Behavioral/Cognitive

Neural Substrates of Intention–Consequence Integration and Its Impact on Reactive Punishment in Interpersonal Transgression

Hongbo Yu,[1] Jia Li,[1] and Xiaolin Zhou[1,2,3]

[1]Center for Brain and Cognitive Sciences and Department of Psychology, Peking University, Beijing 100871, People's Republic of China, [2]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, People's Republic of China, and [3]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, People's Republic of China

When evaluating interpersonal transgressions, people take into account both the consequential damage and the intention of the agent. The intention and consequence, however, do not always match, as is the case with accidents and failed attempts. We combined an interactive game and functional MRI to investigate the neural substrates underlying the processing of intention and consequence, and its bearing on reactive punishment. The participant interacted with anonymous partners, who decided to deliver pain stimulation either to himself/herself or to the participant to earn a monetary reward. In some cases, the decision was reversed by the computer. After pain delivery, the partner's intention was revealed. Unbeknownst to the partner, the participant was then allowed to punish the partner by reducing his/her monetary reward. Behaviorally, the punishment was lower in the accidental condition (unintended harm relative to intended harm) but higher in the failed-attempt condition (unintended no-harm relative to intended no-harm). Neurally, the left amygdala/hippocampus was activated in the conditions with blameworthy intention (i.e., intentional harm and failed attempt). The accidental (relative to intentional) harm activated the right temporoparietal junction (TPJ) and the anterior inferior frontal gyrus (IFG), while the failed attempt (relative to genuine no-harm) activated the anterior insula (AI) and the posterior IFG. Effective connectivity analysis revealed that in the unintentional conditions (i.e., accidental and failed attempt) the IFG received input from the TPJ and AI, and sent regulatory signals to the amygdala. These findings demonstrate that the processing of intention may gate the emotional responses to transgression and regulate subsequent reactive punishment.

Key words: **amygdala; dynamic causal modeling; insula; intention; punishment; temporoparietal junction**

Introduction

"justice does not only weigh what is done but what is in the heart" (Heloise, quoted from Levitan, 2007, p. 59). This quotation highlights a characteristic feature of human moral cognition and practice: when evaluating the actions of others and forming reactive attitudes or behaviors, humans take into account not only the consequences of the action, but also the intention of the agent. This is crucial especially when the consequences are not intended: whether it is everyday social interaction or in a federal court, an accidental harm is more likely to be forgiven, while a failed attempt is more likely to be condemned (Buckholtz and Marois, 2012).

Neuroscience research has begun to reveal the brain mechanisms underlying reactive punishment (Seymour et al., 2007). Some of these studies used interpersonal games (e.g., the Ultimatum Game) to elicit interpersonal transgression and second-person reactive punishment (Sanfey et al., 2003; Liljeholm et al., 2014), while others asked participants to evaluate the blameworthiness of a protagonist in a scenario as an "impartial" third party (Young et al., 2007; Buckholtz et al., 2008; Treadway et al., 2014). These studies have consistently identified several brain structures and processes essential to forming proper responses in real or imagined transgressions, such as the temporoparietal junction (TPJ) in inferring intention, and the amygdala and anterior insula (AI) in experiencing a negative effect. However, each of these approaches has certain limitations, rendering them inadequate to address how the processing of intention and emotion-laden consequence interact to initiate moral evaluation and punishment. In economic games, the transgression is always intentional, and the desire to punish is confounded with the consideration of self-interest, as the punishment is costly for the victim (but see Liljeholm et al., 2014). The scenario-based approach has the strength of independently manipulating the intention and consequence of transgression, and thus succeeds in elucidating the neural substrates of judicial decision making, but evaluating the

Received Aug. 23, 2014; revised Feb. 10, 2015; accepted Feb. 13, 2015.

Author contributions: H.Y., J.L., and X.Z. designed research; H.Y. and J.L. performed research; H.Y. and J.L. analyzed data; H.Y. and X.Z. wrote the paper.

This work was supported by the Natural Science Foundation of China (Grants 91232708, 31170972) and the National Basic Research Program of China (973 Program: 2010CB833904). We thank Xirui Cui, Huihui Zheng, Yunlu Yin, and Dr. Li Hu for their assistance in data collection and data analysis; and Dr. Micheal Treadway and two anonymous reviewers for their helpful comments on previous versions of this manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Xiaolin Zhou, Department of Psychology, Peking University, 5 Yiheyuan Road, Beijing 100871, People's Republic of China. E-mail: xz104@pku.edu.cn.

