI images. Lastly, the

997 functional images were spatially smoothed using an isotropic 6 mm full-width at a
998 half-maximum (FWHM) Gaussian kernel.

999 *First-level analysis*

1000 We first sought to identify neural regions that tracked trial-by-trial empathy
1001 ratings. To do so, we interrogated event-related general linear models (GLMs) in the
1002 baseline session. We included the onsets and durations of (1) the lightning bolt
1003 indicating the level of pain intensity; (2) the videos of recipients undergoing
1004 electrical stimulations, parametrically modulated by the trial-by-trial empathy ratings
1005 of participants; and (3) participant ratings. These regressors were convolved with the
1006 canonical hemodynamic response function and its time derivatives. The model also
1007 contained six (three translation and three rotation) regressors to account for motion.

1008 To examine neural activity correlating with observational prediction errors, we
1009 investigated GLMs for the observational empathy learning session. We included the
1010 onsets and durations of: (1) the cues indicating the beginning of the observation
1011 phase or self-rating phase; (2) the electric bolt indicating the level of pain intensity
1012 (modelling painful and non-painful stimulations separately); (3) the videos of
1013 recipients receiving electrical stimulations (modelling separately for the painful and
1014 non-painful stimulation videos in the observation phase and self-rating phase); (4)
1015 the prediction of the demonstrator rating (modelled separately for the painful and
1016 non-painful stimulations); (5) the ratings of the demonstrator (modelled separately
1017 for the painful and non-painful stimulations), parametrically modulated by

observational prediction errors derived from the reinforcement learning model (see computational model for details); and (6) participant ratings. These regressors were again convolved with the canonical hemodynamic response function and its time derivatives, and the model contained six (three translation and three rotation) regressors to account for motion. The results are based on the original (raw) ratings. Using normalized ratings revealed similar results.

Second-level analysis

First, we assessed the regions tracking the trial-by-trial empathy ratings in the baseline session. We brought the first-level contrast images created by the parametric modulator of empathy ratings to the second level and tested against zero in a one-sample t-test.

Next, we investigated the regions encoding observational prediction errors. First, we investigated the high and low empathy groups separately and identified regions encoding observational prediction errors (i.e., by setting the prediction error regressor to "1") or inverse observational prediction errors (i.e., by setting the prediction error regressor to "-1") in one-sample t-tests at the second level.

We then collapsed all contrast images created by the observational prediction error parametric modulator from the first level and compared them between high and low empathy group at the second level. Imaging results were obtained in whole-brain analyses, using a combined voxel-level threshold of $P_{\text{uncorrected}} < 0.001$ and a family-wise error (FWE) corrected cluster-level threshold of $P < 0.05$.

Psychophysiological interaction (PPI) analyses

To examine how neural activity related to observational prediction errors influences neural responses in the self-rating phase and lead to the differential responses between high and low empathy groups, we performed psychophysiological interaction (PPI) analyses (*52*, *53*). We used the generalized PPI (gPPI) toolbox (https://www.nitrc.org/projects/gppi), which has the benefit of accommodating multiple task conditions in the same connectivity model (*54*). Given that multiple regions were associated with the differential encoding of observational learning prediction errors between groups (i.e., **Table 1C**), we first conducted a multi-region PPI analysis (*32*) to identify brain regions that changed their functional connectivities with other regions depending on the individual size of the *W1* parameter, i.e., the parameter associated with the change in empathy across participants in the behavioral analyses, **Figure 3G** and **3H**). To do so, we defined regions of interest (ROIs) using the full set of activated clusters related to the differential processing of observational prediction errors between groups (**Table 1C**). Next, we used each of these ROIs as a seed and obtained the respective connectivity strengths with other regions across the whole brain (264 regions based on an established template (*33*) when participants watched others in pain in the self-rating phase (vs. the implicit baseline). Finally, we correlated the connectivity strength with the *W1* parameter. To prevent arbitrariness in the definition of the seed region, we defined it with different thresholds, ranging from 0.001 to 0.05, which led to similar conclusions (see (*32*) for a similar approach).

The multi-region PPI analysis revealed that the connectivity between the left TPJ and the rest of the brain showed the strongest modulation by the *W1* parameter. As such, we focused on the left TPJ in a follow-up PPI analysis. We extracted the time series of the left TPJ (the region tracking the observational prediction error) as the physiological regressor. Psychological regressors were then convolved onset regressors and parametric modulators. Psychophysiological interaction (PPI) terms were created by multiplying the time series from the psychological regressors with the physiological variable. All of the above were performed for each participant separately, and individual gPPI models were created by including the physiological variables, the psychological regressors, and the PPI terms (*54*).

The physiological, psychological, and psychophysiological interaction regressors as well as six motion parameters were then entered into the GLM. We first used this GLM to determine regions in which connectivity strength with the left TPJ was modulated by watching painful videos in the self-rating phase (vs. the implicit baseline) or the observation phase (vs. implicit baseline for a control analysis) in 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 when participants watched painful videos in the self-rating phase or in the observation phase, and a weight of 0 on all other regressors at the first level. Next, we determined regions whereby connectivity strength to the left TPJ was modulated by the weight given to observational prediction errors. To do so, we conducted second-

level covariate analyses in which the contrast image obtained for the first-level gPPI analysis was entered into a full-factorial design, with the individual *W1* parameters in the first session (see computational models for details) as the covariates. We entered the $W1_{pos}$ for the high empathy group and $W1_{neg}$ for the low empathy group. We tested the functional connectivity that was differentially associated with the *W*1 parameter in the high and low empathy group. Imaging results were determined in whole-brain analyses, using a combined voxel-level threshold of $P_{\text{uncorrected}} < 0.001$ and a FWE-corrected cluster-level threshold of $P < 0.05$.

The PPI analysis revealed that the individual *W1* parameters modulated the connectivity between the left TPJ and the left AI in the self-rating phase. In additional analyses, we aimed to specify the function of the AI that was identified in the PPI analysis. Using the identified AI region (**Figure 5**, upper panel) as a mask for small-volume-correction (FWE-SVC < 0.05), first we regressed the individual *W1* parameters against the the neural activity to the painful videos in the self-rating phase using a second-level regression. Second, we compared the neural activity tracked by the trial-by-trial empathy ratings between baseline session and generalization session between high and low empathy groups

Using MarsBaR (http://marsbar.sourceforge.net), we extracted beta values of identified clusters to visualize the correlations of the left TPJ with the left AI, and with vmPFC, and the weight parameters for high and low empathy groups respectively. Specifically, we plotted the connectivity strength for the left AI and

vmPFC identified by the PPI analysis (**Figure 5**). We also extracted the activation of the left AI (**Figure 5**) when watching others in pain in the baseline session to reveal the functional role of the AI.

Acknowledgement:

This work was supported by the German Research Foundation (GH, HE 4566/6-1; HE 4566/3-2 to GH), Swiss NSF (grants 100014_165884, 100019_176016 and IZKSZ3_162109 to PNT), the National Natural Science Foundation of China (project 32230043 to SH), the Ministry of Science and Technology of China (2019YFA0707103 to SH) and the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (E3CX1225 to YZ). The authors thank the High-performance Computing Platform of Peking University for assistance with this study.

Author contributions:

Conceptualization: YZ, GH

Methodology: YZ, SH, PK, PNT, GH

Investigation: YZ

Visualization: YZ

Supervision: PNT, GH

Writing—original draft: YZ, GH

Writing—review & editing: YZ, SH, PK, PNT, GH

Competing interests: Authors declare that they have no competing interests.