DOI:10.1523/JNEUROSCI.3536-14.2015

 0270-6474/15/354917-09$15.00/0

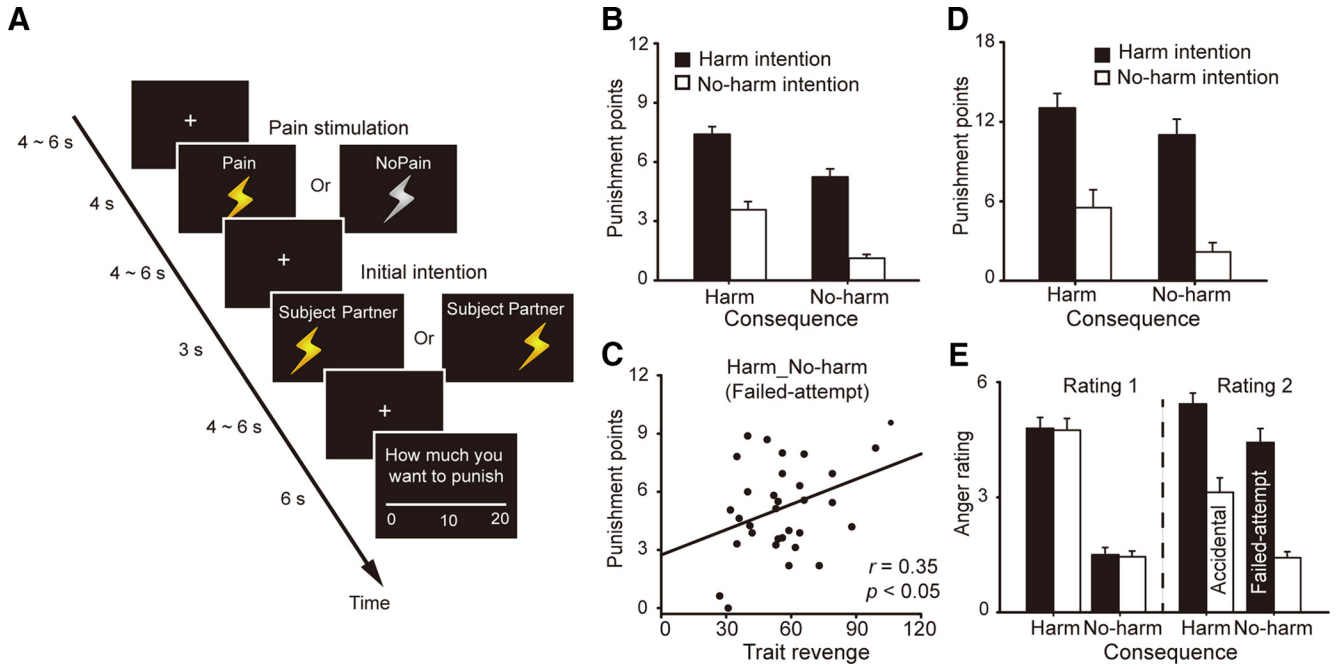

Figure 1. Experimental task design and behavioral results. ***A***, The timing of the fMRI experiment. At the beginning of each round, the participant either received a pain stimulation (Harm), indicated by a yellow lightning cue, or received no stimulation (No-harm), indicated by a gray lightning cue. This was the result of the partner's decision or was caused by an interfering computer program, which reversed the partner's initial decision. After a variable interval, the partner's initial decision was revealed (i.e., delivering the pain to the participant or to the partner himself/herself). The initial decision could be blameworthy (i.e., delivering stimulation to the participant; Harm) or innocent (i.e., delivering stimulation to the partner himself/herself; No-harm). After another variable interval, the participant was asked to punish the partner by subtracting the money points that the partner earned in that round, with the knowledge that the partner was unaware of this punishment. The punishment cost the participant nothing. Note that in the behavioral experiment, the participant was asked to rate their anger before and after the revelation of the partner's intention. The participant had 4 s for each anger rating. ***B***, Punishment amount in the four conditions in the fMRI experiment. ***C***, In the fMRI experiment, the punishment amount in the No-harm_Harm (i.e., failed-attempt) condition positively correlated with the trait revenge score. ***D***, ***E***, Punishment amount and anger ratings in the four conditions in the behavioral experiment. Error bars represent the SE.

blameworthiness of the protagonist as an "impartial" third party could differ fundamentally from experiencing a transgression directly (Schilbach et al., 2013). Moreover, there has been no work on how revenge-related personalities could modulate the neural processing of intention and its contribution to punishment behavior, although personality does play an important role in forming responses to transgression (Stuckless and Goranson, 1992).

We sought to elucidate the neural processing of intention in interpersonal transgression and how this processing regulates the affective system in response to interpersonal harm. To this end, we asked the participant to interact with anonymous partners (confederates), who could deliberately or accidentally choose to physically harm the participant or himself/herself to earn a monetary reward. The participant then had the opportunity to punish the partner without the partner knowing. We addressed the following three questions: (1) what are the differences between the neural processing of an innocent and a blameworthy intention?; (2) how do personality traits (e.g., revenge and forgiveness) modulate the neural processing of intention?; and (3) what are the neural circuits through which the processing of intention, innocent or blameworthy, regulates the emotional responses to transgression and the formation of punishment behavior?

Materials and Methods

Participants

Forty-five healthy, right-handed undergraduate and graduate students took part in the fMRI scanning experiment. Because of excessive head movements, 12 were excluded from fMRI data analysis, leaving 33 participants (mean age, 22.0 years; age range, 18–25 years; 16 females) for data analysis. Another 16 undergraduate students (mean age, 19.9 years; age range, 18–22 years; 9 females) took part in the behavioral experiment. None of the participants reported any history of psychiatric, neurological, or cognitive disorders. Informed written consent was obtained from each participant before the experiments. The study was performed in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Department of Psychology, Peking University.

Procedure

fMRI experiment. Each participant came to the scanning room individually. Upon arrival he/she met three confederates and was told that they would later play together in an interactive game via intranet and that the participant undergoing MRI scanning (i.e., the participant) would play a different role in the game than the others (i.e., the confederates). At least one confederate was of the same sex as the participant, and at least one was of the opposite sex. An intraepidermal needle electrode was attached to the left wrist of the participant for cutaneous electrical stimulation (Inui et al., 2002). The intensity of pain stimulation was calibrated to a subjective pain rating of "7" on a scale of 0 ("no feeling") to 10 ("unbearable").

In the scanning session, the participant performed the task described in Figure 1*A*. The participant was told that in each trial, he or she would be paired with a partner (randomly chosen from among the three confederates). To avoid any influence of the previous encounter on the current trial, we explicitly told the participant that the partner in each trial may or may not be the same partner as in the last trial. This partner then decided between whether to shock the participant or to shock himself/herself, and was rewarded with 20 monetary points after doing so. Critically, the decision by the partner could be reversed by the computer system in certain trials (40% for both the "pain" and "no pain" trials, unknown to the participants). At the beginning of each trial, an image of lightning informed the participant whether he/she would receive a shock (if the figure was yellow) or not (if the figure was gray), after which the shock was delivered. After a jittered interval, the partner's initial decision (i.e., intention) was revealed. After another jittered interval, the partici-

Table 1. Behavioral modeling for the fMRI experiment

Models	AIC	Predictors	*t* value	*p* value
1	9806	$Consequence_n$	8.50	1.04×10^{-9}
		$Intention_n$	9.18	1.75×10^{-10}
		$Consequence_{n-1}$	1.24	0.22
		$Intention_{n-1}$	0.96	0.34
2	9743	$Consequence_n$	8.53	9.42×10^{-10}
		$Intention_n$	9.26	1.42×10^{-10}
		$Consequence_{n-1}$	1.21	0.23
		$Intention_{n-1}$	0.87	0.39
		Consequence × $intention_{n-1}$	−0.15	0.88
3	**9486**	**$Consequence_n$**	**8.49**	**1.06×10^{-9}**
		$Intention_n$	**9.19**	**1.73×10^{-10}**
		$Consequence_{n-1}$	**−1.18**	**0.25**
		$Intention_{n-1}$	**0.92**	**0.37**
		Consequence × $intention_n$	**1.01**	**0.32**
4	9521	$Consequence_n$	8.56	8.89×10^{-10}
		$Intention_n$	9.26	1.43×10^{-10}
		$Consequence_{n-1}$	−1.20	0.24
		$Intention_{n-1}$	0.97	0.34
		Consequence × $intention_n$	1.57	0.13
		Consequence × $intention_{n-1}$	1.53	0.14

The model with the smaller AIC is better. The winning model (model 3) is indicated in bold.

Table 2. Behavioral modeling for the behavioral experiment

Models	AIC	Predictors	*t* value	*p* value
1	2668	Consequence	−0.14	0.44
		Intention	3.73	1.01×10^{-3}
		Rating 1	0.14	0.44
		Rating 2	7.54	8.86×10^{-7}
2	**2627**	**Consequence**	**−0.68**	**0.51**
		Intention	**3.48**	**3.37×10^{-3}**
		Rating 1	**1.27**	**0.22**
		Rating 2	**5.33**	**8.47×10^{-5}**
		Consequence × intention	**−0.53**	**0.60**
3	2824	Consequence	3.29	2.20×10^{-3}
		Intention	4.75	2.60×10^{-4}
		Rating 2–rating 1	4.24	7.20×10^{-4}
4	2784	Consequence	3.15	6.63×10^{-3}
		Intention	4.50	4.21×10^{-4}
		Rating 2–rating 1	2.30	0.04
		Consequence × intention	0.72	0.48

The model with the smaller AIC is better. The winning model (model 2) is indicated in bold.

pant was given an opportunity to punish the partner by reducing the partner's money points (0–20). This served as an on-line measure of reactive attitudes (revenge or forgiveness). Therefore, we had a two (intention: harm vs no-harm) by two (consequence: harm vs no-harm) design with the following four experimental conditions (intention_consequence): harm_harm (intentional harm); no-harm_harm (accidental); harm_no-harm (failed attempt); and no-harm_no-harm (genuine no-harm).