Data and code availability: The data and codes support the findings of the current study is available at

https://osf.io/n49y3/?view_only=60dd2d738b2646d6ada135aa1913f7dd

References:

1. M. L. Hoffman, Development of moral thought, feeling, and behavior. *Am. Psychol.* **34**, 958 (1979).

2. G. Thomas, G. R. Maio, Man, I feel like a woman: when and how gender-role motivation helps mind-reading. *J. Pers. Soc. Psychol.* **95**, 1165 (2008).

3. E. Weisz, D. C. Ong, R. W. Carlson, J. Zaki, Building empathy through motivation-based interventions. *Emotion*. **21**, 990–999 (2021).

4. E. C. Nook, D. C. Ong, S. A. Morelli, J. P. Mitchell, J. Zaki, Prosocial conformity: Prosocial norms generalize across behavior and empathy. *Pers. Soc. Psychol. Bull.* **42**, 1045–1062 (2016).

5. M. L. Hoffman, Interaction of affect and cognition in empathy. *Emot. Cogn. Behav.*, 103–131 (1984).

6. A. Bandura, Observational learning. *Int. Encycl. Commun.* (2008).

7. C. J. Burke, P. N. Tobler, M. Baddeley, W. Schultz, Neural mechanisms of observational learning. *Proc. Natl. Acad. Sci.* **107**, 14431–14436 (2010).

8. P. Kang, C. J. Burke, P. N. Tobler, G. Hein, Why We Learn Less from Observing Outgroups. *J. Neurosci.* **41**, 144–152 (2021).

9. S. Suzuki, E. L. S. Jensen, P. Bossaerts, J. P. O'Doherty, Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proc. Natl. Acad. Sci.* **113**, 3755–3760 (2016).

10. C. J. Charpentier, J. P. O'Doherty, The application of computational models to social neuroscience: promises and pitfalls. *Soc. Neurosci.* **13**, 637–647 (2018).

11. S. Suzuki, N. Harasawa, K. Ueno, J. L. Gardner, N. Ichinohe, M. Haruno, K. Cheng, H. Nakahara, Learning to simulate others' decisions. *Neuron*. **74**, 1125–1137 (2012).

12. L. Zhang, F. I. Kandil, K. Zhao, X. Fu, C. Lamm, C. C. Hilgetag, J. Glaescher, A causal role of the human left temporoparietal junction in computing social influence during goal-directed learning. *bioRxiv* (2022).

13. J. Debiec, A. Olsson, Social fear learning: from animal models to human function. *Trends Cogn. Sci.* **21**, 546–555 (2017).

14. S. Keum, H.-S. Shin, Neural basis of observational fear learning: a potential model of affective empathy. *Neuron*. **104**, 78–86 (2019).

15. A. Olsson, K. I. Nearing, E. A. Phelps, Learning fears by observing others: the neural systems of social fear transmission. *Soc. Cogn. Affect. Neurosci.* **2**, 3–11 (2007).

16. Zhang, Gläscher, A brain network supporting social influences in human decision-making. *Sci. Adv.*, 20 (2020).

17. G. Hein, J. B. Engelmann, M. C. Vollberg, P. N. Tobler, How learning shapes the empathic brain. *Proc. Natl. Acad. Sci.* **113**, 80–85 (2016).

18. M. H. Davis, Measuring individual differences in empathy: Evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* **44**, 113 (1983).

19. Y. Zhou, B. Lindström, A. Soutschek, P. Kang, P. N. Tobler, G. Hein, Learning from ingroup experiences changes intergroup impressions. *J. Neurosci.* **42**, 6931–6945 (2022).

20. J. Stöber, The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *Eur. J. Psychol. Assess.* **17**, 222 (2001).

21. A. Mehrabian, C. A. Stefl, Basic temperament components of loneliness, shyness, and conformity. *Soc. Behav. Personal. Int. J.* **23**, 253–263 (1995).

22. A. R. Wagner, R. A. Rescorla, Inhibition in Pavlovian conditioning: Application of a theory. *Inhib. Learn.*, 301–336 (1972).

23. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proc. Natl. Acad. Sci.* **111**, 12252–12257 (2014).

24. G.-J. Will, R. B. Rutledge, M. Moutoussis, R. J. Dolan, Neural and computational processes underlying dynamic changes in self-esteem. *Elife*. **6**, e28098 (2017).

25. Y. Fan, N. W. Duncan, M. de Greck, G. Northoff, Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neurosci. Biobehav. Rev.* **35**, 903–911 (2011).

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

27. C. Lamm, J. Decety, T. Singer, Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*. **54**, 2492–2502 (2011).

28. S. G. Shamay-Tsoory, The neural bases for empathy. *The Neuroscientist*. **17**, 18–24 (2011).

29. T. Singer, B. Seymour, J. O'doherty, H. Kaube, R. J. Dolan, C. D. Frith, Empathy for pain involves the affective but not sensory components of pain. *Science*. **303**, 1157–1162 (2004).

30. X. Xu, X. Zuo, X. Wang, S. Han, Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses. *J. Neurosci.* **29**, 8525–8529 (2009).

31. C. Sripada, M. Angstadt, D. Kessler, K. L. Phan, I. Liberzon, G. W. Evans, R. C. Welsh, P. Kim, J. E. Swain, Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks. *NeuroImage*. **89**, 110–121 (2014).

32. G. Lois, M. F. Gerchen, P. Kirsch, P. Kanske, S. Schönfelder, M. Wessa, Large-scale network functional interactions during distraction and reappraisal in remitted bipolar and unipolar patients. *Bipolar Disord.* **19**, 487–495 (2017).

33. J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, Functional network organization of the human brain. *Neuron*. **72**, 665–678 (2011).

34. A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*. **185**, 1124–1131 (1974).

35. C. D. Batson, A. A. Powell, "Altruism and prosocial behavior" in *Handbook of psychology: Personality and social psychology, Vol. 5.* (John Wiley & Sons, Inc., Hoboken, NJ, US, 2003), pp. 463–484.

36. J. Decety, I. B.-A. Bartal, F. Uzefovsky, A. Knafo-Noam, Empathy as a driver of prosocial behaviour: highly conserved neurobehavioural mechanisms across species. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150077 (2016).

37. A. Mahmoodi, H. Nili, D. Bang, C. Mehring, B. Bahrami, Distinct neurocomputational mechanisms support informational and socially normative conformity. *PLOS Biol.* **20**, e3001565 (2022).

38. J. Ou, Y. Wu, Y. Hu, X. Gao, H. Li, P. N. Tobler, Testosterone reduces generosity through cortical and subcortical mechanisms. *Proc. Natl. Acad. Sci.* **118**, e2021745118 (2021).

39. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*. **76**, 412–427 (2013).

40. L. Koban, M. Jepma, S. Geuter, T. D. Wager, What's in a word? How instructions, suggestions, and social information change pain and emotion. *Neurosci. Biobehav. Rev.* **81**, 29–42 (2017).

41. A. Mehrabian, C. A. Stefl, Basic temperament components of loneliness, shyness, and conformity. *Soc. Behav. Personal. Int. J.* **23**, 253–263 (1995).

42. F. Faul, E. Erdfelder, A. Buchner, A.-G. Lang, Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods*. **41**, 1149–1160 (2009).

43. G. Hein, J. B. Engelmann, P. N. Tobler, Pain relief provided by an outgroup member enhances analgesia. *Proc. R. Soc. B Biol. Sci.* **285**, 20180501 (2018).

44. V. A. Mathur, T. Harada, T. Lipke, J. Y. Chiao, Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage*. **51**, 1468–1475 (2010).

45. M. R. Jordan, D. Amir, P. Bloom, Are empathy and concern psychologically distinct? *Emotion*. **16**, 1107 (2016).

46. X. Han, S. Zhou, N. Fahoum, T. Wu, T. Gao, S. Shamay-Tsoory, M. J. Gelfand, X. Wu, S. Han, Cognitive and neural bases of decision-making causing civilian casualties during intergroup conflict. *Nat. Hum. Behav.* **5**, 1214–1225 (2021).