Behavioral experiment. To make sure that the aggressive emotional reaction was indeed modulated by a partner's intention, we performed a behavioral experiment in which we directly asked the participant to evaluate his/her angry feeling before and after they knew the partner's initial intention. The procedure and setup of the behavioral experiment was similar to the fMRI experiment, except that an emotion (anger) self-report was inserted before and after the revelation of the partner's intention. The self-report was on a 7-point Likert scale (0 = not angry at all; 6 = extremely angry). At the end of each trial, the participant had an opportunity to punish the partner by reducing the partner's money points (0–20), which was identical to the fMRI experiment.

Analysis of the behavioral data

fMRI experiment. We fit a set of linear regression models for the trial-wise punishment amount (Table 1). Predictors included consequence (harm: 1; no_harm: 0) and intention (harm: 1; no_harm: 0) of the current trial and of the last trial. We ran four models, each with different interaction terms. One model included no interaction term (model 1). Two models included either the interaction between consequence and intention for the last trial (model 2) or for the current trial (model 3). The last model (model 4) included both of the two interactions. Model goodness was assessed using the Akaike information criterion (AIC; Burnham and Anderson, 2004), which reflects both model fitness and complexity. A lower AIC indicates that a model is considered to be closer to the truth. Parameters were estimated based on the best model (lowest AIC).

Behavioral experiment. Similar to the fMRI experiment, we fit a set of four models to account for the trial-wise variance in punishment amount. Given that no learning effect was observed in the fMRI experiment (see Results), we included only predictors corresponding to the current trial. We were interested in the following four variables: consequence (harm: 1; no_harm: 0), intention (harm: 1; no_harm: 0), and the two anger ratings (rating 1 and rating 2). In two of these models, rating 1 and rating 2 were entered as separate regressors, while in the other two, only the change of anger (rating 1 to rating 2) was entered as a regressor. Orthogonal to this categorization, the other dimension was the inclusion of the interaction between consequence and intention (Table 2). In addition to the amount of punishment, we also examine how anger ratings are modulated by intention using standard paired-sample *t* tests, comparing the first and second ratings in the accidental and failed-attempt conditions, respectively.

fMRI data acquisition

Images were acquired using a Siemens 3.0 tesla Trio scanner with a standard head coil at the Key Laboratory of Cognition and Personality (Ministry of Education) of Southwest University, People's Republic of China. T2*-weighted functional images were acquired in 36 axial slices parallel to the anterior commissural–posterior commissural line with no interslice gap, affording full-brain coverage. Images were acquired using an EPI pulse sequence (TR = 2200 ms; TE = 30 ms; flip angle = 90°; FOV = 192 mm × 192 mm; slice thickness = 3 mm).

FMRI data analysis

Analysis on BOLD activation. Image preprocessing and analysis were conducted with the Statistical Parametric Mapping software SPM8 (Wellcome Trust Department of Cognitive Neurology, London, UK). Images were slice-time corrected, motion corrected, normalized to MNI (Montreal Neurological Institute) space, spatially smoothed using an 8 mm FWHM Gaussian filter, and temporally filtered using a high-pass filter with 1/128 Hz cutoff frequency. In the first-level (within-participant) statistical analysis, we modeled the pain delivery, the partner's initial intention, and punishment response as separate regressors in a GLM. The initial intention event was further divided into four regressors corresponding to the four conditions. The pain delivery events were separately modeled using two regressors, one corresponding to pain trials and the other corresponding to no-pain trials. The 6 rigid body parameters, their squares, and the derivatives of these 12 parameters (altogether, 24 parameters) were included to account for head motion artifacts. In the second-level (group-level) analysis the four contrast maps corresponding to the four conditions from each participant were fed into a flexible factorial design. We defined two contrasts corresponding to the following two types of intentions controlling consequence (intention_consequence): no-harm_harm > harm_harm (innocent intention); and harm_no-harm > no-harm_no-harm (blameworthy intention). Additionally, we defined a main effect contrast corresponding to the valence of the intention (harm_harm + harm_no-harm > no-harm_harm + no-harm_no-harm), which can be viewed as the contrast between the conditions in which the participants were prompted to punish and the conditions in which the participants withheld punishment. The statistical threshold was set at ≥20 voxels, each significant at $p < 0.001$ (uncorrected).

To test the functional dissociation within the inferior frontal gyrus (IFG; see Results), we extracted the parameter estimates from five voxels equally distributed along the anterior IFG (aIFG)-to-posterior IFG (pIFG) axis (MNI coordinates: [−47, 44, −11], [−47, 41, −8], [−47, 38,

−5], [−47, 35, −2], and [−47, 32, 1]) and computed for each voxel the intention–consequence mismatch effect (intention–consequence mismatch condition − intention–consequence match condition) for the actual harm and the actual no-harm consequences (i.e., accidental vs intentional harm and failed-attempt vs genuine no-harm). We then performed a two-by-five repeated-measures ANOVA to formally test whether the mismatch effect was modulated by rostrocaudal position (McNamee et al., 2013). To test the correlation between personality traits and intention-related brain activations, we extracted the parameter estimates from the right AI (rAI) and the right TPJ (rTPJ), and computed the Pearson correlation between the parameter estimates and the trait forgiveness score (Rye et al., 2001) and the trait revenge score (Stuckless and Goranson, 1992).

Effective connectivity analysis. To address our third question, we investigated the effective connectivity underlying intention processing in the interpersonal transgressive context. Here we used the dynamic causal modeling (DCM), which allowed us to build and compare neural connectivity models based on competing hypotheses concerning the structure and dynamics of the network. Dynamic causal modeling is defined by the following three sets of parameters (Friston et al., 2003): (1) the "intrinsic" connectivity represents the latent connectivity among brain regions in the absence of experimental perturbations; (2) the "modulatory" connectivity represents the changes imposed on the intrinsic connectivity caused by experimental perturbations; and (3) "input" represents the driving influence on brain regions by external perturbations.

Our dynamic causal modeling was designed to answer a specific question, namely, whether the coupling between regions responsible for the (counterfactual) intention processing influenced the amygdala or was influenced by the amygdala. Volumes of interest (VOI) were extracted based on the group-level analyses described above (no-harm_harm > harm_harm; harm_no-harm > no-harm_no-harm; and harm_harm + harm_no-harm > no-harm_harm + no-harm_no-harm). The whole-brain analysis allowed us to establish the peak voxels within the amygdala, aIFG, pIFG, rAI, and rTPJ at a group level for the contrasts of interest; 4 mm spherical VOIs were extracted and adjusted for the session average. Forty-two models were constructed corresponding to the intention revelation stage. The structure of these models varied in the following three orthogonal dimensions: direction of information flow (from amygdala, to amygdala, or bilateral); input regions (the intention-sensitive regions, the amygdala, or both); and direct intrinsic connectivity between the rAI and the amygdala. Figure 4 summarizes the structures of the models. Note that for the sake of simplicity, models with direct intrinsic connectivity between the rAI and the amygdala were not shown in the figure. Each of these alternative structures formed a model family, which contained three individual models differing in the modulatory effect. Specifically, in one of these three models, the modulatory effects of external perturbations (accidental and failed attempt) were set on the intrinsic connectivities between the intention-sensitive regions (rAI and rTPJ) and the IFG; in another, the modulatory effects were set on the intrinsic connectivities between the IFG and the amygdala; in the third, both of the above modulatory effects were included.

These models were compared using Bayesian model selection (BMS), which uses a Bayesian framework to calculate the "model evidence" of each model. The model evidence represents the trade-off between model simplicity and accuracy (Penny et al., 2004). Here, BMS was implemented using a random-effects analysis (i.e., assuming that the model structure might vary across participants) that is robust to the presence of outliers (Stephan et al., 2009). When comparing model families, all models within a family were averaged using Bayesian model averaging, and the exceedance probabilities were calculated for each model family

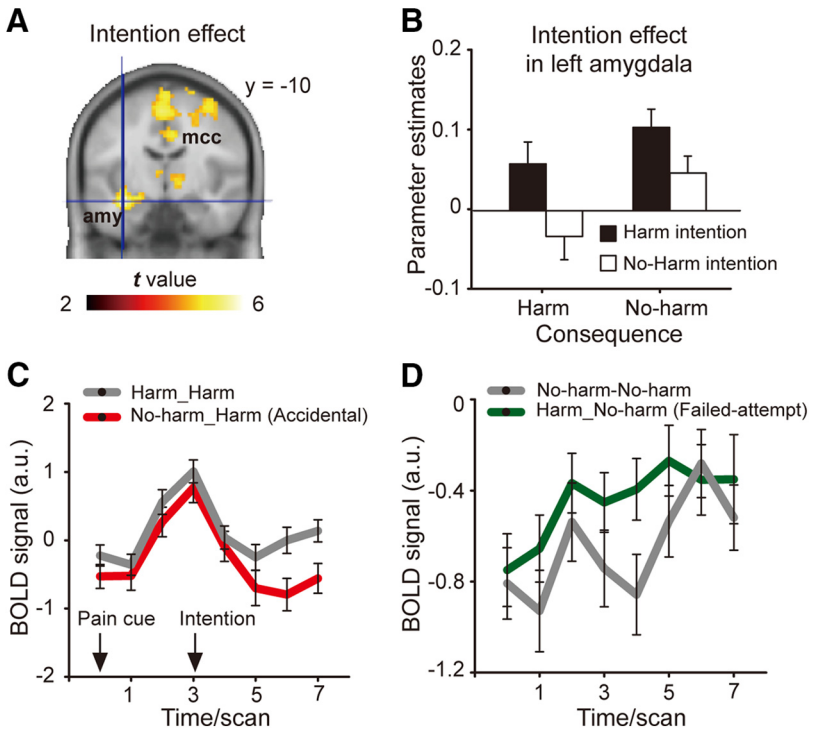

Figure 2. Brain activations revealed by the intention effect. ***A***, Statistical parametric map displaying the left amygdala activation (blameworthy > innocent). ***B***, Parameter estimates of the left amygdala. ***C***, ***D***, Time course extracted from the left amygdala cluster. Error bars represent the SE. amy, Amygdala; mcc, middle cingulate cortex.

Table 3. Brain activations revealed by the whole-brain contrasts ($p < 0.001$ uncorrected at voxel level, cluster contains >20 voxels)

Regions (BA)	Hemisphere	Maximum *t* value	Cluster size (voxels)	MNI coordinates *x*	*y*	*z*
Intention effect						
LOFC (47)	L	4.72	56	−30	32	−20
MCC (24)	R	5.18	1035	6	2	31
Precentral (6)	R	4.80	209	45	−4	46
Amygdala/hipp (20)	L	6.01	191	−36	−10	−17
Thalamus	L	4.35	132	−3	−16	10
MTG (21)	R	4.63	89	60	−46	1
No-harm_Harm > Harm_Harm						
aIFG (47)	L	4.82	90	−45	44	−11
IFG *tri.* (45)	L	4.00	34	−54	23	1
MFG (9)	R	4.37	78	36	17	43
	L	3.65	24	−39	11	55
SMA (6)	R	4.06	46	−6	14	55
TPJ (39)	R	4.75	172	60	−58	40
Harm_No-harm > No-harm_No-harm						
pIFG (47)	L	4.94	355	−48	32	1
IFG *tri.* (48)	L	4.44	177	−45	26	28
MFG (6)	L	4.61	165	−45	5	55
AI (47)	R	3.98	47	33	32	1
SMA (8)	L/R	4.61	215	0	20	61
Midbrain	R	4.64	228	12	−10	−2

Whole-brain contrasts were $p < 0.001$ (uncorrected at the voxel level). The cluster contains >20 voxels. IFG *tri.*, Inferior frontal gyrus *pars triangularis*; MFG, middle frontal gyrus; LOFC, lateral orbitofrontal cortex; hipp, hippocampus; MTG, middle temporal gyrus; MCC, middle cingulate cortex; hipp, hippocampus; BA, Brodmann area; L, left; R, right.