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

48. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron*. **69**, 1204–1215 (2011).

49. A. Huebner, C. Wang, A Note on Comparing Examinee Classification Methods for Cognitive Diagnosis Models. *Educ. Psychol. Meas.* **71**, 407–419 (2011).

50. C.-C. Ting, S. Palminteri, M. Lebreton, J. B. Engelmann, The elusive effects of incidental anxiety on reinforcement-learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **48**, 619 (2022).

51. J. Ashburner, K. J. Friston, Unified segmentation. *Neuroimage*. **26**, 839–851 (2005).

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

53. D. G. McLaren, M. L. Ries, G. Xu, S. C. Johnson, A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage*. **61**, 1277–1286 (2012).

54. K. C. Aberg, E. E. Kramer, S. Schwartz, Interplay between midbrain and dorsal anterior cingulate regions arbitrates lingering reward effects on memory encoding. *Nat. Commun.* **11**, 1829 (2020).

Supplementary Materials for

The social transmission of empathy relies on observational reinforcement learning

Yuqing Zhou *et al.*

*Corresponding author. Yuqing Zhou, zhouyq@psych.ac.cn
Grit Hein, Hein_G@ukw.de

This PDF file includes:

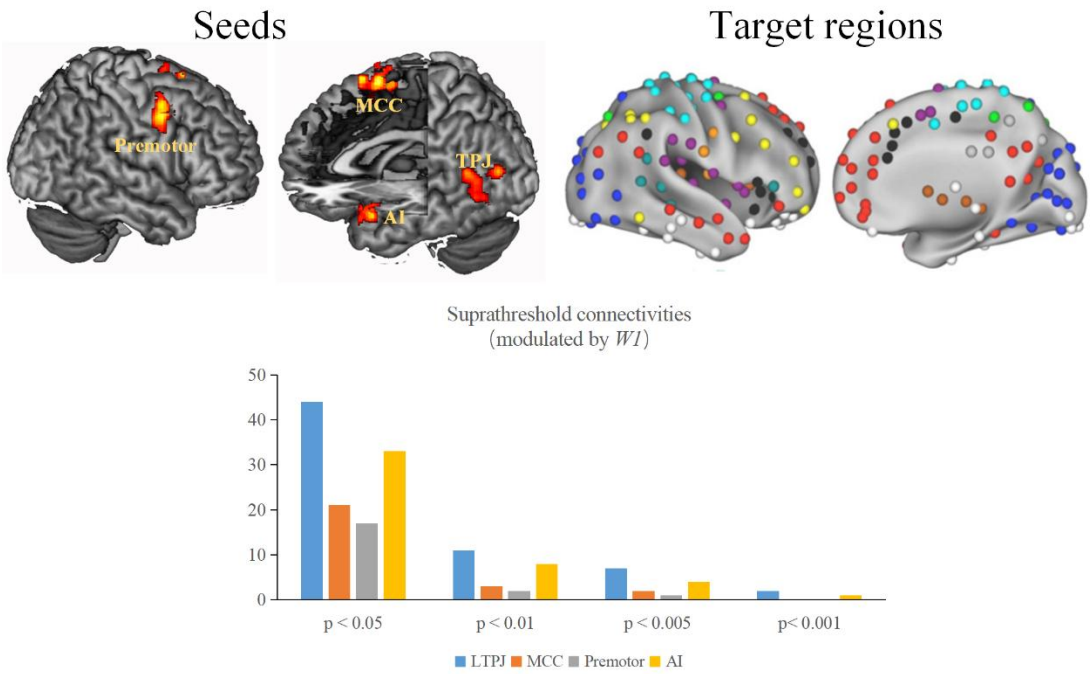

Figure S1. Multi-region PPI analysis. The upper panel shows the seeds for the multi-region PPI analysis and the target regions. The lower panel shows the number of connectivities modulated by the strength of observational learning (i.e., *W1* parameter), collapsed over high and low empathy conditions.