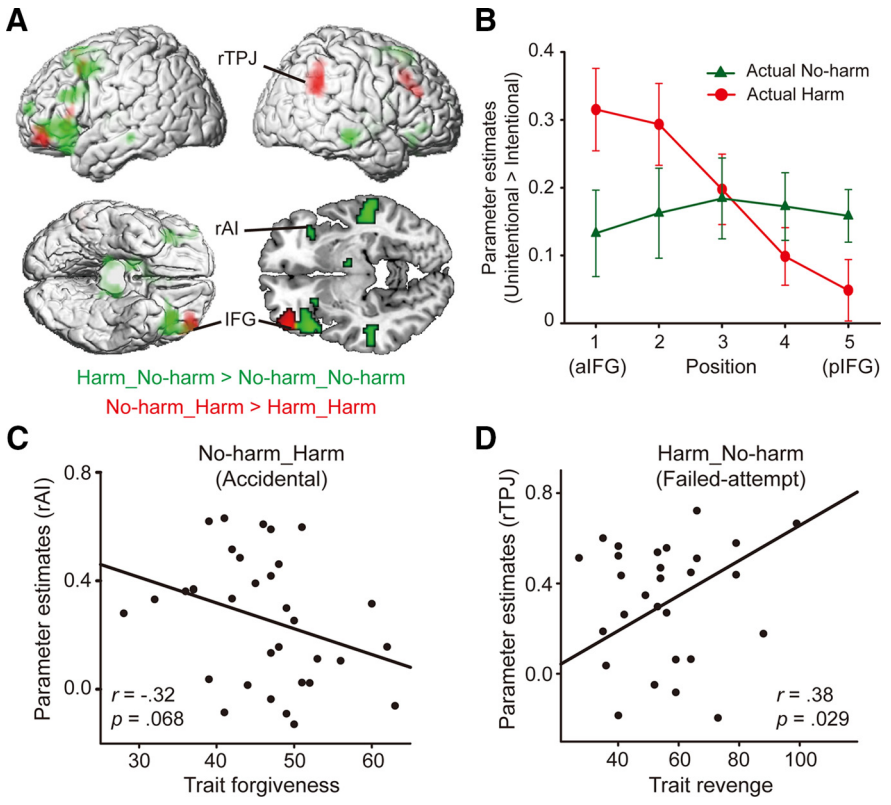

Figure 3. Brain activations associated with the processing of innocent and blameworthy intentions. ***A***, The activations of the whole-brain contrast Harm_No-harm > No-harm_No-harm is shown in green, and the contrast No-harm_Harm > Harm_Harm is shown in red. ***B***, Plot of the mismatch effect versus rostrocaudal location. Error bars indicate the SE. ***C,D***, The activation in the rAI in the No-harm_Harm (i.e., accidental) condition correlated negatively with the trait forgiveness scores ($r = -0.32, p = 0.068$), whereas the activation in the rTPJ in the Harm_No-harm (i.e., failed-attempt) condition correlated positively with the trait revenge scores ($r = 0.38, p = 0.029$).

(Penny et al., 2010). Model parameters were estimated based on the winning model.

Results

Behavioral results

fMRI experiment

As can be seen from Table 1, the model with the interaction between consequence and intention of the current trial (model 3) has the lowest AIC. Parameters that were estimated based on this model showed that both the consequence and the intention were significantly predictive of the punishment amount. Specifically, both a harmful intention and a harmful consequence increased the punishment amount (Fig. 1*B*). It is clear from Figure 1*B* that when the consequence and the partner's intention mismatched, as is the case with accidental harm (no-harm_harm) and failed attempt (harm_no-harm), the participants increased or reduced their punishment behavior according to the partner's initial intention (Buckholtz et al., 2008; Treadway et al., 2014). No influence of the previous trial on the current punishment amount was observed, indicating that no learning occurred and that the participant treated each trial independently (Table 1). Moreover, the punishment amount in the failed-attempt condition positively correlated with participants' trait revenge scores ($r = 0.35$, $p < 0.05$), suggesting that individuals who are more likely to desire revenge were also more likely to deploy a severe punishment on unsuccessful transgressors (Fig. 1*C*).

Behavioral experiment

To directly test our hypothesis that individuals can regulate their emotional reactions to interpersonal transgression according to the intention underlying others' behavior, we compared the first (before intention revelation) and second (after intention revelation) anger ratings for the accidental and failed-attempt conditions (Fig. 1*E*). Compared with the first rating, the second rating was significantly lower (i.e., downregulated) for the accidental condition ($t_{(15)} = 6.35, p < 0.001$) and was significantly higher (i.e., upregulated) for the failed-attempt condition ($t_{(15)} = 7.01, p < 0.001$). We then tested how such changes in emotional responses influence participants' punishment (Fig. 1*D*). Among our linear regression models linking the anger rating and punishment, the best model was the one including the two anger ratings as separate predictors and the interaction term between consequence and intention (model 2; Table 2). Parameters that were estimated based on this model showed that the intention and the second anger rating (after intention revelation) were independently predictive of the punishment amount. Specifically, the participant punished more when the partner's intention was to harm relative to not to harm, and when the second anger rating was high relative to low.

Whole-brain analysis of fMRI: intention stage

Consistent with previous findings (Treadway et al., 2014), the contrast corresponding to intention (blameworthy > innocent), harm_harm + harm_no-harm > no-harm_harm + no-harm_no-harm, revealed activations in the left amygdala extending to the hippocampus and the surrounding medial temporal cortex (Fig. 2*A*); other regions identified by this contrast included the middle cingulate cortex, the thalamus, and the right middle temporal gyrus (Table 3). The amygdala/hippocampus activation survived the whole-brain voxel-level FWE correction. For a comparison, we conducted a small volume correction analysis (8 mm radius) around the peak coordinates of the left amygdala reported in the study by Treadway et al. (2014). We found a significant activation cluster within the small volume that survived voxel-level FWE correction ($p_{FWE} < 0.001$, 59 voxels; peak coordinates in the MNI space: $x = -36, y = -10, z = -17$). Moreover, we extracted the regional parameter estimates from 27 voxels around the left amygdala peak coordinates reported in the study by Treadway et al. (2014) and conducted an intention-by-consequence repeated-measures ANOVA. We found a significant main effect of intention ($F_{(1,32)} = 10.34, p < 0.005$). This pattern mirrored both our behavioral pattern and our amygdala activation pattern. These findings indicated that the current finding concerning the left amygdala was similar to the one found in the study by Treadway et al. (2014). As Figure 2*B* showed, the pattern of activation in the amygdala was similar to the pattern of behavioral response: downregulated in the accidental condition and upregulated in the failed-attempt condition. The time course showed that such regulations occurred approximately one or two scans (~4 s) after the presentation of intention (Fig. 2*C*,*D*).