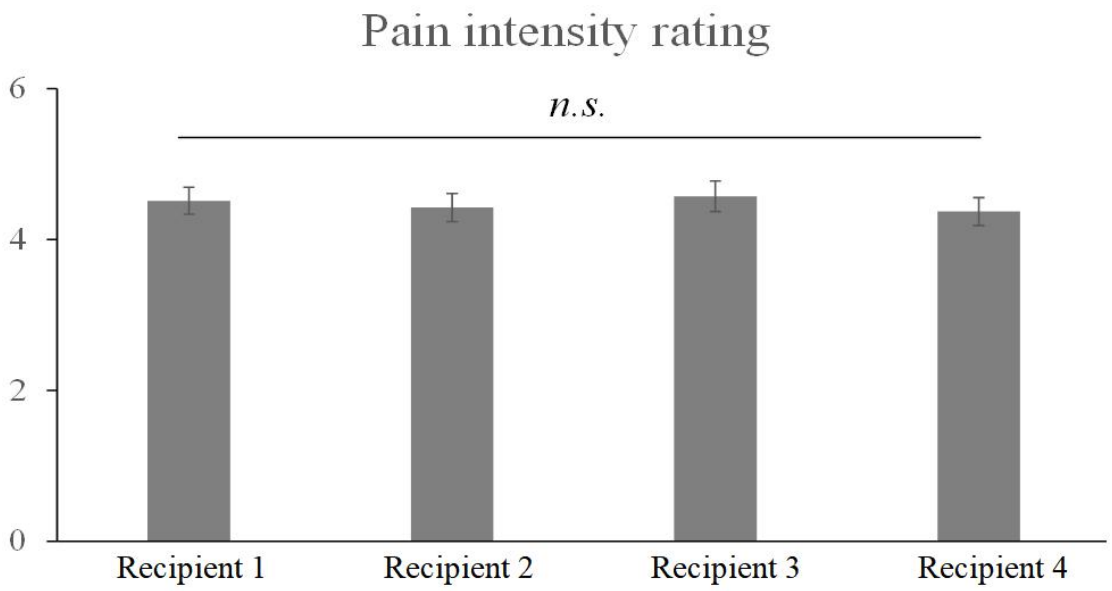

Figure S2. **Rating scores from an independent group of female participants (N = 37).** The pain intensity ratings were matched between recipients.

Table S1. Predicting change in empathy ratings from baseline ratings, social desirability, and conformity.

ΔEmpathy ratings			
Predictors	**β (SE)**	**T-value**	**P-value**
High empathy group			
Baseline rating	0.096 (0.063)	0.43	0.675
Conformity	0.145 (0.075)	0.66	0.515
Social desirability	-0.219 (0.321)	-1.04	0.308
Low empathy group			
Baseline rating	0.126 (0.122)	0.59	0.558
Conformity	-0.158 (0.205)	-0.74	0.469
Social desirability	0.091 (0.677)	0.41	0.683

Table S2. Means and standard deviations for computational model parameters of the winning model for high and low empathy condition in the fMRI study.

Computational Parameter	**High empathy**	**Low empathy**
$W1_{pos}$	0.46 (0.35)	-0.18 (0.49)
$W2_{pos}$	0.41 (0.34)	-0.03 (0.37)
$W1_{neg}$	-0.10 (0.38)	0.25 (0.39)
$W2_{neg}$	-0.18 (0.47)	0.31 (0.36)
Forgetting parameter (γ)	0.80 (0.27)	0.90 (0.14)
Learning rate (α)	0.46 (0.14)	0.47 (0.13)

Note: The parameters *W1* and *W2,* which capture the weight of the influence of observational prediction errors on changes in participants' empathy ratings in the first/second half of the observational learning session.