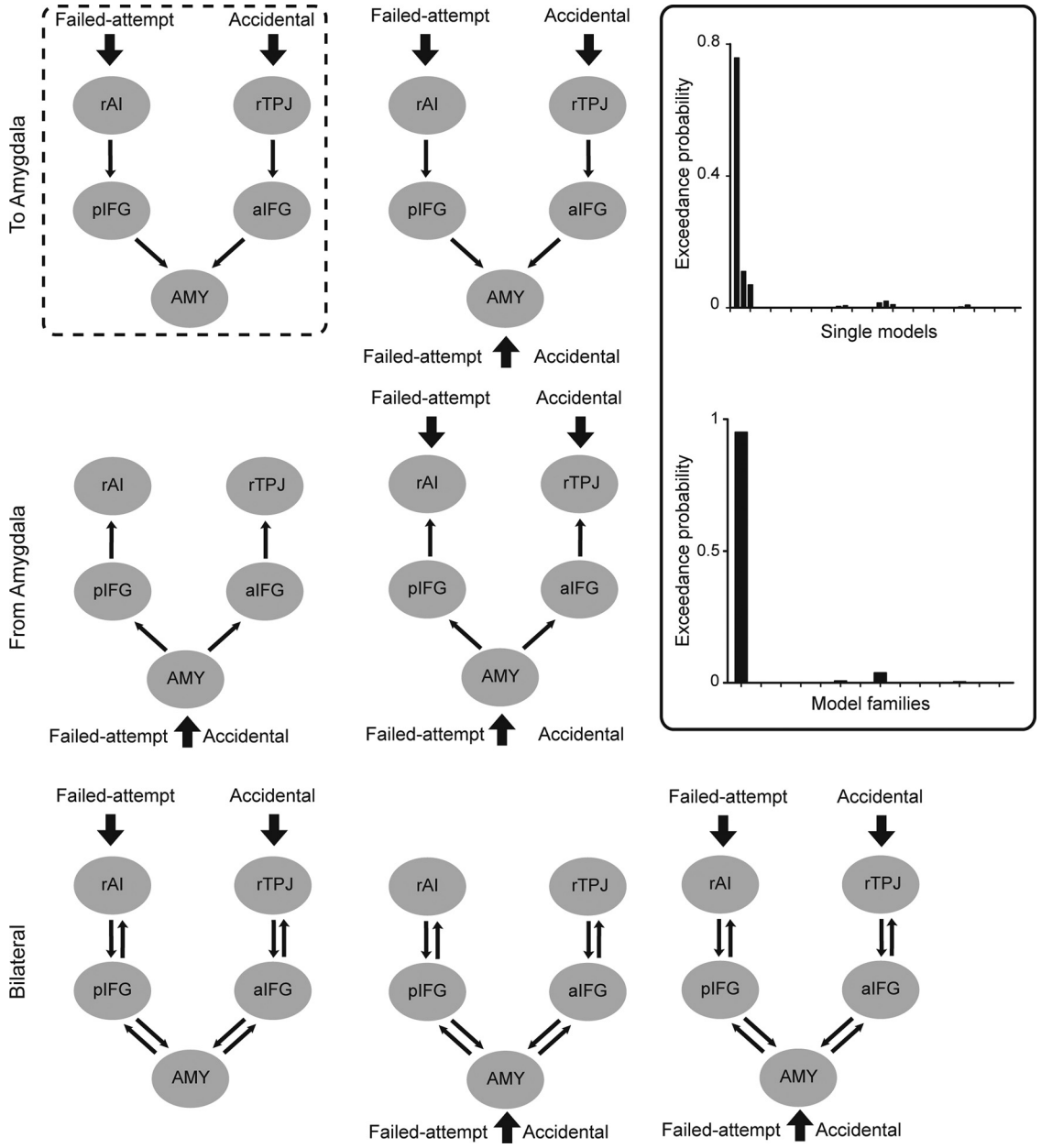

Figure 4. Model structures in the DCM analysis. Shown here are seven model families without direct intrinsic connectivity between the rAI and the amygdala (AMY). Another seven model families are similar to the ones shown here, but all have the direct intrinsic connectivity between the rAI and the amygdala. The structure of the winning family is highlighted in the dashed-line box. In the solid-line box are the exceedance probabilities of individual models and model families. rAI, right anterior insula; aIFG, anterior inferior frontal gyrus; pIFG, posterior inferior frontal gyrus; rTPJ, right temporoparietal junction.

We then examined the integration of consequence and intention separately for the innocent (no-harm_harm > harm_harm) and blameworthy intention (harm_no-harm > no-harm_no-harm). Relative to the harm_harm condition, the no-harm_harm (i.e., accidental) activated the anterior inferior frontal cortex, the supplementary motor area (SMA), and the rTPJ in adjacent to the right angular gyrus (Fig. 3*A*, red). Relative to the no-harm_no-harm condition, the harm_no-harm (i.e., failed attempt) activated the left posterior IFC, the right AI, the SMA, and the thalamus (Fig. 3*A*, green). As can be seen from Figure 3*A*, the processing of accidental and failed attempts exhibited a rostrocaudal dissociation in the left IFG. To formally test this dissociation, we extracted the parameter estimates from five voxels between the aIFG and the pIFG, and computed for each voxel the mismatch effect (intention–consequence mismatch condition − intention–consequence match condition) both in the actual harm and actual no-harm consequences. We found a significant position-by-consequence interaction ($F_{(4,128)} = 9.26, p < 0.001$), such that the mismatch effect in the harm consequence (i.e., accidental > intentional harm) decreased from aIFG to pIFG, whereas the mismatch effect in the no-harm consequence (i.e., failed attempt > genuine no-harm) increased from aIFG to pIFG. The effect of position was significant for both the harm and no-harm consequences with F values >2.58, p values <0.05.

The individual differences analysis revealed that the activation in the rTPJ in the failed attempt condition negatively correlated with trait forgiveness ($r = -0.32$, $p = 0.068$; Fig. 3*C*), while the activation in the rAI in the accidental condition positively corre-

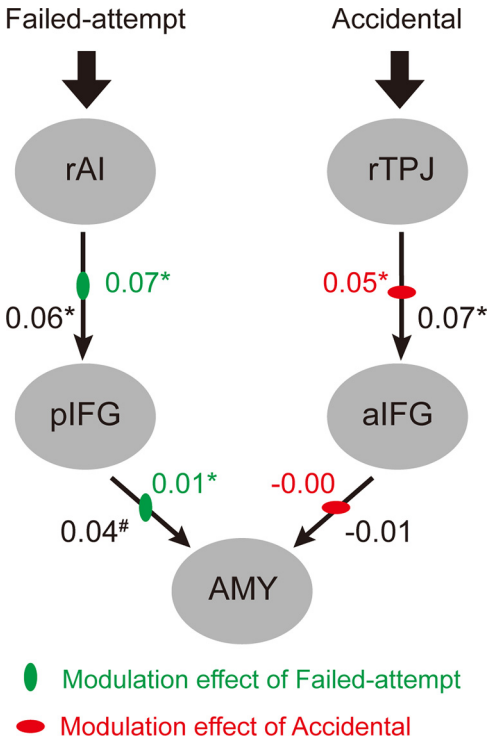

Figure 5. Parameters (connectivity strength) of the winning model. The numbers shown in black indicate the strength of the intrinsic connectivity, and the numbers shown in color indicate the strength of the modulatory effects of the failed-attempt (green vertical ellipse) and the accidental (red horizontal ellipse) conditions. Asterisks indicate significant planned one-sample *t* test with reference to zero (*$p < 0.05$, #$p < 0.1$). AMY, amygdala; rAI, right anterior insula; aIFG, anterior inferior frontal gyrus; pIFG, posterior inferior frontal gyrus; rTPJ, right temporo-parietal junction.

lated with trait revenge ($r = 0.38$, $p < 0.05$; Fig. 3*D*). To further explore the potential source of the individual differences in the activation in rAI and rTPJ, we performed backward stepwise multiple regression analysis using each participant's rAI activation in the accidental condition and rTPJ activation in the failed-attempt condition as the dependent variable in the two regression models, respectively. For the accidental condition, we initially included five predictors (Machiavelli score, self-construal score, forgiveness trait, punishment amount in the accidental condition, and amygdala activation in the accidental condition). Following stepwise reduction, the resulting model captured 21% of the variance ($F = 3.90$, $p = 0.031$) in the rAI activation (accidental condition) with one trait variable (forgiveness trait: $t = -2.14$, $p = 0.041$) and one brain variable (amygdala activation, $t = 1.98$, $p = 0.058$). Similarly, for the failed-attempt condition, we initially included five predictors (Machiavelli score, self-construal score, revenge trait, punishment amount in the failed-attempt condition, and amygdala activation in the failed-attempt condition). Following stepwise reduction, the resulting model captured 15% of the variance ($F = 5.26$, $p = 0.029$) with one trait variable (vengeful trait: $t = 2.29$, $p = 0.029$). The regression findings were consistent with the hypothesis-driven correlation analysis. Together, these findings indicated that (1) participants with higher trait forgiveness evidenced lower rAI activation when finding out an innocent intention after being accidentally harmed, and (2) the participant with higher trait revenge showed higher rTPJ activation when finding out an aggressive intention.