Table S3. Results of linear mixed models predicting the change of empathy ranting and the comparison between experiments.

Experiment	Regressors	Statistic value	
		χ^2	p
fMRI & Non-social control	Empathy Condition	1.24	0.26
	PE	27.70	< 0.001
	Experiment	0.51	0.48
	Empathy Condition × PE	0.55	0.46
	Empathy Condition × Experiment	0.0002	0.98
	PE × Experiment	5.34	0.021
	Empathy Condition × PE × Experiment	0.29	0.59
fMRI & Behavioral replication	Empathy Condition	0.22	0.64
	PE	53.62	< 0.001
	Experiment	0.0001	0.97
	Empathy Condition × PE	0.25	0.62
	Empathy Condition × Experiment	0.50	0.48
	PE × Experiment	0.55	0.46
	Empathy Condition × PE × Experiment	0.06	0.81

Table S4. Sample characteristics of the three studies.

Variables	Study 1	Study 2	Study 3	ANOVA	
	Mean ± SD	Mean ± SD	Mean ± SD	*F-value*	*p*
Age	21.1 ± 2.1	20.8 ± 2.4	20.7± 1.9	0.421	0.657
IRI	96.7 ± 10.3	97.2 ± 10.0	99.1 ± 11.9	0.768	0.466
Contagion	22.6 ± 3.8	22.4 ± 3.5	23.0 ± 3.7	0.416	0.661
Empathy	23.3± 4.5	22.2 ± 3.9	22.7 ± 4.3	0.872	0.420
SDS	10.1 ± 3.3	9.1 ± 2.7	9.2 ± 3.0	1.814	0.166
Conformity	53.9 ± 10.3	54.0 ± 12.7	51.5 ± 13.6	0.706	0.495

IRI = Interpersonal Reactivity Index; SDS = Social Desirability Scale; Contagion = Behavioral Contagion; Empathy = Empathy Index.

Table S5. Results of the questionnaire and behavioral measures within studies.

	Variables	High empathy group	Low empathy group	T test	
		Mean ± SD	Mean ± SD	*T-value*	*P*
Study 1	Age	21.1 ± 2.3	20.8 ± 1.8	-0.066	0.948
	IRI	97.6 ± 11.3	95.6 ± 9.3	0.698	0.488
	Contagion	23.4 ± 4.2	21.7 ± 3.4	1.572	0.122
	Empathy	24.3 ± 5.0	22.3 ± 3.7	1.630	0.109
	SDS	9.8 ± 3.7	10.5 ± 2.8	-0.720	0.475
	Conformity	56.5 ± 10.9	51.3 ± 9.1	1.862	0.068
Study 2	Age	20.7 ± 2.2	20.9 ± 2.6	-0.310	0.758
	IRI	97.7 ± 10.6	96.8 ± 9.6	0.356	0.723
	Contagion	22.4 ± 3.7	22.3 ± 3.4	0.118	0.907
	Empathy	22.1 ± 4.2	22.3 ± 3.8	-0.185	0.854
	SDS	8.9 ± 2.7	9.3 ± 2.8	-0.570	0.571
	Conformity	51.3 ± 12.9	56. 4 ± 12.3	-1.516	0.135
Study 3	Age	21.0 ± 1.9	20.5 ± 1.9	0.962	0.341
	IRI	99.6 ± 10.9	98.7 ± 13.0	0.280	0.780
	Contagion	22.4 ± 2.8	23.6 ± 4.4	-1.206	0.234
	Empathy	22.2 ± 2.8	23.0 ± 5.4	-0.658	0.514
	SDS	8.8 ± 3.1	9.6 ± 3.0	-0.981	0.331
	Conformity	53.0 ± 15.2	50.0 ± 12.0	0.805	0.425

IRI = Interpersonal Reactivity Index; SDS = Social Desirability Scale; Contagion = Behavioral Contagion; Empathy = Empathy Index.