Effective connectivity results

Once the brain structures sensitive to intention processing were identified (rAI and pIFG for the failed-attempt condition; rTPJ and aIFG for the accidental condition), we set out to test whether the activity of these intention-sensitive regions influenced the activity in the amygdala (after the true intention was revealed) or were influenced by the amygdala. We built and compared 42 models differing in the intrinsic connectivity (from amygdala or toward amygdala), modulatory effect, and input (Fig. 4). The winning family had unilateral intrinsic connectivity from the intention-sensitive structures (rAI and rTPJ) via the IFG toward the amygdala, and its input was from the intention-sensitive regions (Fig. 4). As can be seen from Figure 5, the failed-attempt condition enhanced the connectivity from the intention-sensitive region (rAI) to the pIFG, and from the pIFG to the amygdala, which is consistent with the upregulation hypothesis; namely, a blameworthy intention could increase the activation in the amygdala via the modulation of the prefrontal cortex. Moreover, the strength of the modulatory effect of failed-attempt on the connectivity from the pIFG to the amygdala positively correlated with the punishment amount in the failed-attempt condition ($r = 0.34$, $p = 0.053$). For the accidental condition, we observed a significant modulatory effect on the connectivity from the intention-sensitive region (rTPJ) to the aIFG. For the connectivity from the aIFG to the amygdala, the direction of the modulatory effect was as predicted (decreasing), although its strength did not reach significance.

Discussion

Combining an interactive game and fMRI, we investigated the neural processing of innocent and blameworthy intentions and its bearing on proper adjustment of reactive punishment in an interpersonal transgression context. Unlike prior work (Buckholtz et al., 2008; Young and Saxe, 2009; Treadway et al., 2014), we placed the participants in a second-person position and measured both behavioral and neural responses to different intentions and consequences of transgressions. Specifically, we were interested in the neural mechanism by which the processing of intention gates the activations of the emotional system responsible for reactive punishment. In line with a previous study (Treadway et al., 2014), we found a corticolimbic circuit responsible for the processing of intention to regulate the neural processing of punishment. Improving on the previous research, we showed that this circuit not only functions to suppress the responses to the emotionally salient negative consequence, as in the case of unintentional or accidental harm, but also augments neural responses to the blameworthy but unrealized intention to harm. These complementary findings confirmed the critical role of the corticolimbic circuit in mediating the impact of the neural processing of intention on the formation of proper responses to interpersonal transgression (Buckholtz and Marois, 2012).

Moral and legal judgments are influenced by the emotional content of the event being judged (Greene et al., 2001; Darley, 2009). However, emotions are not the sole determinant of human cognition and action; the emotion-driven responses are susceptible to the judgment of reason. In the context of interpersonal transgression, inference of the agent's intention, or "what is in the heart" as quoted in the Introduction, interacts with the emotional value of the action and consequence to form a proper reaction: to forgive an accidentally inflicted harm and to condemn a blameworthy although unrealized intent (Young et al., 2007). By incorporating both aspects in an interactive game, we were able to demonstrate, in each case, that punishment severity was gated by

the processing of agent intention. Moreover, for the first time in this line of research (Young et al., 2007, 2010; Buckholtz et al., 2008; Young and Saxe, 2009; Treadway et al., 2014), we showed that revenge- and forgiveness-related personalities influence, both at the behavioral and the neural levels, the processing of intention and the formation of punishment. These new findings also reflect the strength of the interactive paradigm, which is more lifelike and closer to everyday social interactions (Schilbach et al., 2013).

Although the neural processing of accident and failed attempts are similar in certain aspects, they differ in important ways. One notable difference is the extent to which these two types of processing rely on the rTPJ. Previous evidence is mixed concerning whether the rTPJ is specifically involved in one type of processing or similarly involved in both types of processing of intention. On the one hand, Young et al. (2007) showed that reading scenarios in which the protagonist had blameworthy intention but was unable to actualize the bad intention (i.e., failed-attempt) activated the rTPJ. Such activation was not observed for the accidental condition (see also Young and Saxe, 2008; Young et al., 2010). On the other hand, combining multivariate pattern analysis with fMRI, Koster-Hale et al. (2013) found that the rTPJ activation pattern could distinguish the condition in which the participants read an accidental harm-related scenario from the condition in which the participants read an intentional harm-related scenario. This inconsistency is difficult to reconcile, and we acknowledge that our findings are not decisive in this respect due to the differences in the nature of the tasks used (second-person interpersonal interaction vs third-party imagination). However, our data do suggest that the recruitment of the rTPJ in intention processing and moral judgment depends on the degree to which such processing is required and is modulated by individual differences in vindictiveness-related personalities (e.g., revenge trait). Specifically, previous studies have shown that pardoning an accidental harm is more difficult than blaming a failed-attempt because the former relies more on the ability to use the intention to suppress an emotionally salient negative consequence (i.e., the actual harm) than the latter, where no emotionally salient event exists that competes with the blameworthy intention (Young et al., 2007). In line with this, developmental studies showed that children first use mental state information to assign blame for attempted harms and only later learn to mitigate blame for accidents (Baird and Astington, 2004). Indeed, our individual difference analysis revealed that people with a highly vengeful trait both showed higher rTPJ activation and more severe punishment in the failed-attempt condition, suggesting that heightened sensitivity to blameworthy intention may result from a revengeful disposition and lead to more severe retributive practices.

Another difference between the processing of accidental harm and failed attempt is reflected in the rostrocaudal distribution of sensitivity to the two types of processing in the left IFG (Fig. 3*B*). Given that in our task, the partner's intention was revealed after the transgression consequence, the recruitment of the IFG may reflect the retrieval of the consequence (harm vs no-harm) and its integration with the intention at the time that it was revealed. The functional dissociation is in line with the rostrocaudal gradient consistently implicated in previous studies. It is proposed that the more rostral regions support more abstract, and probably higher-level cognitive processes (Badre and Wagner, 2007). This dissociation again suggests that the processing of intention in moral judgment does not rely on a homogeneous neural substrate; rather, it is modulated by the emotional valence of the consequence and the intention behind the action.

The rAI activation was higher in the failed-attempt than in the no_harm condition, and its activation in the accidental condition was modulated by individual differences in trait forgiveness. The right AI has been consistently associated with the processing of negative experiences in social interactions, such as broken promises (Baumgartner et al., 2009), social exclusion (Eisenberger, 2012), aggression (Krämer et al., 2007), interpersonal guilt (Yu et al., 2014), and being treated unfairly (Sanfey et al., 2003). More recently, Liljeholm et al. (2014) demonstrated that the right AI tracked the interaction between intentionality and harmful consequence in an interpersonal transgression context, such that intentionally inflicted bodily harm (e.g., delivering aversive salty tea to the participants) elicited higher AI responses than unintentionally inflicted harm. Thus, it is not surprising to observe higher activation of the right AI in the failed-attempt condition, where the participant perceived the partner's malicious intent. Interestingly, the activation of the right AI in the accidental-harm condition was attenuated in those who had higher trait forgiveness (i.e., those who readily forgave harms as long as they were not intentional). This finding lends support to the idea that letting go of the negative reactive attitude is an important component of forgiveness (Griswold, 2007; Murphy, 2003).

Both consistent with and complimentary to a previous finding (Treadway et al., 2014), we showed that the cortical processing of intention not only serves to suppress inappropriate affective responses to unintentional harm (as in Treadway et al., 2014), but such processing also drives proper affective responses to blameworthy intention in the absence of actual damage. The latter mechanism is crucial for the long-term survival of social beings in that it serves as a warning signal to the individual that certain partners are dangerous and should be avoided in future encounters (Algoe, 2012). Specifically, our effective connectivity analysis suggested that the IFG functions as an interface where the information concerning the partner's intention triggers a downregulation or upregulation of the emotion-related brain structures (i.e., the amygdala) that may drive the reactive punishment decision (Buckholtz et al., 2008; Treadway et al., 2014). Note that the modulatory effects are fairly large in relation to the intrinsic connectivity (Fig. 5). This means that the top-down modulatory effects range between 25% and >100%. These modulatory effects are, in proportional terms, quite substantial.

To conclude, by combining an interactive game and fMRI, we demonstrated that corticolimbic circuits gate the influence of the processing of agent intention on the understanding and evaluation of interpersonal transgression. The left inferior frontal cortex, which does not consistently show up in the previous studies with third-party moral judgment, plays a critical role in our second-person paradigm and exhibits a rostrocaudal dissociation in relation to the types of intention (innocent vs blameworthy). The results have implications for our conceptualization of reactive attitudes and moral judgments by demonstrating how the processing of affect and intention may be integrated at the brain and behavioral levels.

References

Algoe SB (2012) Find, remind, and bind: the functions of gratitude in everyday relationships. Soc Personal Psychol Compass 6:455–469. CrossRef

Badre D, Wagner AD (2007) Left ventrolateral prefrontal cortex and the cognitive control of memory. Neuropsychologia 45:2883–2901. CrossRef Medline

Baird JA, Astington JW (2004) The role of mental state understanding in the

development of moral cognition and moral action. New Dir Child Adolesc Dev 2004:37–49. CrossRef Medline

Baumgartner T, Fischbacher U, Feierabend A, Lutz K, Fehr E (2009) The neural circuitry of a broken promise. Neuron 64:756–770. CrossRef Medline

Buckholtz JW, Marois R (2012) The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat Neurosci 15:655–661. CrossRef Medline

Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, Marois R (2008) The neural correlates of third-party punishment. Neuron 60: 930–940. CrossRef Medline

Burnham KP, Anderson DR (2004) Multimodel inference understanding AIC and BIC in model selection. Sociol Methods Res 33:261–304. CrossRef

Darley JM (2009) Morality in the law: the psychological foundations of citizens' desires to punish transgressions. Annu Rev Law Soc Sci 5:1–23. CrossRef

Eisenberger NI (2012) The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. Nat Rev Neurosci 13:421–434. CrossRef Medline

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. Neuroimage 19:1273–1302. CrossRef Medline

Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108. CrossRef Medline

Griswold C (2007) Forgiveness: a philosophical exploration. Cambridge, UK: Cambridge UP.

Inui K, Tran TD, Hoshiyama M, Kakigi R (2002) Preferential stimulation of Adelta fibers by intra-epidermal needle electrode in humans. Pain 96: 247–252. CrossRef Medline

Koster-Hale J, Saxe R, Dungan J, Young LL (2013) Decoding moral judgments from neural representations of intentions. Proc Natl Acad Sci U S A 110:5648–5653. CrossRef Medline

Krämer UM, Jansma H, Tempelmann C, Münte TF (2007) Tit-for-tat: the neural basis of reactive aggression. Neuroimage 38:203–211. CrossRef Medline

Levitan W (2007) Abelard and Heloise: the letters and other writings. Indianapolis, IN: Hackett

Liljeholm M, Dunne S, O'Doherty JP (2014) Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. J Neurosci 34:11339–11348. CrossRef Medline

McNamee D, Rangel A, O'Doherty JP (2013) Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nat Neurosci 16:479–485. CrossRef Medline

Murphy JG (2003) Getting even: forgiveness and its limits. New York: Oxford UP.

Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. Neuroimage 22:1157–1172. CrossRef Medline

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP (2010) Comparing families of dynamic causal models. PLoS Comput Biol 6:e1000709. CrossRef Medline

Rye MS, Loiacono DM, Folck CD, Olszewski BT, Heim TA, Madia BP (2001) Evaluation of the psychometric properties of two forgiveness scales. Curr Psychol 20:260–277. CrossRef

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. Science 300:1755–1758. CrossRef Medline

Schilbach L, Timmermans B, Reddy V, Costall A, Bente G, Schlicht T, Vogeley K (2013) Toward a second-person neuroscience. Behav Brain Sci 36: 393–414. CrossRef Medline

Seymour B, Singer T, Dolan R (2007) The neurobiology of punishment. Nat Rev Neurosci 8:300–311. CrossRef Medline

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017. CrossRef Medline

Strawson PF (2008) Freedom and resentment and other essays. New York: Routledge

Stuckless N, Goranson R (1992) The vengeance scale: development of a measure of attitudes toward revenge. J Soc Behav Pers 7:25–42.

Treadway MT, Buckholtz JW, Martin JW, Jan K, Asplund CL, Ginther MR, Jones OD, Marois R (2014) Corticolimbic gating of emotion-driven punishment. Nat Neurosci 17:1270–1275. CrossRef Medline

Young L, Saxe R (2008) The neural basis of belief encoding and integration in moral judgment. Neuroimage 40:1912–1920. CrossRef Medline

Young L, Saxe R (2009) Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. Neuropsychologia 47:2065–2072. CrossRef Medline

Young L, Cushman F, Hauser M, Saxe R (2007) The neural basis of the interaction between theory of mind and moral judgment. Proc Natl Acad Sci U S A 104:8235–8240. CrossRef Medline

Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proc Natl Acad Sci U S A 107:6753–6758. CrossRef Medline

Yu H, Hu J, Hu L, Zhou X (2014) The voice of conscience: neural bases of interpersonal guilt and compensation. Soc Cogn Affect Neurosci 9:1150–1158. CrossRef Medline